

On estimation of the probability mass function and the cumulative distribution function of a natural discrete one parameter polynomial exponential distribution

Indrani Mukherjee^a, Sudhansu S. Maiti^a, Rama Shanker^b

^aDepartment of Statistics

Visva-Bharati University

Santiniketan-731 235

India

indranimukherjee.stat@gmail.com, dssm1@rediffmail.com

^bDepartment of Statistics

Assam University

Silchar-788011

India

shankerrama2009@gmail.com

ABSTRACT

In this paper, a new natural discrete analog of the one parameter polynomial exponential (OPPE) distribution as a mixture of a number of negative binomial distributions has been proposed and is called as a natural discrete one parameter polynomial exponential (NDOPPE) distribution. This distribution is a generalized version of natural discrete Lindley (NDL) distribution, proposed and studied in literature. Maximum likelihood estimator (MLE) and uniformly minimum variance unbiased estimator (UMVUE) of the probability mass function (PMF) and the cumulative distribution function (CDF) of the NDOPPE distribution have been derived. The estimators have been compared with respect to their mean squared errors (MSEs). Simulation study has been conducted to verify the consistency of the estimators. Three real data illustrations have been reported.

Keywords: Goodness of fit; Maximum likelihood estimator; Natural discrete Lindley distribution; Simulation; Uniformly minimum variance unbiased estimator.

2010 Mathematics Subject Classification: 62F10, 62N05.

1 Introduction

Some standard discrete distributions have been mentioned and the estimators of their probability mass functions (PMF) and cumulative distribution functions (CDF) are studied in (Maiti and Mukherjee, 2017). Sometimes these models are too restrictive. As for example, Poisson model is not appropriate because it imposes the restriction of equidispersion in the modeled data. Similarly, binomial model imposes the restriction of under-dispersion. As a result, various models prescribed using discrete concentration and discrete analog approaches that are

less restrictive (see, (Nakagawa and Osaki, 1975), (Famoye, 1993), among others). The most recent discrete distributions are due to (Stein and Dattero, 1984), (Roy, 2002), (Roy, 2003), (Roy, 2004), (Krishna and Pundir, 2009), (Jazi, Lai and Alamatsaz, 2010), (Gomez-Deniz, 2010), (Gomez-Deniz and Calderin-Ojeda, 2011), (Maiti, Dey and Sarkar (Mondal), 2018), some among others.

To generate a natural discrete analog of the one parameter polynomial exponential (OPPE) distribution, we use the well-known fact that the geometric and the negative binomial distributions are the natural discrete analogs of the exponential and the gamma distributions respectively. This is the motivation behind proposing a new natural discrete analog to the OPPE distribution by mixing discrete counterparts of the exponential and gamma distributions.

The PMF of the random variable X of a natural discrete one parameter polynomial exponential (NDOPPE) distribution can be written as

$$f(x) = h(\theta)p(x)(1 - \theta)^x, \quad x = 0, 1, \dots, \quad 0 < \theta < 1, \quad (1.1)$$

where, $h(\theta) = \frac{1}{\sum_{k=0}^r a_k \frac{\Gamma(k+1)}{\theta^{k+1}}}$, $p(x) = \sum_{k=0}^r a_k \Gamma(k+1) \binom{x+k}{x}$.

The distribution can also be written as

$$f(x) = h(\theta) \sum_{k=0}^r a_k \frac{\Gamma(k+1)}{\theta^{k+1}} f_{NB}(x; k+1, \theta), \quad (1.2)$$

where $f_{NB}(x; k+1, \theta)$ is the PMF of a negative binomial distribution with number of success $(k+1)$ and success probability is θ . a_k 's are known non-negative constants and r is a known positive integer. The distribution is a finite mixture of $(r+1)$ negative binomial distributions. The CDF is given by

$$F(x) = h(\theta) \sum_{k=0}^r a_k \frac{\Gamma(k+1)}{\theta^{k+1}} I_{\bar{\theta}}(k+1, x+1), \quad \bar{\theta} = 1 - \theta, \quad (1.3)$$

where $I_p(a, b) = \frac{1}{B(a, b)} \int_0^p x^{a-1} (1-x)^{b-1} dx$.

