MedCrave
Step into the World of Research

Biometrics & Biostatistics International Journal

A Discrete Lindley Distribution with Applications in Biological Sciences

Abstract

In this paper, a discrete Lindley distribution, a discrete analogue of continuous Lindley distribution using infinite series method of discretization, has been proposed and investigated. Its generating functions, moments and moments based measures including coefficients of variation, skewness, kurtosis, index of dispersion have been obtained and discussed. Both the method of moments and the method of maximum likelihood estimation have been discussed for estimating its parameter. Applications of the proposed distribution have been explained through two examples of observed real datasets from biological sciences.

Keywords: Lindley distribution; Discretization; Moment generating function; Moments; Estimation; Goodness of fit

Research Article

Volume 7 Issue 2 - 2018

Berhane Abebe and Rama Shanker*

Department of Statistics, College of Science, Eritrea Institute of Technology, Eritrea

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

Received: January 24, 2018 | **Published:** February 09, 2018

Introduction

In the last few decades many papers appeared in the statistical literature on the discretization of continuous distributions. The main reasons for discretizing a continuous distributions are two fold, namely, (i) the discrete analogue of a continuous distribution provide probability mass function (pmf) that can compete with the classical discrete distributions commonly used in the statistical analysis of count data and (ii) the discrete analogue of a continuous distribution avoids the use of a continuous distribution in the case of strictly discrete data.

According to Lai [1], discretization of a continuous lifetime model is an interesting and intuitively appealing approach to derive a discrete lifetime model corresponding to the continuous one. It has been observed that many times in the real world the original variables may be continuous in nature but discrete by observation and , therefore, it is reasonable and convenient to model the situation by an appropriate discrete distribution generated from the underlying continuous distribution preserving one or more important characteristics including probability density function (pdf), moment generating function (mgf), moments, hazard rate function, mean residual life function etc., of the continuous distribution.

There are several methods available in Statistics literature to derive a discrete distribution from a continuous distribution. One of the first proposed discretization methods is based on the definition of pmf that depends on an infinite series. The method of discretization by an infinite series was firstly considered by Good [2] who has proposed the discrete Good distribution to model the population frequencies of species and the estimation of parameters. A random variable Y is said to have a discrete Good distribution if its pmf can be expressed as

$$P(Y=y)=\frac{\alpha^{y} y^{\beta}}{\sum_{j=1}^{\infty} \alpha^{j} j^{\beta}} \; ;y=0,1,2,\ldots., \; \beta\in R \text{ and } \alpha\in(0,1) \quad (1.1)$$

The method of infinite series is characterized by the following definition

Definition 1.1: Let X be a continuous random variable having pdf $f_X(x)$ with support on R. Then the corresponding discrete random variable Y has pmf given by

$$P(Y=y)=P(y;\theta)=\frac{f_X(y;\theta)}{\sum_{j=-\infty}^{\infty} f_X(j;\theta)} \; ;y\in Z \; , \quad (1.2)$$

where θ may be the vector of parameters indexing the distribution of X.

This method of discretizing a continuous distribution has been studied by several researchers including Kulasekara and Tonkyn [3], Doray and Luong [4], Sato et al. [5], Nekoukhou et al. [6], are some among others, who proposed a version of the method when the continuous random variable of interest is defined on R_+. Thus, if the random variable X is defined on R_+, the pmf of Y can be defined as

$$P(Y=y)=P(y;\theta)=\frac{f_X(y;\theta)}{\sum_{j=0}^{\infty} f_X(j;\theta)} \; ;y\in Z_+ \quad (1.3)$$

Note that the discrete distribution generated using the method of infinite series may not always have a compact form due to the normalizing constant.

In the present paper, a discrete Lindley distribution (DLD), a discrete analogue of continuous Lindley distribution proposed by Lindlley [7] has been introduced using a discretization method based on an infinite series. Its moment generating function, moments and moments based measures have been obtained and discussed. Both the method of moment and the method of maximum likelihood estimates give the same estimator of the parameter of DLD. The applications and the goodness of fit of the proposed distribution have been explained using two real datasets from biological sciences and the fit has been compared with other discrete distributions.

