On Some Aspects of Generalized Extended Yule Distribution: Properties and Applications

C. Satheesh Kumar* and S. Harisankar

Department of Statistics, University of Kerala, Thiruvananthapuram, India

Abstract: Martinez-Rodriguez (Comp. Statist. Dat. Anal., 2011) studied an extended version of the Yule distribution, namely "the extended Yule distribution (EYD)" which they obtained as a mixture of geometric distribution and generalized beta distribution. Through the present paper, we propose a generalized version of the EYD and named it "the generalized extended Yule distribution (GEYD)". Several statistical properties of the distribution are obtained, including probability generating function (p.g.f), moments, recursion formulae etc. The maximum likelihood estimation of the parameters of the GEYD is discussed and fitted to two real-life data sets for illustrating its usefulness compared to the existing models. Further, the generalized likelihood ratio test procedure is considered for testing the significance of the parameters of the GEYD.

Keywords: Probability generating function, Model selection, Maximum likelihood estimation, Generalized likelihood ratio test.

1. INTRODUCTION

Yule [1] considered a distribution through the following probability generating function (p.g.f), for $\rho > 0$.

$$G(t) = \frac{\rho+1}{\rho}\ {}_2F_1[1,1;\rho+2;\ t] \quad (1)$$

A distribution with p.g.f (1) was later known in the literature as "the Yule distribution"; hereafter, we denoted it as YD (ρ). Simon [2] considered the Yule distribution as a model in the sociological field (number of words in a text by their frequency of occurrence, number of scientists by number of papers published, distribution of cities by population), in the biological field (distribution of biological genera by number of species) and the economics field (distribution of incomes by size). Kendall [3] utilized YD for describing certain types of bibliographic data sets. Haight [4] used YD to model word frequency data, and Xekalaki [5] applied the YD (ρ) in an econometric context. Jones and Handcock [6,7] considered it as the underlying mechanism in the formation of social networks. It has also been studied by Dorogovtsev *et al.* [8] and Levene *et al.* [9] in the context of modelling the growth of the internet. Xekalaki and Panaretos [10] derived the YD as a discrete analog of the Pareto distribution. Singh and Vasudeva [11] characterized the exponential distribution via the Yule distribution.

In [12-16] studied certain modified versions of the Yule distribution and described some of their applications. Martinez-Rodriguez [17] considered an extended version of YD (ρ) through the name "the extended Yule distribution (EYD)", which possess the following p.g.f, for $\rho > -2$ and $0 < \lambda < 1$.

$$H(t) = \frac{\lambda t}{\lambda}\ {}_2F_1[1,1;\rho+2] \quad (2)$$

The EYD belongs to the family of distributions generated by the Gaussian hypergeometric function, and it can be expressed as a generalized beta mixture of a geometric distribution. The EYD has a similar genesis to YD, so it keeps the more relevant properties of the YD, but it also has a new parameter that allows control of the right tail of the distribution, and the effect of infinite variance is not possible. The main objective of the present paper is to develop a generalized version of EYD so as to make its modeling suitable for more complex data sets having heavy tails. The proposed class of distribution was termed "the generalized extended Yule distribution (GEYD)" and it has been obtained through compounding generalized geometric distribution with the generalized beta distribution. Further, it can be noted that the GEYD is over-dispersed (variance more significant than the mean) for $\rho \geq -1$ and under-dispersed (variance smaller than the mean) if $-2 < \rho < -1$. This indicates the utility of the model to both over-dispersed and under-dispersed data sets. The GEYD is fitted to two real-life data sets and observed that the GEYD allows better fits compared to the other related generalized version of YD available in the literature.

The rest of the paper is organized as follows. In Section 2, we present a genesis of the GEYD and derive its important properties such as its p.g.f., expressions for its mean and variance, recursion formulae for its probabilities, raw moments, and

*Address correspondence to this author at the Department of Statistics, University of Kerala, Thiruvananthapuram, India;
Email: drcsatheeshkumar@gmail.com

E-ISSN: 2409-5761/20

factorial moments. In Section 3, we discuss the estimation of the parameters of the GEYD by the method of maximum likelihood. In Section 4, the GEYD has been fitted to three real-life data sets for establishing the importance of the proposed distribution, and in section 5, we consider the generalized likelihood ratio test procedure for testing the significance of the parameters of the GEYD. In Section 6, we carried out a simulation study to examine the performance of the maximum likelihood estimators of the parameters of the distribution.