We have mentioned two special cases which are given below,

1. $r = 0$, $a_0 = 1$ gives the geometric distribution ,
2. $r = 1$, $a_0 = 1$, $a_1 = 1$ gives a natural discrete Lindley (NDL) distribution [c.f. (Ahmed and Afify, 2019)]. The PMF and the CDF in this case are given by

$$f(x) = \frac{\theta^2}{(1+\theta)} (2+x)(1-\theta)^x, \quad x = 0, 1, \dots; \theta \in (0, 1)$$

and

$$F(x) = 1 - \frac{1+\theta+\theta x}{(1+\theta)} (1-\theta)^x, \quad x = 0, 1, \dots; \theta \in (0, 1),$$

respectively.

The problem of estimation of the PMF and the CDF is interesting for many reasons. For example, the PMF and the CDF can be used for estimation of differential entropy, Rényi entropy,

Kullback-Leibler divergence, Fisher's Information, Cumulative residual entropy, quantile function, Lorenze curve, Hazard rate function, Mean remaining life function etc.

Most of the times, emphasis is given to infer the parameters involved in the distribution and the study is concentrated on measuring the efficiency of these estimators. No such effort has been made to find out estimator of the PMF and the CDF of these discrete random variables. Plugging in the MLE of parameter gives, by invariance property, the MLE of the PMF and the CDF. It is to be noted that all these estimators are biased. Mere substitution of the UMVUE of parameter(s) will not provide the UMVUE of the PMF and that of the CDF. Therefore, comparing UMVUE of the PMF and that of the CDF (the only unbiased estimators in our study) with other estimators seem to be an interesting study.

Similar type of studies have appeared in recent literature for some continuous distributions by (Asrabadi, 1990), (Bagheri, Alizadeh, Baloui and Nadarajah, 2014), (Dixit and Jabbari, 2010), (Dixit and Jabbari, 2011), (Jabbari and Jabbari, 2010), (Maiti and Mukherjee, 2017), (Mukherjee and Maiti, 2018), some among others.

The paper is organized as follows. Section 2 deals with MLE of the PMF and the CDF of a NDOPPE distribution. Section 3 is devoted to finding out the UMVUE of the PMF and the CDF. In section 4, simulation study results are reported and comparisons are made. Real-life data set is analyzed in section 5. In section 6, concluding remarks are given.

2 MLE of the PMF and the CDF

Let $X_1, X_2, \dots, X_n$ be a random sample of size n drawn from the PMF in (1.2). Here we try to find the MLE of θ which is denoted as $\hat{\theta}$. The log-likelihood of θ is given by

$$\begin{aligned} l(\theta) &= \ln L(\theta|X) \\ &= n \ln h(\theta) + \sum_{i=1}^n \ln p(X_i) + \ln(1 - \theta) \sum_{i=1}^n X_i. \end{aligned}$$

Now,

$$\begin{aligned} \frac{dl(\theta)}{d\theta} &= 0 \\ \text{i.e. } h(\theta) \sum_{k=0}^r a_k \frac{\Gamma(k+2)}{\theta^{k+2}} &= \frac{\bar{X}}{1 - \theta}. \end{aligned} \quad (2.1)$$

Since, the above equation is not of a closed form, we have to solve (2.1) numerically to obtain the MLE of θ . Theoretical expression for the mean squared error (MSE) of the MLEs are not available. MSE will be studied through simulation.

3 UMVUE of the PMF and the CDF

In this section, we obtain the UMVUE of the PMF and the CDF of a NDOPPE distribution. Also, we obtain the MSEs of these estimators.

Theorem 3.1. Let, $X_1, X_2, \dots, X_n \sim f_{NDOPPE}(x, \theta)$. Then the distribution of $T = X_1 + X_2 + \dots + X_n$ is

$$f(t) = h^n(\theta) \sum_{y_0} \dots \sum_{y_r} c(n, y_0, \dots, y_r) \binom{t + \sum_{k=0}^r (k+1)y_k - 1}{t} (1-\theta)^t,$$

$t = 0, 1, \dots$ with $y_0 + \dots + y_r = n$ and $c(n, y_0, \dots, y_r) = \frac{n!}{y_0! \dots y_r!} \prod_{k=0}^r (a_k \Gamma(k+1))^{y_k}$.

