STATISTICA, anno LXXVIII, n. 2, 2018

ON A MODIFIED YULE DISTRIBUTION

C. Satheesh Kumar [1]
Department of Statistics, University of Kerala, Trivandrum-695 581, India
Sivasankarapanicker Harisankar
Department of Statistics, University of Kerala, Trivandrum-695 581, India

1. INTRODUCTION

Yule (1925) introduced a one parameter discrete distribution namely "the Yule distribution (YD)", which he obtained by compounding a shifted geometric distribution with the exponential distribution through the following probability mass function (p.m.f.)

$$g(x)=\frac{\rho\Gamma(\rho+1)\Gamma(x)}{\Gamma(\rho+x+1)}, \tag{1}$$

for $x=1,2,3,\cdots$ with $\rho>0$. The YD has been extensively used for describing data sets of various fields. For example, Simon (1955), Simon (1960) and Haight (1966) used YD to model word frequency data while Kendall (1961) utilized it for describing certain types of bibliographic data sets. Xekalaki (1983) applied the YD in an econometric context.

Mishra (2009) obtaind a generalization of YD through the following p.m.f., in which $\rho>0, k>0$ and $x=1,2,3,\cdots$

$$g_1(x)=\frac{\rho\Gamma(k+x-1)\Gamma(\rho+k)}{\Gamma(\rho+x+k)\Gamma(k)}. \tag{2}$$

A distribution with p.m.f (2) here after we denote as GYD (ρ,k). Clearly, GYD $(\rho,1)$ is YD. Mishra (2009) used the GYD (ρ,k) for describing certain bibiliographical data sets.

Since both the YD and GYD (ρ,k) have been found extensive application in several practical situations, a more generalized class of GYD (ρ,k) is quite relevant. As such, through this paper we consider a modified version of the GYD (ρ,k) which we named as "the modified Yule distribution (MYD)" and study some of its important properties. In Section 2, we present the definition of the MYD and derive its p.g.f., expression of

[1] Corresponding Author. E-mail: drcsatheeshkumar@gmail.com

factorial moments, raw moments, mean, variance and recursion formulae for its probabilities, raw moments and factorial moments. In Section 3, we discuss the estimation of the parameters of MYD by method of maximum likelihood and consider certain test procedures for testing the significance of the additional parameter of the MYD. Here it is also illustrated the practical suitability of MYD with the help of two real life data sets, among them the first one is a bibliographical data sets, the other is a biological data set. It is shown that both the GYD and YD gives fitts while the MYD gives a better fit. In Section 4, we carried out a simulation study for examining the performance of the maximum likelihood estimators of the parameters of the distribution.

Throughout the paper we adopted the following shorter notation, for $i = 0, 1, 2, \cdots$

$$\Omega_i^{-1} = {}_2F_1(1+i, k+i, \rho+k+1+i; \theta), \tag{3}$$

where ${}_2F_1(.)$ is the Gaussian hypergeometric function (GHF) as defined in (2.3) of Kumar and Riyaz (2013). For more details of GHF, see Slater (1966) or Mathai and Haubold (2008).

Further we need the following series representation in the sequel

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

2. Definition and properties

In this section, first we present the definition of the MYD.

Definition 1. *A positive, integer valued random variable X is said to follow "the modified Yule distribution (MYD)", if its p.m.f $h_x = P(X = x)$ is the following, for $x = 1, 2, 3 \cdots, \rho > 0$, $k > 0$ and $0 < \theta \leq 1$*

$$h_x = \frac{\Omega_0\, \Gamma(k+x-1)\, \theta^{x-1} \Gamma(\rho+k+1)}{\Gamma(\rho+x+k)\Gamma(k)}, \tag{5}$$

in which Ω_0 is as defined in (3).

Clearly when $\theta = 1$, the MYD reduces to the GYD with p.m.f. (2) and $\theta = 1, k = 1$ the MYD reduces to the YD with p.m.f. (1). We obtain the p.g.f of the MYD through the following result.