Throughout this paper, we assumed m as a positive integer and adopted the following shorter notation.

$$\Omega_0^{-1} = {}_2F_1[1,1,\rho+2;\lambda_1+\lambda_2], \tag{3}$$

where ${}_2F_1[.]$ is the Gaussian hypergeometric function. For more details regarding Gaussian hypergeometric distribution, see Mathai and Haubold [18]. Further, we need the following series representations in the sequel.

$$\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}A_1(s,r)=\sum_{r=0}^{\infty}\sum_{s=0}^{r}A_1(s,r-s) \tag{4}$$

and

$$\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}A_2(s,r)=\sum_{r=0}^{\infty}\sum_{s=0}^{[\frac{r}{m}]}A_2(s,r-ms) \tag{5}$$

2. A GENESIS OF GEYD AND ITS PROPERTIES

Let X be a generalized geometric random variable with the following p.g.f, in which $\theta>0$, $\lambda_1>0$ and $\lambda_2\geq 0$.

$$G(t)=\frac{{}_1F_0[1;;\theta(\lambda_1\,t+\lambda_2\,t^m)]}{{}_1F_0[1;;\theta(\lambda_1+\lambda_2)]}$$

Assume that the parameter θ follows a generalized beta distribution with parameters ρ, λ_1 and λ_2, with probability density function (p.d.f)

$$f(\theta)=\frac{\Omega_0\,(1-\theta)^{\rho+1}\,{}_1F_0[1;;\theta(\lambda_1+\lambda_2)]}{B(1,\rho+2)},$$

Where B (.,.) is the beta function and $\rho>-2$. Then the unconditional distribution of X is obtained as

$$Q(t)=\frac{\Omega_0}{B(1,\rho+2)}\int_0^1(1-\theta)^{\rho+1}\;{}_1F_0[1;;\theta(\lambda_1\,t+\lambda_2\,t^m)]d\theta \tag{6}$$

$$=\Omega_0\;{}_2F_1[1,1,\rho+2;\lambda_1\,t+\lambda_2\,t^m],$$

in the light of the identity (1.104) of Johnson *et al.* [19].

Now we present the following definition of the proposed class of distribution.

Definition 2.1 A non-negative integer-valued random variable X is said to follow "the generalized extended Yule distribution (GEYD)" if its p.g.f is of the following form, in which $\rho>-2$, $\lambda_1>0$ and $\lambda_2\geq 0$.

$$H(t) = \Omega_0\;{}_2F_1[1,1;\rho+2;\lambda_1 t+\lambda_2 t^m] \tag{7}$$

Clearly, several well-known models are special cases of the GEYD. Some of them are listed below.

1) When $\lambda_2=0$, the p.g.f (7) reduces to the p.g.f of the GYD as given in (2)

2) When ρ = -1, the p.g.f (7) reduces to the p.g.f of the generalized version of a geometric distribution (GGD) with parameters λ_1 and λ_2, which further reduces to the geometric distribution when λ_2=0.

3) When ρ =0, the p.g.f (7) reduces to the p.g.f of the extended zero-inflated logarithmic series distribution (EZILSD) studied by Kumar and Riyaz [20] which further reduces to the zero-inflated logarithmic series distribution of Kumar and Riyaz [21] when λ_2=0.

4) When $\lambda_2=0$ and λ_1 approaches to 1, the p.g.f (7) reduces to the pgf (1) of the YD (ρ).

Now we obtain the p.m.f of the GEYD through the following result.

Proposition 2.2 The p.m.f h_x of GEYD with p.g.f (7) is the following, for $x=0,1,2\cdots$, $\rho>-2$, $\lambda_1>0$ and $\lambda_2\geq 0$ with $\lambda_1+\lambda_2<1$.

$$h_x=\Omega_0\sum_{n=0}^{[\frac{x}{m}]}\frac{[(x-(m-1)n)!]^2}{(\rho+2)_{x-(m-1)n}}\frac{\lambda_1^{x-mn}\lambda_2^n}{(x-mn)!\,n!}, \tag{8}$$

where Ω_0 is as defined in (3).