Proof. The mgf of T is

$$\begin{aligned} M_T(t) &= h^n(\theta) \left[\sum_{k=0}^r \frac{a_k \Gamma(k+1)}{\theta^{k+1}} \cdot \frac{\theta^{k+1}}{\{1 - (1-\theta)e^t\}^{k+1}} \right]^n \\ &= h^n(\theta) \left[\sum_{y_0} \dots \sum_{y_r} c(n, y_0, \dots, y_r) \right. \\ &\quad \times \frac{1}{\theta^{\sum_{k=0}^r (k+1)y_k}} \left\{ \frac{\theta}{1 - (1-\theta)e^t} \right\}^{\sum_{k=0}^r (k+1)y_k} \left. \right]. \end{aligned}$$

Hence, the distribution of T is

$$\begin{aligned} f(t) &= h^n(\theta) \sum_{y_0} \dots \sum_{y_r} c(n, y_0, \dots, y_r) \\ &\quad \times \binom{t + \sum_{k=0}^r (k+1)y_k - 1}{t} (1-\theta)^t, \end{aligned}$$

where, $c(n, y_0, \dots, y_r) = \frac{n!}{y_0! \dots y_r!} \prod_{k=0}^r (a_k \Gamma(k+1))^{y_k}$. □

Lemma 3.1. The conditional distribution of X_1 given $X_1 + X_2 + \dots + X_n = T$ is

$$\begin{aligned} f_{X_1|T}(x|t) &= \frac{p(x)}{A_n(t)} \sum_{q_0} \dots \sum_{q_r} c(n-1, q_0, \dots, q_r) \\ &\quad \times \binom{t-x + \sum_{k=0}^r (k+1)q_k - 1}{t-x}, \quad x = 0, \dots, t, \end{aligned}$$

where,

$$A_n(t) = \sum_{y_0} \dots \sum_{y_r} c(n, y_0, \dots, y_r) \binom{t + \sum_{k=0}^r (k+1)y_k - 1}{t}$$

and

$$c(n-1, q_0, \dots, q_r) = \frac{(n-1)!}{q_0! \dots q_r!} \prod_{k=0}^r (a_k \Gamma(k+1))^{q_k},$$

with $q_0 + q_1 + \dots + q_r = n-1$.

Proof.

$$\begin{aligned} f_{X_1|T}(x|t) &= \frac{f(x)f(t-x)}{f(t)} \\ &= \frac{p(x)}{A_n(t)} \sum_{q_0} \dots \sum_{q_r} c(n-1, q_0, \dots, q_r) \\ &\quad \binom{t-x + \sum_{k=0}^r (k+1)q_k - 1}{t-x}. \end{aligned}$$

□

Theorem 3.2. Let, $T = t$ be given. Then

$$\begin{aligned}\hat{f}(x) &= \frac{p(x)}{A_n(t)} \sum_{q_0} \dots \sum_{q_r} c(n-1, q_0, \dots, q_r) \\ &\times \binom{t-x + \sum_{k=0}^r (k+1)q_k - 1}{t-x}, \quad x = 0, \dots, t,\end{aligned}\quad (3.1)$$

is UMVUE for $f(x)$ and

$$\begin{aligned}\hat{F}(x) &= \frac{1}{A_n(t)} \sum_{q_0} \dots \sum_{q_r} c(n-1, q_0, \dots, q_r) \\ &\times \sum_{w=0}^x p(w) \binom{t-w + \sum_{k=0}^r (k+1)q_k - 1}{t-w}, \quad x = 0, \dots, t,\end{aligned}\quad (3.2)$$

is UMVUE for $F(x)$, where, $p(w) = \sum_{k=0}^r a_k \Gamma(k+1) \binom{w+k}{w}$.

Remark: When $r = 1$, $a_0 = 1$, $a_1 = 1$, the above expressions (3.1) and (3.2) reduce to

$$\begin{aligned}\hat{f}(x) &= \frac{(2+x)}{\sum_{k=0}^n \binom{n}{k} \binom{2n-k+t-1}{t}} \sum_{k=0}^{n-1} \binom{n-1}{k} \binom{t-x + 2(n-1) - k - 1}{t-x} \\ &, \quad x = 0, \dots, t\end{aligned}\quad (3.3)$$

and

$$\begin{aligned}\hat{F}(x) &= \sum_{w=0}^x \frac{(2+w)}{\sum_{k=0}^n \binom{n}{k} \binom{2n-k+t-1}{t}} \sum_{k=0}^{n-1} \binom{n-1}{k} \\ &\times \binom{t-w + 2(n-1) - k - 1}{t-w}, \quad x = 0, \dots, t,\end{aligned}\quad (3.4)$$

respectively.