Proposition 2. *The p.g.f of the MYD with p.m.f (5) is the following*

$$H(t) = \Omega_0\; t\; {}_2F_1(1, k; \rho+k+1; \theta t). \tag{6}$$

PROOF. By definition, the p.g.f of the MYD with p.m.f (5) is given by

$$\begin{aligned} H(t) &= \sum_{x=1}^{\infty} h_x t^x \\ &= \sum_{x=1}^{\infty} \frac{\Omega_0\ \Gamma(k+x-1)\,\theta^{x-1}\,\Gamma(\rho+k+1)t^x}{\Gamma(\rho+x+k)\Gamma(k)} \\ &= \sum_{x=0}^{\infty} \frac{\Omega_0\,(1)_x\,\Gamma(k+x)\Gamma(\rho+k+1)(\theta t)^x}{\Gamma(k)\Gamma(\rho+k+x+1)x!} \end{aligned}$$

since $(a)_r = \frac{\Gamma(a+r)}{\Gamma(a)}$, and $(1)_x = x!$. Thus we have

$$H(t) = \Omega_0\, t \sum_{x=0}^{\infty} \frac{(1)_x (k)_x}{(\rho+k+1)x} \frac{(\theta t)^x}{x!}, \tag{7}$$

which gives (6). □

PROPOSITION 3. *The characteristic function $\psi(t)$ of the MYD is the following, for any $t \in R$ and $i = \sqrt{-1}$*

$$\psi(t) = \Omega_0\, e^{it}\ {}_2F_1(1,k;\rho+k+1;\theta e^{it}). \tag{8}$$

PROPOSITION 4. *For any positive integer r, the r^{th} factorial moment $\mu_{[r]}$ of the MYD exist finitely and is given by*

$$\mu_{[r]} = \Omega_0 \frac{(1)_r (k)_{r-1}\,\theta^{r-1}}{(\rho+k+1)_{r-1}} \left[\frac{\theta(K+r-1)\Omega_r^{-1}}{\rho+r+k} + \Omega_{r-1}^{-1} \right]. \tag{9}$$

PROOF. The factorial moment generating function F(t) of the MYD with p.g.f (6) is given by

$$F(t) = \sum_{r=0}^{\infty} \mu_{[r]} \frac{t^r}{r!} \tag{10}$$

$$= \Omega_0\,(t+1)\ {}_2F_1[1,k;\rho+k+1;\theta(t+1)]. \tag{11}$$

On expanding the Gauss hypergeometric function in (11), we have

$$= \Omega_0(t+1) \sum_{r=0}^{\infty} \frac{(1)_r (k)_r \theta^r}{(\rho+k+1)_r\ r!} (t+1)^r. \tag{12}$$

By applying binomial expansion in (12) we obtain

$$F(t) = \Omega_0(t+1) \sum_{r=0}^{\infty} \frac{(1)_r (k)_r \theta^r}{(\rho+k+1)_r\ r!} \sum_{m=0}^{r} \binom{r}{m} t^{r-m}, \tag{13}$$

which on simplification gives

$$F(t) = \Omega_0\left[\sum_{r=1}^{\infty}\sum_{m=0}^{\infty}\frac{(1)_{r+m-1}(k)_{r+m-1}\theta^{r+m-1}r!}{(\rho+k+1)_{r+m-1}(r+m-1)!}\binom{r+m-1}{m}\frac{t^r}{r!} + \sum_{r=0}^{\infty}\sum_{m=0}^{\infty}\frac{(1)_{r+m}(k)_{r+m}\theta^{r+m}r!}{(\rho+k+1)_{r+m}(r+m)!}\binom{r+m}{m}\frac{t^r}{r!}\right]. \quad (14)$$

In the light of the series representation (4) of Kumar and Nair (2014) On equating the co-efficients of $t^r(r!)^{-1}$ on the right hand side expression of (10) and (14), we get

$$\mu_{[r]} = \Omega_0\left[\sum_{m=0}^{\infty}\frac{(1)_{r+m-1}(k)_{r+m-1}\theta^{r+m-1}r!}{(\rho+k+1)_{r+m-1}(r+m-1)!}\binom{r+m-1}{m} + \sum_{m=0}^{\infty}\frac{(1)_{r+m}(k)_{r+m}\theta^{r+m}r!}{(\rho+k+1)_{r+m}(r+m)!}\binom{r+m}{m}\right], \quad (15)$$

Since $(A)_{n+m} = (A)_n(A+n)_m$ and on using (3) in (15), we get (9). □

PROPOSITION 5. *For any positive integer r, the r^{th} raw moment μ_r of MYD is*

$$\mu_r = \Omega_0\sum_{v=0}^{r}\sum_{m=0}^{v}\binom{r}{v}S(v,m)\theta^m\frac{m!(k)_m}{(\rho+k+1)_m}\Omega_m^{-1}, \quad (16)$$

where $S(n,r)$ is the Stirling numbers of the second kind (see Riordan, 1968).