Proof. From (7), we have the following:

$$H(t)=\Omega_0\;{}_2F_1[1,1;\rho+2;\lambda_1\,t+\lambda_2\,t^m] \tag{9}$$

$$=\sum_{x=0}^{\infty}h_x\,t^x \tag{10}$$

On expanding the gauss hypergeometric function in (9), we get

$$H(t)=\Omega_0\sum_{x=0}^{\infty}\frac{x!}{(\rho+2)_x}[\lambda_1\,t+\lambda_2\,t^m]^x \tag{11}$$

By applying binomial theorem in (11) to get the following.

$$H(t)=\Omega_0\sum_{x=0}^{\infty}\frac{x!}{(\rho+2)_x}\sum_{n=0}^{x}\binom{x}{n}(\lambda_1 t)^{x-n}(\lambda_2 t^m)^n$$
$$=\Omega_0\sum_{x=0}^{\infty}\sum_{n=0}^{\infty}\frac{(x+n)!}{(\rho+2)_{x+n}}\binom{x+n}{n}(\lambda_1 t)^x(\lambda_2 t^m)^n, \qquad (12)$$

in the light of (4). Now, applying (5) in (12), we obtain

$$H(t)=\Omega_0\sum_{x=0}^{\infty}\sum_{n=0}^{[\frac{x}{m}]}\frac{[x-(m-1)n]![x-(m-1)n]!}{(\rho+2)_{x-(m-1)n}\,(x-mn)!\,n!}\lambda_1^{x-mn}\lambda_2^n t^x. \qquad (13)$$

Equating the coefficients of t^x on the right-hand side expressions of (10) and (13), we get (8)

Proposition 2.3 The characteristic function $\psi(t)$ of the GEYD is the following, for any $t\in R$ and $i=\sqrt{-1}$.

$$\psi(t)=\Omega_0\ {}_2F_1[1,1;\rho+2;\lambda_1 e^{it}+\lambda_2 e^{mit}] \qquad (14)$$

Proposition 2.4 The mean and variance of the GEYD are the following, in which $\lambda^*=\lambda_1+m\lambda_2$

$$Mean=\frac{\lambda^*\Omega_0}{(\rho+2)}\Omega_1 \qquad (15)$$

and

$$Variance=\frac{\lambda^*\Omega_0}{\rho+2}\left[\frac{4\lambda^*\Omega_2}{\rho+3}+\Omega_1\left(1-\frac{\lambda^*\Omega_0\Omega_1}{\rho+2}\right)\right]. \qquad (16)$$

Proof. It follows from the fact that

Mean= $H^{(1)}(1)$

and

Variance= $H^{(2)}(1)+H^{(1)}(1)-[H^{(1)}(1)]^2$,

where

$H^{(r)}(1)=\frac{d^r H(t)}{dt^r}/t=1.$

Remark 2.5 From (15) and (16), it is seen that the GEYD is over-dispersed if and only if

$4(\rho+2)\Omega_2-(\rho+3)\Omega_1\Omega_0>0$,

for all values of the parameters ρ,λ_1 and λ_2 and the GEYD is under-dispersed otherwise.

Proposition 2.6 For $x\geq 1$, the following is a simple recursion formula for probabilities $h_x=h_x(1,1;\rho+2)$ of the GEYD with p.g.f (7).

$$\Omega_1(\rho+2)(x+1)h_{x+1}(1,1;\rho+2)=$$
$$\Omega_0[\lambda_1 h_x(2,2;\rho+3)+m\lambda_2 h_{x-m+1}(2,2;\rho+3)] \qquad (17)$$

These recurrence relations are helpful for computing probabilities of the GEYD while fitting the distribution to the data.

Proof. From (7), we have

$$H(t)=\sum_{x=0}^{\infty}h_x(1,1;\rho+2)t^x=\Omega_0\ {}_2F_1[1,1;\rho+2;\lambda_1 t+\lambda_2 t^m] \qquad (18)$$

Differentiating the equation (18) with respect to t, we get

$$\sum_{x=0}^{\infty}(x+1)h_{x+1}(1,1;\rho+2)t^x=\frac{\Omega_0(\lambda_1+m\lambda_2 t^{m-1})}{\rho+2}\ {}_2F_1[2,2;\rho+3;\lambda_1 t+\lambda_2 t^m]. \qquad (19)$$

In (18), by replacing $1,1$ and $\rho+2$ with $2,2$ and $\rho+3$, respectively, we obtain

$$\Omega_1\ {}_2F_1[2,2;\rho+3;\lambda_1 t+\lambda_2 t^m]=\sum_{x=0}^{\infty}h_x(2,2;\rho+3)t^x. \qquad (20)$$

Substitute (20) in (19) to get

$$\sum_{x=0}^{\infty}(x+1)h_{x+1}(1,1;\rho+2)t^x=\frac{\Omega_0(\lambda_1+m\lambda_2 t^{m-1})}{\Omega_1(\rho+2)}\sum_{x=0}^{\infty}h_x(2,2;\rho+3)t^x$$
$$=\frac{\Omega_0}{\Omega_1(\rho+2)}\left(\lambda_1\sum_{x=0}^{\infty}h_x(2,2;\rho+3)t^x+m\lambda_2\sum_{x=0}^{\infty}h_x(2,2;\rho+3)t^{x+m-1}\right). \qquad (21)$$

Equating the coefficients of t^x on both sides of (21), we get (17).