A Discrete Lindley Distribution

The pdf and the cdf of a continuous random variable X having Lindley distribution introduced by Lindlley [7] are given by

$$f_1(x;\theta)=\frac{\theta^2}{\theta+1}(1+x)\,e^{-\theta x};\quad x>0,\quad \theta>0 \qquad (2.1)$$

$$F_1(x;\theta)=1-\left[1+\frac{\theta x}{\theta+1}\right]e^{-\theta x};\quad x>0,\quad \theta>0 \qquad (2.2)$$

Ghitany et al. [8] has discussed its various mathematical and statistical properties including its shapes for varying values of parameter, moments based measures, hazard rate function, mean residual life function, stochastic ordering, mean deviations, order statistics, Renyi entropy measure, Bonferroni and Lorenz curves and stress-strength reliability along with estimation of parameter and application for modeling waiting time in a bank. Shanker et al. [9] have detailed comparative study on applications of Lindley distribution and exponential distribution for modeling lifetime data and observed that both distributions are competitor to each other for modeling lifetime data from various fields of knowledge. In fact, in some datasets exponential distribution gives much closer fit than the Lindley distribution and in some datasets Lindley distribution gives better fit than exponential distribution.

Using the definition (1.1), the pmf of the discrete random variable Y corresponding to a continuous random variable X following Lindley distribution (2.1) can be obtained as

$$P_1(Y=y)=P_1(y;\theta)=\frac{\left(e^{\theta}-1\right)^2}{e^{2\theta}}(1+y)e^{-\theta y};\ y=0,1,2,\ldots,\ \theta>0 \qquad (2.3)$$

We would call this distribution, a discrete Lindley distribution (DLD). The nature and behavior of DLD for varying values of its parameter θ has been shown graphically in figure 1.

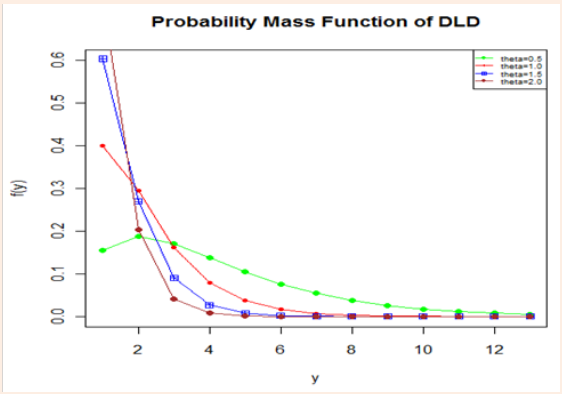

Figure 1: The pmf plot of DLD for varying values of the parameter θ.

The survival function, $S(y;\theta)$ and the cumulative distribution function (cdf), $F(y;\theta)$ of DLD can be obtained as

$$S(y;\theta)=\left[\frac{\left(e^{\theta}-1\right)y+\left(2e^{\theta}-1\right)}{e^{2\theta}}\right]e^{-\theta y};y=0,1,2,\ldots,\ \theta>0 \qquad (2.4)$$

$$F_2(y;\theta)=1-\left[\frac{\left(e^{\theta}-1\right)y+\left(2e^{\theta}-1\right)}{e^{2\theta}}\right]e^{-\theta y};y=0,1,2,\ldots,\ \theta>0 \qquad (2.5)$$

Graph of cumulative distribution function of DLD for varying values of its parameter θ has been shown in figure 2.

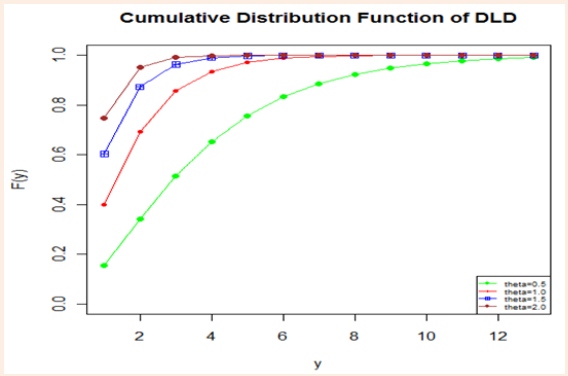

Figure 2: The cdf plot of DLD for varying values of the parameter θ.