PROOF. By definition, the characteristic function of the MYD is given by

$$\psi(t) = \sum_{r=0}^{\infty}\mu_r\frac{(it)^r}{r!} \quad (17)$$

$$= \Omega_0 e^{it}{}_2F_1(1,k,\rho+k+1;\theta e^{it}). \quad (18)$$

On expanding (18), we get

$$\psi(t) = \Omega_0\sum_{n=0}^{\infty}\frac{(1)_n(k)_n}{(\rho+k+1)_n}\frac{\theta^n}{n!}\sum_{r=0}^{\infty}\frac{(it)^r}{r!}(n+1)^r. \quad (19)$$

Equating the coefficients of $(it)^r(r!)^{-1}$ on right hand side of (17) and (19) to get

$$\mu_r = \Omega_0\sum_{n=0}^{\infty}\frac{(1)_n(k)_n}{(\rho+k+1)_n}\frac{\theta^n}{n!}(n+1)^r \quad (20)$$

By applying binomial expansion and Stirling numbers of second kind in (20) we have

$$\mu_r = \Omega_0\sum_{n=0}^{\infty}\frac{(1)_n(k)_n}{(\rho+k+1)_n}\frac{\theta^n}{n!}\sum_{v=0}^{r}\binom{r}{v}\sum_{m=0}^{v}S(v,m)(n)_m, \quad (21)$$

Now, on rearranging the terms in (21) we obtain (16), in the light of (3). □

PROPOSITION 6. *Mean and variance of the MYD are*

$$Mean = 1 + \frac{k\theta\Omega_0}{\Omega_1(\rho+k+1)} \tag{22}$$

and

$$Variance = \frac{\theta\Omega_0 k}{\rho+k+1}\left[\frac{1}{\Omega_1} + \theta\left(\frac{2(k+1)}{(\rho+k+2)\Omega_2} - \frac{k\Omega_0}{(\rho+k+1)\Omega_1^2}\right)\right]. \tag{23}$$

REMARK 7. *From (22) and (23), it is seen that the MYD is under-dispersed if and only if*

$$(\rho+k+2)\left[(\rho+k+1)^2 + \theta^2 k^2\right]\Omega_2 > 2\,\theta^2\,k(k+1)\,(\rho+k+1)\Omega_0,$$

for all values of the parameters ρ, θ and k and the MYD is over-dispersed otherwise.

PROPOSITION 8. *For $x \geq 1$, the following is a simple recursion formula for probabilities $h_x = h_x(1,k;\rho+k+1)$ of the MYD with p.g.f (6).*

$$h_{x+1}(1,k,\rho+k+1) = \frac{k\ \theta\ \Omega_0\ h_x(2,k+1;\rho+k+2)}{\Omega_1\ x\ (\rho+k+1)}. \tag{24}$$

PROOF. From (6), we have

$$H(t) = \sum_{x=0}^{\infty} h_x(1,k;\rho+k+1)t^x = \Omega_0\ t\ {}_2F_1(1,k;\rho+k+1;\theta t). \tag{25}$$

Differentiating the above equation with respect to t, we get

$$\begin{aligned}\sum_{x=0}^{\infty}(x+1)h_{x+1}(1,k;\rho+k+1)t^x &= \frac{\Omega_0\ \theta\ k\ t}{(\rho+k+1)}\ {}_2F_1(2,k+1;\rho+k+2;\theta t) \\ &\quad + \Omega_0\ {}_2F_1(1,k;\rho+k+1;\theta t).\end{aligned} \tag{26}$$

In (6) on replacing $1, k$ and $\rho+k+1$ by $2, k+1$ and $\rho+k+2$ respectively, we obtain

$${}_2F_1(2,k+1;\rho+k+2;\theta t) = \Omega_1{}^{-1}t^{-1}\sum_{x=0}^{\infty} h_x(2,k+1;\rho+k+2)t^x. \tag{27}$$