Now we derive certain recurrence relations for raw moments and factorial moments of the GEYD. Those recurrence relations are useful for evaluating the moments of the distribution of any order.

Proposition 2.7 The following is a simple recursion formula for raw moments

$$\mu_{r+1}(1,1;\rho+2)=\frac{\Omega_0}{\rho+2}\sum_{s=0}^{r}\binom{r}{s}(\lambda_1+m^{s+1}\lambda_2)\mu_{r-s}(2,2;\rho+3) \qquad (22)$$

Proof. By definition, the characteristic function of the GEYD is given by

$$\psi(t) = \sum_{r=0}^{\infty} \mu_r(1,1;\ \rho+2)\,\frac{(it)^r}{r!} \tag{23}$$

$$= \Omega_0\ {}_2F_1[1,1;\rho+2;\lambda_1\, e^{it} + \lambda_2\, e^{mit}]$$

By using (14) with $1,1$ and $\rho+2$ replaced by $2,2$ and $\rho+3$ respectively, we obtain

$$\Omega_1\ {}_2F_1[2,2;\ \rho+3;\lambda_1\, e^{it} + \lambda_2\, e^{mit}] = \sum_{r=0}^{\infty} \mu_r(2,2;\ \rho+3)\,\frac{(it)^r}{r!}. \tag{24}$$

Differentiate (23) with respect to t to get

$$\sum_{r=0}^{\infty} i\ \mu_{r+1}(1,1;\rho+2)\,\frac{(it)^r}{r!} = \Omega_0\ \frac{i\,(\lambda_1\, e^{it} + m\,\lambda_2\, e^{mit})}{\rho+2}$$
$$\times\ {}_2F_1[2,2;\rho+3;\lambda_1\, e^{it} + \lambda_2\, e^{mit}], \tag{25}$$

which on simplification in the light of (25) gives

$$(\rho+2)\sum_{r=0}^{\infty} \mu_{r+1}(1,1;\ \rho+2)\,\frac{(it)^r}{r!} = \Omega_0\,[(\lambda_1\, e^{it} + m\,\lambda_2\, e^{mit})$$
$$\times \sum_{r=0}^{\infty} \mu_r(2,2;\ \rho+3)\,\frac{(it)^r}{r!}]. \tag{26}$$

On expanding the exponential functions in (26) and applying (4) to obtain

$$(\rho+2)\sum_{r=0}^{\infty} \mu_{r+1}(1,1;\ \rho+2)\,\frac{(it)^r}{r!} = \Omega_0\left[\lambda_1 \sum_{r=0}^{\infty}\sum_{s=0}^{r} \mu_{r-s}(2,2;\ \rho+3)\,\frac{(it)^r}{(r-s)!\,s!}\right.$$
$$\left. + \lambda_2 \sum_{r=0}^{\infty}\sum_{s=0}^{r} m^{s+1}\,\mu_{r-s}(2,2;\ \rho+3)\,\frac{(it)^r}{(r-s)!\,s!}\right]. \tag{27}$$

Equating the coefficients of $(it)^r\,(r!)^{-1}$ on both sides of (27), we get (22)

Proposition 2.8 The following is a simple recursion formula for factorial moments $\mu_{[r]} = \mu_{[r]}(1,1;\rho+2)$ of the GEYD, for $r \geq 0$.

$$(\rho+2)\mu_{[r+1]}(1,1;\ \rho+2) = \Omega_0\ \Omega_1^{-1}\left[\lambda_1\ \mu_{[r]}(2,2;\ \rho+3) + \right.$$
$$\left. \times\, m\,\lambda_2 \sum_{x=0}^{m-1}\binom{m-1}{x}\mu_{[r-x]}(2,2;\ \rho+3)\right]$$

Proof. The factorial moment generating function F(t) of the GEYD with p.g.f (7) is given by

$$F(t) = H(1+t) \tag{28}$$

$$= \sum_{r=0}^{\infty} \mu_{[r]}\,\frac{t^r}{r!} \tag{29}$$

$$= \Omega_0\ {}_2F_1[1,1;\rho+2;\lambda_1\,(t+1) + \lambda_2(t+1)^m]. \tag{30}$$