The MSE of $\hat{f}(x)$ is given by

$$\begin{aligned}MSE(\hat{f}(x)) &= E(\hat{f}^2(x)) - f^2(x) \\ &= \sum_{t=x}^{\infty} \left[\frac{p(x)}{A_n(t)} \sum_{q_0} \sum_{q_1} \dots \sum_{q_r} c(n-1, q_0, q_1, \dots, q_r) \right. \\ &\quad \left. \binom{t-x + \sum_{k=0}^r (k+1)q_k - 1}{t-x} \right]^2 f(t) - f^2(x).\end{aligned}\quad (3.5)$$

Using Theorem 3.1 and equation (1.2) in (3.5), we can get the value of the MSE of UMVUE of the PDF.

The MSE of $\hat{F}(x)$ is given by

$$\begin{aligned}MSE(\hat{F}(x)) &= E(\hat{F}^2(x)) - F^2(x) \\ &= \sum_{t=x}^{\infty} \left[\sum_{w=0}^x \frac{p(w)}{A_n(t)} \sum_{q_0} \dots \sum_{q_r} c(n-1, q_0, \dots, q_r) \right. \\ &\quad \left. \times \binom{t-w + \sum_{k=0}^r (k+1)q_k - 1}{t-w} \right]^2 f(t) \\ &\quad - F^2(x),\end{aligned}\quad (3.6)$$

where, $p(w) = \sum_{k=0}^r a_k \Gamma(k+1) \binom{w+k}{w}$. Similarly, using Theorem 3.1 and equation (1.3) in (3.6), we can get the value of the MSE of UMVUE of the CDF.

Theoretical MSE of the UMVUE of the PMF and the CDF have been shown in Figure 1, for $a_0 = 1$, $a_1 = 1$, $x = 2$, $\theta = 0.01$ and $r = 1$. It is clear from the graph that MSE (in this case, variance) decreases as sample size increases.

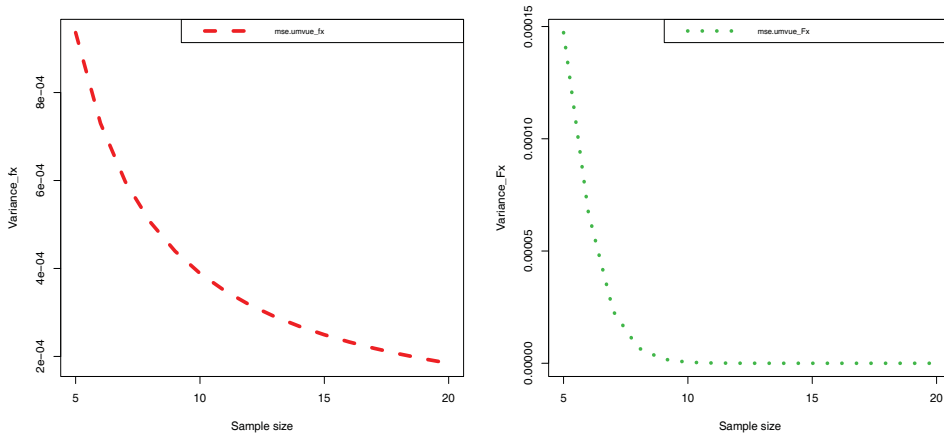


Figure 1: Graph of theoretical MSE of the UMVUE of the PMF and that of the CDF of a NDL distribution for $x = 2$, $\theta = 0.01$ and $r = 1$.

4 Simulation

Generation of random sample $X_1, X_2, \dots, X_n$ is distributed in the following algorithm:

1. Generate $U_i \sim Uniform(0, 1), i = 1(1)n$.
2. If $\frac{\sum_{k=0}^{j-1} a_k \frac{\Gamma(k+1)}{\theta^{k+1}}}{\sum_{k=0}^r a_k \frac{\Gamma(k+1)}{\theta^{k+1}}} < U_i \leq \frac{\sum_{k=0}^j a_k \frac{\Gamma(k+1)}{\theta^{k+1}}}{\sum_{k=0}^r a_k \frac{\Gamma(k+1)}{\theta^{k+1}}}$, $j = 2, \dots, r$, then set $X_i = V_i$,
where $V_i \sim NB(j+1, \theta)$ and if $U_i \leq \frac{a_0 \frac{1}{\theta}}{\sum_{k=0}^r a_k \frac{\Gamma(k+1)}{\theta^{k+1}}}$, then set $X_i = V_i$, where $V_i \sim geo(\theta)$.