Substituting (25) and (27) in (26) we get

$$\begin{aligned}&\sum_{x=0}^{\infty}(x+1)h_{x+1}(1,k;\rho+k+1)t^x = \\ &\frac{\Omega_0\theta k}{\Omega_1(\rho+k+1)}\sum_{x=0}^{\infty} h_x(2,k+1;\rho+k+2)t^x + \sum_{x=0}^{\infty} h_{x+1}(1,k;\rho+k+1)t^x.\end{aligned} \tag{28}$$

Simplifying and equating the coefficients of t^x on both sides of (28), we get (24). □

PROPOSITION 9. *The following is a simple recursion formula for raw moments* $\mu_r = \mu_r(1,k;\rho+k+1)$ *of the MYD for* $r \geq 0$.

$$\begin{aligned}\mu_{r+1}(1,k;\rho+k+1) &= \frac{\Omega_0\,\theta\,k}{\Omega_1(\rho+k+1)}\sum_{s=0}^{r}\frac{r!}{s!(r-s)!}\mu_{r-s}(2,k;\rho+k+2)\\ &\quad+\mu_r(1,k;\rho+k+1). \qquad (29)\end{aligned}$$

PROOF. On differentiating (17) and (18) with respect to t, we get

$$\frac{\partial\psi(t)}{\partial t} = \frac{i\,\Omega_0\,(e^{it})^2\,\theta}{\rho+k+1}\,{}_2F_1(2,k+1;\rho+k+2;\theta e^{it}) + i\,e^{it}\,\Omega_0\,{}_2F_1(1,k;\rho+k+1;\theta e^{it}). \qquad (30)$$

By using (8) with $1,k$ and $\rho+k+1$ by $2,k+1$ and $\rho+k+2$ respectively, we obtain

$$e^{it}\,{}_2F_1(2,k+1;\rho+k+2;\theta e^{it}) = \Omega_1{}^{-1}\sum_{r=0}^{\infty}\mu_r(2,k+1;\rho+k+2)\frac{(it)^r}{r!}. \qquad (31)$$

Substituting (18) and (31) in (30) we get

$$\begin{aligned}\sum_{r=0}^{\infty}\mu_{r+1}(1,k;\rho+k+1)\frac{(it)^r}{r!} &= \frac{\Omega_0\theta k}{\Omega_1(\rho+k+1)}\\ &\quad\times\sum_{r=0}^{\infty}\sum_{s=0}^{r}\mu_{r-s}(2,k+1;\rho+k+2)\frac{(it)^r}{(r-s)!s!}\\ &\quad+\mu_r(1,k,\rho+k+1)\frac{(it)^r}{r!}. \qquad (32)\end{aligned}$$

On equating the coefficients of $\frac{(it)^r}{r!}$ on both sides of (32), we get (29). □

PROPOSITION 10. *The following is a simple recursion formula for factorial moments* $\mu_{[r]} = \mu_{[r]}(1,k;\rho+k+1)$ *of the MYD, for* $r \geq 0$.

$$\begin{aligned}\mu_{[r+1]}(1,k;\rho+k+1) &= \frac{\Omega_0\,\theta\,k}{\Omega_1(\rho+k+1)}\mu_{[r]}(2,k+1;\rho+k+2)\\ &\quad+\sum_{j=0}^{r}\frac{(-1)^j\,r!}{(r-j)!}\mu_{[r-j]}(1,k;\rho+k+1). \qquad (33)\end{aligned}$$

PROOF. The factorial moment generating function F(t) of the MYD with p.g.f (6) is given by

$$\begin{aligned}F(t) &= \sum_{r=0}^{\infty}\mu_{[r]}\frac{t^r}{r!} \qquad (34)\\ &= \Omega_0(t+1)\,{}_2F_1[1,k;\rho+k+1;\theta(t+1)]. \qquad (35)\end{aligned}$$

Differentiating the above equation with respect to t, we get

$$\sum_{r=0}^{\infty}\mu_{[r]}(1,k;\rho+k+1)\frac{t^r}{r!} = \Omega_0\,{}_2F_1[1,k;\rho+k+1;\theta(t+1)]+\frac{\Omega_0\,(t+1)\theta k}{(\rho+k+1)} \times{}_2F_1[2,k+1;\rho+k+2;\theta(t+1)]. \tag{36}$$