From (28) with $1,1$ and $\rho+2$ changed by $2,2$ and $\rho+3$ respectively, we have

$$\Omega_1\ {}_2F_1[2,2;\ \rho+3;\lambda_1\,(t+1) + \lambda_2(t+1)^m] = \sum_{r=0}^{\infty} \mu_r(2,2;\ \rho+3)\,\frac{t^r}{r!}. \tag{31}$$

On differentiating (28) with respect to t, we get

$$\sum_{r=0}^{\infty} \mu_{[r+1]}(1,1;\ \rho+2)\,\frac{t^r}{r!} = \frac{\Omega_0\,[\lambda_1 + m\lambda_2\,(t+1)^{m-1}]}{\rho+2}$$
$$\times\ {}_2F_1[2,2;\rho+3;\lambda_1\,(t+1) + \lambda_2(t+1)^m],$$

Applying binomial theorem and the series representation (5) in (31) to obtain

$$(\rho+2)\sum_{r=0}^{\infty} \mu_{[r+1]}(1,1;\ \rho+2)\,\frac{t^r}{r!} = \Omega_0\ \Omega_1^{-1}\,[\lambda_1 \sum_{r=0}^{\infty} \mu_{[r]}(2,2;\ \rho+3)\,\frac{t^r}{r!} +$$
$$\times\, m\,\lambda_2 \sum_{r=0}^{\infty}\sum_{x=0}^{m-1}\binom{m-1}{x}\mu_{[r]}(2,2;\ \rho+3)\,\frac{t^{x+r}}{r!}]$$

By equating the coefficients of $t^r\,(r!)^{-1}$ on above equation, we get the results.

3. ESTIMATION

In this section, we discuss the estimation of the parameters ρ, λ_1 and λ_2 of the GEYD by the method of maximum likelihood, and thereafter the generalized likelihood ratio test procedure is utilized for testing the significance of the parameters ρ, λ_1 and λ_2 of the GEYD.

Let $a(x)$ be the observed frequency of x events based on the observations from a sample with independent components and let y be the highest value of the x observed. The likelihood function of the sample is

$$L = \prod_{x=0}^{y}\ [h_x]^{a(x)}, \tag{32}$$

which implies

$$\ln L = \sum_{x=0}^{y}\ a(x)\,\ln h_x. \tag{33}$$

Let $\hat{\rho}$, $\hat{\lambda}_1$ and $\hat{\lambda}_2$ be the MLEs of ρ, λ_1 and λ_2, respectively. Now, the MLEs of the parameters are obtained by solving the following likelihood equations obtained from (33) on differentiation with respect to ρ, λ_1 and λ_2 respectively and equating to zero. Then

$$\frac{\partial \log L}{\partial \rho} = 0 \tag{34}$$

or equivalently

$$\sum_{x=0}^{y} a(x)\left[\sum_{j=0}^{[\frac{x}{m}]}\left[\frac{1}{\varepsilon(x;\lambda_1,\lambda_2)}\frac{\binom{x-j}{j}(1)_{x-j}\,\lambda_1^{x-mj}\,\lambda_2^{j}}{(\rho+2)_{x-j}}[\upsilon(\rho+2+x-j)-\upsilon(\rho+2)]\right.\right.$$
$$\left.\left.+\Omega_0\sum_{r=0}^{\infty}\frac{(1)_r\,(\lambda_1+\lambda_2)^r}{(\rho+2)_r}[\upsilon(\rho+2)-\upsilon(\rho+r+2)]\right]\right]=0, \tag{35}$$

$$\frac{\partial \log L}{\partial \lambda_1} = 0$$

or equivalently

$$\sum_{x=0}^{y} a(x) - \Omega_0 \sum_{r=0}^{\infty}\frac{(1)_r\, r\,(\lambda_1+\lambda_2)^{r-1}}{(\rho+2)_r} + \sum_{j=0}^{[\frac{x}{m}]}\left[\frac{1}{\varepsilon(x;\lambda_1,\lambda_2)}\frac{\binom{x-j}{j}(1)_{x-j}\,(x-mj)\,\lambda_1^{x-mj-1}\,\lambda_2^{j}}{(\rho+2)_{x-j}}\right]=0,$$