Since $\frac{P(y+1;\theta)}{P(y;\theta)}=\left[1+\frac{1}{1+y}\right]e^{-\theta}$ is a decreasing function of $y\geq 0$, $P(y;\theta)$ is log-concave and therefore, the DLD has an increasing hazard rate. Further, $\left[P(y;\theta)\right]^2\geq P(y-1;\theta)\cdot P(y+1;\theta)$ for $y\geq 0$, which implies unimodality, by theorem 3 of Keilson and Gerber [10]. The interrelationship between log-concavity, unimodality and increasing hazard rate of discrete distributions are available in Grandell [11].

Moments, Skewness, Kurtosis and Index of Dispersion

The probability generating function (G(t)) and the moment generating function (M(t)) of DLD can be obtained as

$$G(t)=\left(\frac{e^{\theta}-1}{e^{\theta}-t}\right)^2,\ \text{for } t\neq e^{\theta}, \qquad (3.1)$$

and

$$M(t)=\frac{\left(e^{\theta}-1\right)^2}{e^{2\theta}\left(1-e^{-(\theta-t)}\right)^2},\ \text{for } t\neq\theta. \qquad (3.2)$$

It can be easily verified that the function in (3.2) is infinitely differentiable with respect to t, since it involves exponential terms of its argument. This means that all moments about origin μ_r', $r\geq 1$ of DLD can be obtained. The first four moments about origin of DLD can thus be obtained as

$$\mu_1'=\frac{2}{\left(e^{\theta}-1\right)}$$

$$\mu_2'=\frac{2\left(e^{\theta}+2\right)}{\left(e^{\theta}-1\right)^2}$$

$$\mu_3'=\frac{2\left(e^{2\theta}+7e^{\theta}+4\right)}{\left(e^{\theta}-1\right)^3}$$

$$\mu_4'=\frac{2\left(e^{3\theta}+18e^{2\theta}+33e^{\theta}+8\right)}{\left(e^{\theta}-1\right)^4}$$

Using the relationship $\mu_r=E\left(Y-\mu_1'\right)^r=\sum_{k=0}^{r}\binom{r}{k}\mu_k'\left(-\mu_1'\right)^{r-k}$ between central moments and moments about origin, the central moments of DLD are obtained as

$$\mu_2=\frac{2e^{\theta}}{\left(e^{\theta}-1\right)^2}$$

$$\mu_3=\frac{2e^{\theta}\left(e^{\theta}+1\right)}{\left(e^{\theta}-1\right)^3}$$

$$\mu_4=\frac{2e^{\theta}\left(e^{2\theta}+10e^{\theta}+1\right)}{\left(e^{\theta}-1\right)^4}$$

The expressions for coefficient of variation (C.V), coefficient of skewness $\left(\sqrt{\beta_1}\right)$, coefficient of kurtosis $\left(\beta_2\right)$ and index of dispersion $\left(\gamma\right)$ of DLD are expressed as

$$C.V=\frac{\sigma}{\mu_1'}=\frac{\sqrt{e^{\theta}}}{\sqrt{2}}$$

$$\sqrt{\beta_1}=\frac{\mu_3}{\left(\mu_2\right)^{3/2}}=\frac{\left(e^{\theta}+1\right)}{\sqrt{2e^{\theta}}}$$

$$\beta_2=\frac{\mu_4}{\mu_2^{\,2}}=\frac{e^{2\theta}+10e^{\theta}+1}{2e^{\theta}}$$

$$\gamma=\frac{\sigma^2}{\mu_1'}=\frac{e^{\theta}}{\left(e^{\theta}-1\right)}.$$

Since $\gamma>1$ always for $\theta>0$, a DLD is over dispersed $(\sigma^2>\mu_1')$. This means that DLD can be used to model over-dispersed data from any fields of knowledge.

Table 1 summarizes the nature and behavior of coefficient of variation (C.V), coefficient of skewness, coefficient of kurtosis and index of dispersion (ID) of DLD for selected values of the parameter θ.