A simulation study is carried out with 1,000(N) repetitions. Here we choose $a_0 = 1$, $a_1 = 1$, $\theta = 0.01$, $x = 2$ and $r = 1$. We compute MSE of the MLE and the UMVUE of the PMF and the CDF of a NDL distribution. From Figure 2, it is clear that MSE decreases with an increasing sample size that shows the consistency property of the estimators.

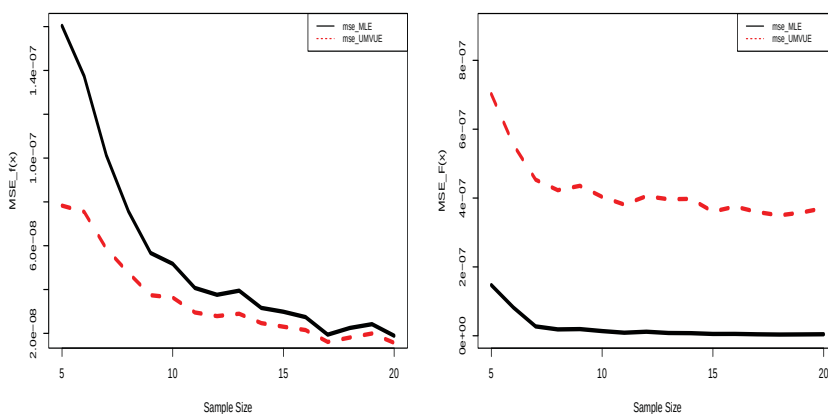


Figure 2: Graph of simulated MSE of the MLE and UMVUE of the PMF and the CDF of a NDL distribution for $x = 2$, $\theta = 0.01$ and $r = 1$.

5 Data Analysis

We have studied three data sets. The first dataset of 123 observations in Table 1 is the number of fires in forests in districts of Greece from period 1 July 1998 to 31 August 1998 and are available in (Bakouch, Jazi and Nadarajah, 2014). The second data set of 47 observations in Table 2 represents the numbers of daily deaths in Egypt due to COVID-19 infections from 8 March to 30 April, 2020 and are available in worldometer website through <https://www.worldometers.info/coronavirus/country/egypt/>. The third data set in Table 3 consists of remission times in weeks for 20 leukemia patients randomly assigned to a certain treatment and are available in (Lawless, 1982).

Table 1: Numbers of fires in Greece

2	4	4	3	3	1	2	4	3	1	1	0	5	5	0	3	1	1	0
1	0	2	0	1	2	0	0	0	0	0	0	0	0	0	1	4	2	2
1	2	1	2	0	2	2	1	0	3	2	1	2	2	7	3	5	2	5
4	5	6	5	4	3	8	43	8	4	4	3	10	5	4	5	12	3	8
12	10	11	6	1	8	9	12	9	4	8	12	11	8	6	4	7	9	15
12	15	15	12	9	16	7	11	9	11	6	5	20	9	8	8	5	7	10
6	6	5	5	15	6	8	5	6										

Table 2: Numbers of daily deaths in Egypt due to COVID-19 infection

1	1	2	2	1	1	2	4	5	1	1	3	6	6	4	1	5	6	6	8
5	7	7	9	9	15	17	11	13	5	14	5	13	9	19	15	11	14	12	11
7	13	10	20	22	21	12													

Table 3: Remission times in weeks for 20 leukemia patients

1	3	3	6	7	7	10	12	14	15
18	19	22	26	28	29	34	40	48	49

Graphs of the estimated PMF and the CDF, fitted to data sets 1, 2 and 3, have been shown in Figures 3-8 respectively.