and

$$\frac{\partial \log L}{\partial \lambda_2} = 0 \tag{36}$$

or equivalently

$$\sum_{x=0}^{y} a(x) - \Omega_0 \sum_{r=0}^{\infty}\frac{(1)_r\, r\,(\lambda_1+\lambda_2)^{r-1}}{(\rho+2)_r} + \sum_{j=0}^{[\frac{x}{m}]}\left[\frac{1}{\varepsilon(x;\lambda_1,\lambda_2)}\frac{\binom{x-j}{j}(1)_{x-j}\,\lambda_1^{x-mj}\,j\,\lambda_2^{j-1}}{(\rho+2)_{x-j}}\right]=0,$$

where

$$\psi(\rho) = [\Gamma(\rho)]^{-1}\frac{d\,\Gamma(\rho)}{d\rho},$$

$$\gamma(x,j) = a(x)\frac{(x-j)!\,(1)_{x-j}\,\lambda_1^{x-mj-1}\,\lambda_2^{j-1}}{g(x;\rho^*)\,(x-mj)!\,j!\,(\rho+2)_{x-j}},$$

$$\upsilon(\rho, x-j) = \psi(\rho) - \psi(\rho+x-j)$$

and

$$\xi(\rho^*) = \lambda_1\,\lambda_2\,\frac{(1)_{x-j}\,(1)_{x-j}}{(\rho+2)_{x-j}}\;\Omega_{x-j}(\lambda_1+\lambda_2).$$

On solving the log-likelihood equations by using some mathematical software, say MATHEMATICA, one can obtain the maximum likelihood estimators of the parameters ρ, λ_1 and λ_2 of the GEYD.

4. APPLICATIONS

For numerical illustration, we have considered two real-life data applications, of which the first data set is from Wagner *et al.* [22], which contains the frequency of direct job changes in a sample of 1962 individuals. The second data set is on the number of European red mites on each leave based on an experiment with 150 leaves from apple trees taken from Bliss *et al.* [23]. We have fitted the EYD, the EZILSD, the geometric distribution (GD), the YD, and the GEYD to these data sets and the results obtained along with the corresponding values of the expected frequencies, Chi-square statistic, degrees of freedom (d.f), Akaike information criterion (AIC) and Bayesian information criterion (BIC) in respect of each of the models are presented in Table **1** and **2**, respectively. Based on the computed values of the Chi-square statistic and information measures, it can be observed that the GEYD gives a better fit to both the data sets considered here compared to the existing models, the EYD, the GD, the EZILSD and the YD.

5. TESTING OF HYPOTHESIS

In this section, we present the generalized likelihood ratio test (GLRT) procedure for testing the significance of the parameters of the GEYD. Here we consider the following tests:

1. Test 1: ${H_0}^{(1)}: \lambda_2 = 0$ against ${H_1}^{(1)}: \lambda_2 \neq 0$

2. Test 2: ${H_0}^{(2)}: \rho = 0$ against ${H_1}^{(2)}: \rho \neq 0$

3. Test 3: ${H_0}^{(3)}: \rho = -1, \lambda_2 = 0$ against ${H_1}^{(3)}: \rho \neq -1, \lambda_2 \neq 0$

The test statistic is

$$-2\,ln\,\Lambda = 2\left(ln\,L\,(\hat{\underline{\Lambda}}; x) - ln\,L\,(\hat{\underline{\Lambda}}^*; x)\right), \tag{37}$$

in which $\hat{\Lambda}$ is the MLE of $\underline{\Lambda} = (\lambda_1, \lambda_2, \rho)$ with no restriction and $\hat{\underline{\Lambda}}^*$ is the MLE of $\underline{\Lambda}$ under H_0. The test statistic $-2\,log\,\Lambda$ is asymptotically distributed as a chi-square with one degree of freedom in the case of Test 1 and 2 and two degrees of freedom in the case of Test 3, respectively. For details of the GLRT, see [24]. We have computed the values of $ln\,L\,(\hat{\underline{\Lambda}}; x)$, $ln\,L\,(\hat{\underline{\Lambda}}^*; x)$ and the test statistic in case of both the data sets are inserted in Table **3**.

From Table **3**, it can be observed that the calculated value of the test statistic is greater than the tabled value in the case of both the data sets, and hence one can conclude that the parameters of the fitted model GEYD are significant in the case of both the data sets at 5 level of significance.