It is obvious from above table that the mean, variance, and index of dispersion of DLD are decreasing for increasing values of the parameter θ, while coefficient of variation, coefficient of skewness and coefficient of kurtosis of DLD are increasing for increasing values of parameter θ. Since $\sigma^2>\mu$, DLD is a suitable model for over-dispersed data.

Table 1: Values of Descriptive statistics of DLD for varying values of parameter θ.

θ	Values of Descriptive Statistics					
	Mean	**Variance**	**C.V**	**Skewness**	**Kurtosis**	**ID**
0.5	3.0380	7.8354	0.9079	1.4586	6.1276	2.5415
1	1.6400	1.8413	1.1658	1.5947	6.5431	1.5820
1.5	0.5744	0.7394	1.4969	1.8310	7.3524	1.2872
2	0.3131	0.3620	1.9221	2.1822	8.7622	1.1565
2.5	0.1789	0.1948	2.4680	2.6706	11.1323	1.0894
3	0.1048	0.1103	3.1690	3.3268	15.0677	1.0524
3.5	0.0623	0.0642	4.0691	4.1920	21.5728	1.0314
4	0.0373	0.0380	5.2248	5.3205	32.3082	1.0187

Parameter Estimation

Method of Moment Estimate (MOME): Equating the population mean to the corresponding sample mean, the MOME $\tilde{\theta}$ of θ of DLD is given by

$\tilde{\theta}=\ln\left(\frac{\bar{y}+2}{\bar{y}}\right)$, where $\bar{y}$ is the sample mean.

Maximum Likelihood Estimate (MLE): Let $\left(y_1,y_2,y_3,\ldots,y_n\right)$ be a random sample from DLD (2.3)). The likelihood function, L of (2.3) is given by

$$L=\left(\frac{\left(e^{\theta}-1\right)^2}{e^{2\theta}}\right)^n e^{-n\theta\bar{y}}\prod_{i=1}^{n}\left(1+y_i\right).$$

The natural log likelihood function is thus obtained as

$$\ln L=2n\ln\left(e^{\theta}-1\right)+\sum_{i=1}^{n}\ln\left(1+y_i\right)-n\theta\bar{y}-2n\theta.$$

The maximum likelihood estimates (MLE) $\hat{\theta}$ of parameter θ is the solution of the log-likelihood equation $\frac{d\ln L}{d\theta}=0$ and is given by

$$\frac{d\ln L}{d\theta}=\frac{2n}{e^{\theta}-1}e^{\theta}-2n-n\bar{y}=0, \text{which gives } \hat{\theta}=\ln\left(\frac{\bar{y}+2}{\bar{y}}\right).$$

This means that both MOME and MLE give the same estimator for the parameter θ.

Goodness of Fit

As we have seen that DLD is over-dispersed and hence it can be applied to model over-dispersed data. In general, the discrete data in biological sciences are over-dispersed and due to this we have taken two examples of observed real datasets from biological sciences. The first dataset is the data regarding number of Homocytometer yeast cell counts per square, available in Gosset [12]. The second dataset is the data regarding the number of European corn borer available in Mc Guire et al. [13]. The goodness of fit of DLD has been compared with one parameter Poisson distribution (PD) which is equi-dispersed and Poisson-Lindley distribution (PLD) which is over-dispersed and proposed by Sankaran [14] as a Poisson mixture of Lindley distribution, introduced by Lindley [7]. It should be noted that Shanker and Hagos [15] has detailed study on applications of PLD in biological sciences and observed that for biological sciences data, PLD is the appropriate choice (Table 2&3).

It is obvious from the values of chi-square that DLD gives much closer fit than both PD and PLD, and hence DLD can be considered an important discrete distribution for modeling discrete data in biological sciences.

Table 2: Observed and expected number of Homocytometer yeast cell counts per square observed by Gosset [12].