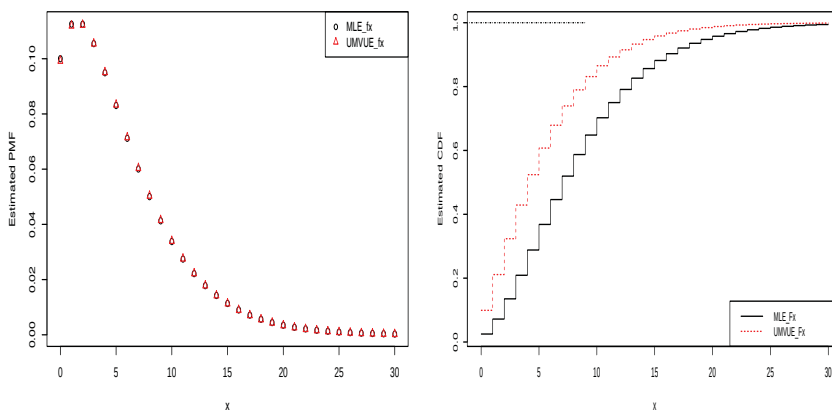


Figure 3: Graph of the estimated PMF and the CDF of a NDL distribution fitted to data set 1.

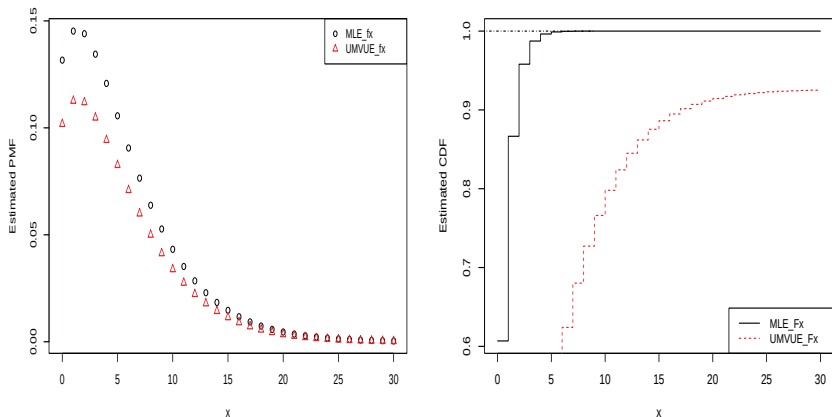


Figure 4: Graph of the estimated PMF and the CDF of a NDOPPE distribution ($a_0 = 1.2$, $a_1 = 1.05$) fitted to data set 1.

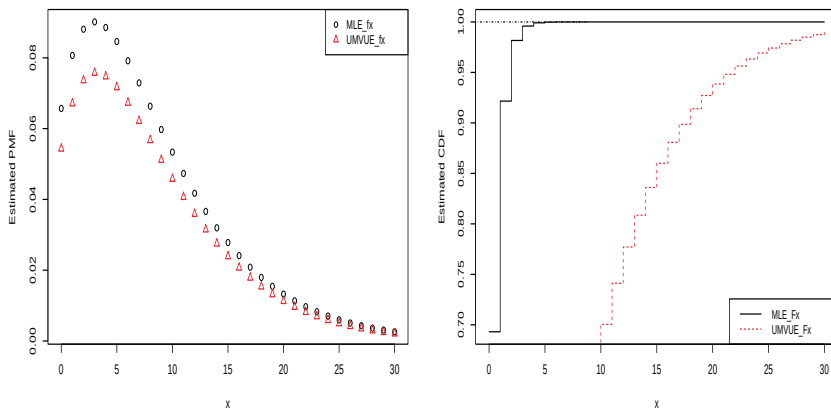


Figure 5: Graph of the estimated PMF and the CDF of a NDL distribution fitted to data set 2.

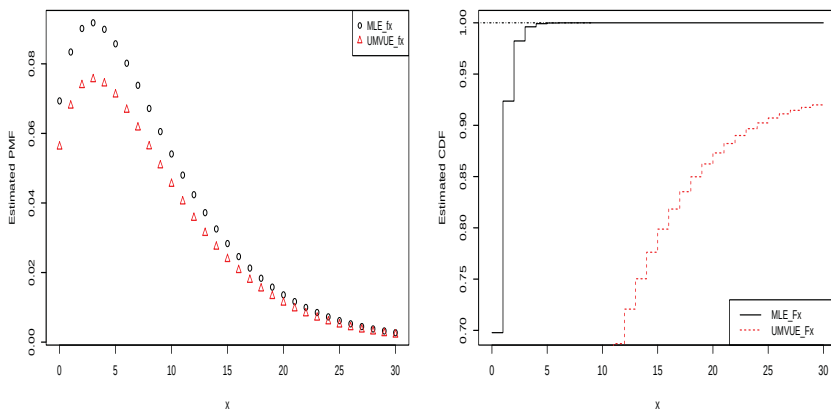


Figure 6: Graph of the estimated PMF and the CDF of a NDOPPE distribution ($a_0 = 1.2$, $a_1 = 1.05$) fitted to data set 2.