Table 1: Observed Frequencies and Computed Values of Expected Frequencies of the EYD, the GD, the EZILSD, the YD and the GEYD by the Method of Maximum Likelihood for the First Data Set

* x	Observed	Expected Frequency by MLE				
		EYD	GD	EZILSD	YD	GEYD m=2
0	1333	1339.72	1273.36	1351.53	1424.47	1328.02
1	404	397.28	446.35	371.63	312.50	388.14
2	133	139.45	156.22	136.26	108.43	140.63
3	43	52.14	54.67	56.21	48.91	48.79
4	25	20.15	19.13	24.73	25.57	29.41
5	10	7.95	6.69	11.33	14.78	12.60
6	4	3.18	3.34	5.34	9.19	6.42
7	4	1.28	1.82	2.65	6.04	3.86
8	1	0.52	0.28	1.25	4.14	2.24
9	2	0.21	0.10	0.62	2.95	1.40
10	2	0.08	0.03	0.31	2.16	0.59
11	0	0.03	0.01	0.15	1.62	0.32
12	1	0.01	0.001	0.07	1.26	0.18
Total	1962	1962	1962	1962	1962	1962
d.f		4	5	5	8	4
Estimates of		ρ =-0.95	λ =1.65	λ_1=0.43	ρ =0.06	ρ =-0.98
parameters		λ =0.34		λ_2=0.10		λ_1=0.75
						λ_2=0.36
χ^2-value		17.96	28.89	11.51	47.29	3.66
AIC		3890.12	3915.12	3887.38	3928	3882.32
BIC		3888.33	3914.23	3885.60	3927.11	3879.65

Table 2: Observed Frequencies and Computed Values of Expected Frequencies of the EYD, the GD, the EZILSD, the YD and the GEYD by the method of Maximum Likelihood for the Second Data Set

* x	Observed	Expected Frequency by MLE				
		EYD	GD	EZILSD	YD	GEYD m=2
0	70	84	68.6	82.91	90	70.56
1	38	35.42	37.5	31.46	25.83	37.47
2	17	15.98	22.8	15.05	13.42	19.70
3	10	7.38	10.49	8.36	8.23	10.44
4	9	3.45	5.26	5.15	5.83	6.71
5	3	1.62	2.54	3.06	2.55	2.87
6	2	0.76	1.56	1.94	1.80	1.29
7	1	0.36	0.82	1.25	1.33	0.64
8	0	0.17	0.43	0.82	1.01	0.32
Total	150	150	150	150	150	150
d.f		2	4	3	4	2
Estimates of		ρ =1.02	λ =0.30	λ_1=0.46	ρ =-0.60	ρ =1.77
parameters		$\lambda = 0.61$		λ_2=0.41		λ_1=0.42
						λ_2=0.34
χ^2-value		3.63	3.80	6.70	13.30	1.29
AIC		456.84	453.68	458.66	476.24	450.08
BIC		453.64	452.58	454.46	475.14	446.78

Table 3: Test Statistic Value and Chi-Square Value for GLRT in Case of Both the Data Sets

Data sets		$\log L(\hat{\Lambda}; x)$	$\log L(\hat{\Lambda}^*; x)$	Test Statistic	d f	Chi square value
Data set 1	1*Test 1	-1938.16	-1943.06	5.90	1	3.84
	1*Test 2	-1938.16	-1941.69	7.06	2	5.99
	1*Test 3	-1938.16	-1956.96	37.60	2	5.99
Data set 2	1*Test 1	-222.04	-226.42	8.76	1	3.84
	1*Test 2	-222.04	-226.33	8.58	2	5.99
	1*Test 3	-222.04	-225.84	7.60	2	5.99

6. CONCLUSION

A new class of distribution namely (the GEYD) is introduced in this paper by mixing the generalized geometric distribution with the generalized beta distribution. The GEYD includes several well-known classes of distributions such as "extended Yule distribution of Martinez-Rodriguez *et al.* [17], generalized geometric distribution, geometric distribution, extended zero-inflated logarithmic series distribution of Kumar and Riyaz [21], zero-inflated logarithmic series distribution and Yule distribution". We studied some of the important properties of the GEYD and shown that it is suitable for modeling both over-dispersed and under-dispersed data sets. We attempted/discussed the estimation of the parameter by the method of MLE and suggested GLRT for testing the significance of the parameters of the model. Several characteristic properties and inferential aspects of the model are yet to study, which we hope to publish shortly via another research article. We hope that the GEYD may attract wider applications in analyzing count data models.

ACKNOWLEDGEMENT

The authors are highly thankful to the Editor-in-Chief and both anonymous referees for several fruitful suggestions, which helped to improve the presentation of the article.

REFERENCES

[1] Yule GU. A mathematical theory of evolution based on the conclusion of Dr. J. C. Wills. Philosophical Society of Royal Society of London, Series B, 1920; 213: 21-87.