On simplification we get

$$\sum_{r=0}^{\infty}\mu_{[r+1]}(1,k;\rho+k+1)\frac{t^r}{r!} = \sum_{r=0}^{\infty}\sum_{j=0}^{\infty}\mu_{[r]}(1,k;\rho+k+1)\frac{t^{(r+j)}(-1)^j}{r!} +\frac{\Omega_0\theta k}{\Omega_1\,(\rho+k+1)}\sum_{r=0}^{\infty}\mu_{[r]}(2,k+1;\rho+k+2)\frac{t^r}{r!}. \tag{37}$$

On equating the coefficients of $\frac{t^r}{r!}$ on both sides of (37), we get (33). □

3. Estimation and testing of the hypothesis

In this section we discuss the estimation of the parameters θ, k and ρ of the MYD by the method of maximum likelihood and certain test procedures for testing the significance of the additional parameter θ of the MYD.

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

$$L=\prod_{x=0}^{y}[h_x]^{a(x)}, \tag{38}$$

which implies

$$\log L=\sum_{x=0}^{y}a(x)\log h_x. \tag{39}$$

Let $\hat{\theta}$, $\hat{\rho}$ and $\hat{k}$ be the MLEs of θ, ρ and k. Noe these MLEs of the parameters are obtained by solving the following likelihood equations

$$\frac{\partial \log L}{\partial \theta}=0, \tag{40}$$

equivalently,

$$\sum_{x=0}^{y}-a(x)\left[\frac{\Omega_0 k}{\Omega_1(\rho+k+1)}+\frac{x-1}{\theta}\right]=0. \tag{41}$$

$$\frac{\partial \log L}{\partial \rho} = 0, \tag{42}$$

equivalently,

$$\sum_{x=0}^{y} a(x) \left\{ -\Omega_0 \left[\sum_{r=0}^{\infty} \theta^r \frac{\Gamma(k+r)\Gamma(\rho+k+1)}{\Gamma(k)\Gamma(\rho+k+r+1)} (\psi(\rho+k+1) - \psi(\rho+r+k+1)) \right] + (\psi(\rho+k+1) - \psi(\rho+x+k)) \right\} = 0. \tag{43}$$

$$\frac{\partial \log L}{\partial k} = 0, \tag{44}$$

equivalently,

$$\sum_{x=0}^{y} a(x) \left\{ -\Omega_0 \sum_{r=0}^{\infty} \theta^r \frac{\Gamma(\rho+k+1)\Gamma(k+r)}{\Gamma(k)\Gamma(\rho+k+r+1} [\psi(\rho+k+1) - \psi(\rho+k+r+1) + \psi(k+r) - \psi(k)] + [(\psi(k+x-1) + \psi(\rho+k+1) - \psi(k) - \psi(\rho+k+x)] \right\} = 0, \tag{45}$$

where $\psi(\beta) = \frac{\partial}{\partial \beta} \log \Gamma(\beta)$. On solving these likelihood equations by using some mathematical softwares such as MATHCAD, MATHEMATICA etc., one can obtain the maximum likelihood estimators of the parameters θ, ρ and k of MYD.

For numerical illustration, we have considered two real life data sets, of which the first data is on the distribution of 1533 biologists according to the number of research papers to their credit in the Review of Applied Entomology, Volume 24, 1936. For details, see Williams (1943) and the second data set is on family epidemics of common cold obtained by Heasman and Reid (1961). We have fitted the MYD, the GYD, and the YD to both these data sets and the results obtained along with the corresponding values of the expected frequencies, chi-square statistic, degrees of freedom (d.f), p-value, Akaike information criterion (AIC) and Bayesian information criterion (BIC) in respect of each of the models are presented in Tables 1 and 2. Based on the computed values of chi-square statistic, p-value, AIC and BIC, it can be observed that the MYD gives the best fit to both the data sets where the existing models the GYD and the YD fails.