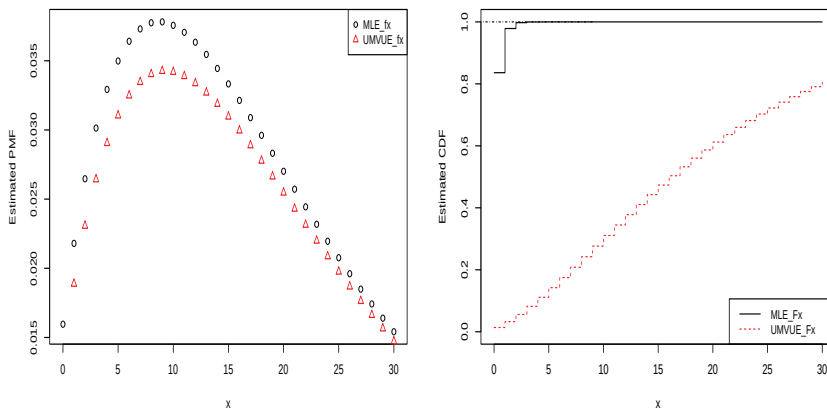
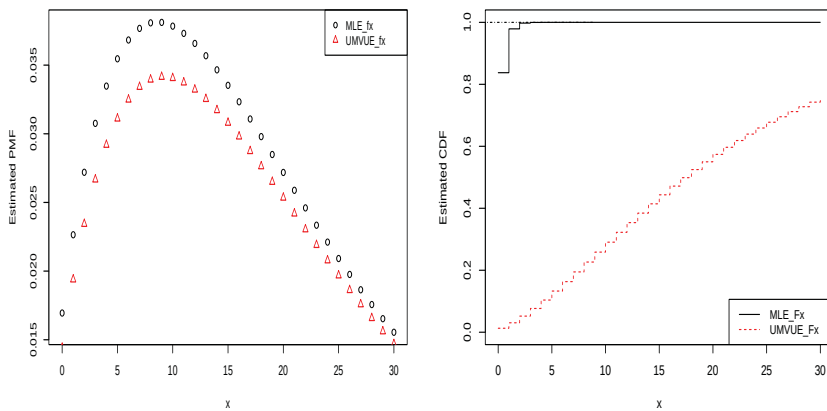


Figure 7: Graph of the estimated PMF and the CDF of a NDL distribution fitted to data set 3.


 Figure 8: Graph of the estimated PMF and the CDF of a NDOPPE distribution ($a_0 = 1.2$, $a_1 = 1.05$) fitted to data set 3.

We use the estimate of the negative log-likelihood values for the model selection criterion. Lower the value of negative log-likelihood indicates better the model fit. From Tables 4, 5 and 6, we found that MLE is better than UMVUE in a negative log-likelihood sense.

Table 4: Model selection criterion for data set 1

Estimators	Negative log-likelihood	
	NDL distribution	NDOPPE distribution ($a_0 = 1.2$, $a_1 = 1.05$)
MLE	340.0195	309.2184
UMVUE	340.1765	340.045

Table 5: Model selection criterion for data set 2

Estimators	Negative log-likelihood	
	NDL distribution	NDOPPE distribution ($a_0 = 1.2, a_1 = 1.05$)
MLE	137.5668	136.686
UMVUE	145.2438	145.3162

Table 6: Model selection criterion for data set 3

Estimators	Negative log-likelihood	
	NDL distribution	NDOPPE distribution ($a_0 = 1.2, a_1 = 1.05$)
MLE	77.13243	76.90162
UMVUE	78.80478	78.79437

6 Concluding Remarks

In this paper, two methods of estimation of the PMF and the CDF of a NDOPPE distribution have been considered. The MLE and the UMVUE have been found out. Simulation study is performed to compare the performances of the proposed methods of estimation. We have studied the performance of the estimators of the PMF and the CDF of a NDL distribution as a particular case of a NDOPPE distribution. From simulation study results, it is found that UMVUE is better than MLE for the PMF and MLE is better than UMVUE for the CDF in MSE sense. From the model selection criterion, it is noticed that MLE is better than UMVUE in negative log-likelihood (i.e. AIC) sense. We analyze three data sets to illustrate the fit of the proposed distribution.