[2] Simon HA. On a class of skew distribution. Biometrika, 1955; 42: 425-440. https://doi.org/10.1093/biomet/42.3-4.425

[3] Kendall MG. Natural law in social sciences. Journal of Royal Statistical Society, Series A, 1961; 124: 1-28. https://doi.org/10.2307/2984139

[4] Haight FA. Some statistical problems in connection with word association data. Journal of Mathematical Psychology 1966; 3(1): 217-233. https://doi.org/10.1016/0022-2496(66)90013-7

[5] Xekalaki E. A property of Yule distribution and its applications. Communication in Statistics: Theory and Methods, 1983; 12: 140-150. https://doi.org/10.1080/03610928308828523

[6] Jones JH. and Handcock MS. An assessment of preferential attachment as a mechanism for human sexual network formation. Proceedings of the Royal Society of London B: Biological Sciences, 2003; 270(1520): 1123-1128. https://doi.org/10.1098/rspb.2003.2369

[7] Handcock MS. and Jones JH. Likelihood-based inference for stochastic models of sexual network formation. Theoretical Population Biology, 2004; 65: 413-422. https://doi.org/10.1016/j.tpb.2003.09.006

[8] Dorogovtsev SN, Mendes JFF. and Samukhin AN. Structure of growing networks with preferential linking. Physical Review Letters, 2000; 85(21): 4633-4636. https://doi.org/10.1103/physrevlett.85.4633

[9] Levene M, Fenner T, Loizou G. and Wheeldon R. A stochastic model for the evolution of the web. Computer Networks, 2002; 39(3): 277-287. https://doi.org/10.1016/s1389-1286(02)00209-8

[10] Xekalaki E. and Panaretos J. On the association of the pareto and the yule distribution. Theory of Probability Its Applications, 1989; 33(1): 191-195. https://doi.org/10.1137/1133028

[11] Singh H. and Vasudeva H. A characterization of exponential distribution by yule distribution. Journal of Indian Statistical Assosiation, 1984; 22: 93-96.

[12] Prasad A. A new discrete distribution. Sankhya: The Indian Journal of Statistics, 1957; 17(4): 353-354.

[13] Irwin JO. The place of mathematics in medical and biological statistic. Journal of Royal Statistical Society, 1968; 126: 1-41. https://doi.org/10.2307/2982445

[14] Mishra A. On a generalized Yule distribution. Assam Statistical Review, 2009; 23: 140-150.

[15] Kumar CS. and Harisankar S. On a modified Yule distribution. Statistica, 2018; 78(2): 169-180.

[16] Kumar CS. and Harisankar S. On some aspects of a general class of Yule distribution and its applications. Communication in Statistics - Theory and Methods, 2019; 49(12): 2887-2897. https://doi.org/10.1080/03610926.2019.1584308

[17] Martinez-Rodriguez AM, Saez-Castillo AJ. and Conde-Sanchez A. Modelling using an extended Yule distribution. Computational Statistics and Data Analysis, 2011; 55(1): 863-873. https://doi.org/10.1016/j.csda.2010.07.014

[18] Mathai AM. and Haubold HJ. Special Functions for Applied Statistics. Springer, New York. 2008. https://doi.org/10.1007/978-0-387-75894-7

[19] Johnson NL, Kemp AW. and Kotz S. Univariate Discrete Distributions (III). John Wiley Sons, New York. 2005; https://doi.org/10.1002/0471715816

[20] Kumar CS. and Riyaz A. An extended zero-inflated logarithmic series distribution and its application. Journal of Applied Statistical Science, 2013; 21(1): 31-42.

[21] Kumar CS. and Riyaz A. On zero-inflated logarithmic series distribution and its application. Statistica, 2013; 73(4): 477-492. https://doi.org/10.1007/s10182-014-0229-1

[22] Wagner GG, Burkhauser RV. and Behringer F. The English language public use file of the German socio-economic panel. Journal of Human Resource, 1993; 28: 429-433.

[23] Bliss CI. and Fisher RA. Fitting the negative binomial distribution to biological data and note on the efficient fitting of the negative binomial. Biometrics, 1953; 9: 176-200. https://doi.org/10.2307/3001850

[24] Rao CR. Minimum variance and the estimation of several parameters. In Mathematical Proceedings of the Cambridge Philosophical Society, 1947; 43: 280-283. Cambridge University Press. https://doi.org/10.1017/s0305004100023471

Received on 25-10-2020 Accepted on 20-11-2020 Published on 24-11-2020

DOI: https://doi.org/10.15377/2409-5761.2020.07.7