TABLE 1
Observed frequencies and computed values of expected frequencies of the MYD, the GYD and the YD by the method of maximum likelihood for the first data set.

x	Observed	YD	GYD	MYD
1	1062	1118.51	1079.50	1024.76
2	263	235.01	270.46	287.46
3	120	91.40	95.88	113.71
4	50	42.80	40.88	52.15
5	22	19.50	20.01	23.33
6	7	11.10	12.03	9.24
7	6	6.90	6.30	6.53
8	2	4.50	3.81	2.70
9	0	2.08	2.53	0.62
10	1	1.20	1.60	0.50
Total	1533	1533	1533	1533
d.f.		6	5	4
Estimates of parameters		ρ=2.7	ρ=3.8 k=1.60	ρ=0.01 k=0.80 θ=0.01
χ^2-value		21.80	14.52	5.83
p-value		0.0013	0.0126	0.2122
AIC		3082.78	3059.5	3046.78
BIC		3081.78	3057.5	3043.78

TABLE 2
Observed frequencies and computed values of expected frequencies of the MYD, the GYD and the YD by the method of maximum likelihood for the second data set.

x	Observed	YD	GYD	MYD
1	156	176.52	170.70	164.01
2	55	39.76	44.85	50.00
3	19	14.63	16.14	17.84
4	10	6.67	6.04	7.46
5	2	4.42	4.27	2.69
Total	242	242	242	242
d.f.		3	2	1
Estimates of parameters		ρ=2.56	ρ=4.22 k=1.93	ρ=0.17 k=1.87 θ=0.49
χ^2-value		12.51	7.87	2.01
p-value		0.0058	0.0195	0.1563
AIC		507.28	498.12	492.12
BIC		505.97	495.51	488.21

3.1. *Testing of the hypothesis*

Here we present two test procedures - the generalized likelihood ratio test (GLRT) and Rao's efficient score test (REST) for testing the significance of the additional parameter θ of the MYD.

Let the null hypothesis be
$H_0 : \theta = 1$ against the alternative hypothesis $H_1 : \theta \neq 1$.

In case of GLRT, the test statistic is

$$-2\log\Lambda = 2\Big(\log L(\hat{\underline{\Lambda}}; x) - \log L(\hat{\underline{\Lambda}}^*; x)\Big), \tag{46}$$

where $\hat{\Lambda}$ is the MLE of $\underline{\Lambda} = (\theta, \rho, k)$ with no restriction and $\hat{\underline{\Lambda}}^*$ is the MLE of $\underline{\Lambda}$ when $\theta = 1$. The test statistic $-2\log\Lambda$ is asymptotically distributed as a chi-square with one degree of freedom. For details see Rao (1947).

In case of REST, the test statistic is $S = T'\Phi^{-1}T$, where T' is

$$T' = \left(\frac{1}{\sqrt{n}}\frac{\partial \log L}{\partial \theta}, \frac{1}{\sqrt{n}}\frac{\partial \log L}{\partial \rho}, \frac{1}{\sqrt{n}}\frac{\partial \log L}{\partial k}\right), \tag{47}$$

where Φ is the Fisher information matrix. The test statistic S follows chi-square with 1 d.f See Rao (1947). Using the data sets given in Table 1 and Table 2, we have computed the values of the test statistic in case of the GLRT and REST and presented in Table 3. For numerical illustration we have computed the values of test statistic in case of the GLRT and REST for the first data set as follows.

$$-2\log\Lambda = 2(1527.75 - 1520.39) = 14.72 \tag{48}$$

and

$$\begin{aligned} S_1 &= \begin{pmatrix} -15.63 & 0.45 & 2.53 \end{pmatrix} \begin{pmatrix} 0.02 & 0.30 & 0.04 \\ 0.30 & 4.0 & 0.80 \\ 0.04 & 0.8 & 0.20 \end{pmatrix} \begin{pmatrix} -15.63 \\ 0.45 \\ 2.53 \end{pmatrix} \\ &= 6.29 \end{aligned}$$

Similarly, GLRT and REST for the second data set is as follows.

$$-2\log\Lambda = 2(247.06 - 243.071) = 7.97 \tag{49}$$

and

$$\begin{aligned} S_2 &= \begin{pmatrix} -11.92 & 0.93 & -0.47 \end{pmatrix} \begin{pmatrix} 0.10 & 0.60 & 0.01 \\ 0.60 & 8.0 & 0.80 \\ 0.01 & 0.8 & 0.50 \end{pmatrix} \begin{pmatrix} -11.92 \\ 0.93 \\ -0.47 \end{pmatrix} \\ &= 7.34 \end{aligned}$$

Since the critical value for the test at 5% level of significance is 3.84 at one degree of freedom, the null hypothesis is rejected in all these cases in respect of both the GLRT and REST.