Acknowledgement

The authors would like to thank the referee for valuable comments and suggestions that have encouraged the authors to make significant improvements in the article. The first author acknowledges the Department of Science and Technology, India, (IF150324), for sponsoring the research work in term of providing Fellowship.

References

- Ahmed, A. N. and Afify, A. Z. 2019. A new discrete analog of the continuous Lindley distribution, with reliability applications, *Proceedings of 62nd ISI World Statistics Congress* **5**: 7–14.
- Asrabadi, B. R. 1990. Estimation in the Pareto distribution, *Metrika* **37**: 199–205.
- Bagheri, S. F., Alizadeh, M., Baloui, J. E. and Nadarajah, S. 2014. Evaluation and comparison of estimations in the generalized exponential-Poisson distribution, *Journal of Statistical Computation and Simulation* **84**(11): 2345–2360.

- Bakouch, H. S., Jazi, M. A. and Nadarajah, S. 2014. A new discrete distribution, *Statistics* **48**(1): 200–240.
- Dixit, U. J. and Jabbari, N. M. 2010. Efficient estimation in the Pareto distribution, *Statistical Methodology* **7**: 687–691.
- Dixit, U. J. and Jabbari, N. M. 2011. Efficient estimation in the Pareto distribution with the presence of outliers, *Statistical Methodology* **8**: 340–355.
- Famoye, F. 1993. Restricted generalized Poisson regression model, *Communications in Statistics-Theory and Methods* **22**(5): 1335–1354.
- Gomez-Deniz, E. 2010. Another generalization of the geometric distribution, *TEST* **19**: 399–415.
- Gomez-Deniz, E. and Calderin-Ojeda, E. 2011. The discrete Lindley distribution: properties and applications, *Journal of Statistical Computation and Simulation* **81**(11): 1405–1416.
- Jabbari, N. M. and Jabbari, N. H. 2010. Efficient estimation of PDF, CDF and r th moment for the exponentiated Pareto distribution in the presence of outliers, *Statistics: A Journal of Theoretical and Applied Statistics* **44**(4): 1–20.
- Jazi, M. A., Lai, C. D. and Alamatsaz, M. H. 2010. A discrete inverse Weibull distribution and estimation of its parameters, *Statistical Methodology* **7**(2): 121–132.
- Krishna, H. and Pundir, P. S. 2009. Discrete Burr and discrete Pareto distributions, *Statistical Methodology* **6**(2): 177–188.
- Lawless, J. F. 1982. *Statistical models and methods for lifetime data*, John Wiley & Sons, Inc., Hoboken, New Jersey.
- Maiti, S. S., Dey, M. and Sarkar (Mondal), S. 2018. Discrete xgamma distributions: properties, estimation and an application to the collective risk model, *Journal of Reliability and Statistical Studies* **11**(1): 117–132.
- Maiti, S. S. and Mukherjee, I. 2017. Estimation of the PMF and CDF of some standard discrete distributions useful in reliability modelling, *International Journal of Agricultural and Statistical Sciences* **13**(2): 735–751.
- Mukherjee, I. and Maiti, S. S. and Das, M. 2018. On estimation of the PMF and CDF of the logarithmic series distribution, *RASHI* **3**(2): 34–44.
- Nakagawa, T. and Osaki, S. 1975. Discrete Weibull distribution, *IEEE Transactions on Reliability* **24**(5): 300–301.
- Roy, D. 2002. Discretization of continuous distributions with an application to stress-strength reliability, *Calcutta Statistical Association Bulletin* **52**: 297–314.
- Roy, D. 2003. The discrete normal distribution, *Communications in Statistics-Theory and Methods* **32**(10): 1871–1883.

Roy, D. 2004. Discrete Rayleigh distribution, *IEEE Transaction in Reliability* **53**(2): 255–260.

Stein, W. E. and Dattero, R. 1984. A new discrete Weibull distribution, *IEEE Transactions on Reliability* **33**(2): 196–197.