Number of Red Mites	Observed Frequency	Expected Frequency		
		PD	PLD	DLD
0	213	202.10	234.00	222.35
1	128	138.00	99.40	113.14
2	37	47.10	40.50	43.18
3	18	10.70	16.00	14.65
4	3	1.80	6.20	4.66
5	1	0.20	2.40	1.42
6	0	0.10	1.50	0.59
Total	400	400.00	400.00	400.00
ML Estimate $(\hat{\theta})$		0.6825	1.9502	1.3687
χ^2		10.0900	11.0610	3.2511
d.f.		2	2	2
p-value		0.0178	0.0114	0.3545

Table 3: Observed and expected number of European corn-borer of McGuire et al [13].

Number of European Corn-Borer	Observed Frequency	Expected Frequency		
		PD	PLD	DLD
0	188	169.40	194.00	184.80
1	83	109.80	79.50	90.47
2	36	35.60	31.30	33.24
3	14	7.80	12.00	10.84
4	2	1.20	4.50	3.32
5	1	0.20	2.70	1.36
Total	324	324.00	324.00	324.00
ML Estimate $(\hat{\theta})$		0.6481	2.0425	1.4075
χ^2		15.2000	1.2970	1.0425
d.f.		2	2	2
p-value		0.0040	0.7297	0.7910

Conclusions

In this paper, a discrete Lindley distribution (DLD), a discrete analogue of continuous Lindley distribution, has been proposed and investigated. Its moment generating function, moments and moments based measures including coefficients of variation, skewness, kurtosis, and index of dispersion have been obtained and their nature have been discussed numerically. Both the method of moments and the method of maximum likelihood estimation have been discussed for estimating its parameter. The applications of DLD have been discussed with two examples of observed real datasets from biological sciences. The DLD gives much closer fit over Poisson distribution (PD) and Poisson Lindley distribution (PLD).

References

1. Lai CD (2013) Issues concerning constructions of discrete lifetime models, Qual Techno Quant Mang 10(2): 251-262.
2. Good LJ (1953) The population frequencies of species and the estimation of population parameters, Biometrika 40: 237- 264.
3. Kulasekara KB, Tonkyn DW (1992) A new discrete distribution with application to survival, dispersal and dispersion, Commun Stat Simul Comput 21: 499-518.
4. Doray LG, Luong A (1997) Efficient estimator s for the Good family, Commun Stat Simul Comput 21: 499-518.
5. Sato H, Ikota M, Aritoshi S, Masuda H (1999) A new defect distribution in meteorology with a consistent discrete exponential formula and its applications, IEEE Trans Semicond Manufactur 12(4): 409-418.
6. Nekoukhou VM, Alamatsaz MH, Bidram H (2012) Discrete Generalized exponential distribution, Communications in Statistics-Theory & Methods 41: 2000-2013.
7. Lindley DV (1958) Fiducial distributions and Bayes' theorem, Journal of the Royal Statistical Society, Series B 20: 102- 107.
8. Ghitany ME, Atieh B, Nadarajah S (2008) Lindley distribution and its Application, Mathematics Computing and Simulation 78: 493-506.
9. Shanker R, Hagos F, Sujatha S (2015) On Modeling of Lifetimes data using Exponential and Lindley distributions: Biometrics & Biostatistics International Journal 2(5): 1-9
10. Keilson J, Gerber H (1971) Some resilts for Discrete Unimodality, Journal of the American Statistical Association 66: 386-389.
11. Grandell J (1997) Mixed Poisson Processes, CRC Press.
12. Gosset WS (1908) The probable error of the mean, Biometrika 6(1): 1-25.
13. Mc Guire JU, Brindley TA, Bancroft TA (1957) The distribution of European corn-borer larvae pyrausta in field corn, Biometrics 13: 65-78.
14. Sankaran M (1970) The discrete Poisson-Lindley distribution, Biometrics 26: 145-149.
15. Shanker R, Hagos F (2015) On Poisson-Lindley distribution and its Applications to Biological Sciences, Biometrics & Biostatistics International Journal 2(4): 1-5.

Citation: Abebe B, Shanker R (2018) A Discrete Lindley Distribution with Applications in Biological Sciences. Biom Biostat Int J 7(2): 00189. DOI: 10.15406/bbij.2018.07.00189