TABLE 3
The computed values of the statistic for GLRT and REST for MYD.

Calculated values of	Data set 1	Data set 2
GLRT	14.72	7.97
REST	6.29	7.34

4. SIMULATION

Here we simulate random variates from the MYD and obtain the bias and standard errors of estimators of the parameters of the distribution by the method of maximum likelihood. We have simulated two sets of observations corresponds to the following two sets of parameters (i) $\theta = 0.45, \rho = 0.25, k = 0.50$ and (ii) $\theta = 0.72, \rho = 0.20, k = 0.45$ of which (i) corresponds to the under-dispersed MYD whereas (ii) corresponds to the over-dispersed MYD. By using simulated observations, we estimated the parameters θ, ρ and k of the MYD and thereby computed the values of the absolute bias and standard errors of each of the estimators. The results obtained are presented in Table 4. From Table 4, it can be observed that both the absolute values of bias and standard errors of the estimators of the parameters are in decreasing order as the sample size increases.

TABLE 4
Bias and standard errors in the parenthesis of the estimators of the parameters θ, ρ and k of the MYD for the simulated data sets.

			MLE	
Parameter set	Sample size	$\hat{\rho}$	$\hat{\theta}$	$\hat{k}$
(i)	$n = 200$	0.0513	0.0870	0.0820
		(0.0075)	(0.0089)	(0.0082)
	$n = 300$	0.0140	0.0068	0.0490
		(0.0045)	(0.0015)	(0.0029)
	$n = 500$	0.0035	0.0014	0.0480
		(0.0043)	(0.0010)	(0.0024)
(ii)	$n = 200$	0.1360	0.0488	0.0440
		(0.0321)	(0.0047)	(0.0071)
	$n = 300$	0.0723	0.0277	0.0210
		(0.0268)	(0.0022)	(0.0054)
	$n = 500$	0.0067	0.0094	0.0190
		(0.0016)	(0.00084)	(0.00086)

ACKNOWLEDGEMENTS

The Authors are highly thankful to the Chief Editor and the anonymous referees for their valuable comments on the earlier version of the paper.

References

F. A. Haight (1966). *Some statistical problems in connection with word association data*. Journal of Mathematical Psychology, 3, no. 1, pp. 217–233.

M. Heasman, D. Reid (1961). *Theory and observation in family epidemics of the common cold*. British journal of preventive & social medicine, 15, no. 1, p. 12.

M. G. Kendall (1961). *Natural law in the social sciences : Presidential address, delivered to the Royal Statistical Society on Wednesday, November 16th, 1960*. Journal of the Royal Statistical Society, 124, no. 1, pp. 1–16.

C. S. Kumar, B. U. Nair (2014). *A three parameter hyper-Poisson distribution and some of its properties*. Statistica, 74, no. 2, pp. 183–198.

C. S. Kumar, A. Riyaz (2013). *On zero-inflated logarithmic series distribution and its modification*. Statistica, 73, no. 4, pp. 477–492.

A. M. Mathai, H. J. Haubold (2008). *Special Functions for Applied Scientists*. Springer-Verlag, New York.

A. Mishra (2009). *On a generalized Yule distribution*. Assam Statistical Review, 23, pp. 140–150.

C. R. Rao (1947). *Minimum variance and the estimation of several parameters*. In *Mathematical Proceedings of the Cambridge Philosophical Society*. Cambridge University Press, Cambridge, vol. 43, no. 2, pp. 280–283.

J. Riordan (1968). *Combinatorial identities*. Wiley, New York.

H. A. Simon (1955). *On a class of skew distribution functions*. Biometrika, 42, no. 3/4, pp. 425–440.

H. A. Simon (1960). *Some further notes on a class of skew distribution functions*. Information and Control, 3, no. 1, pp. 80–88.

L. J. Slater (1966). *Generalized Hypergeometric Functions*. Cambridge University Press, Cambridge.

C. Williams (1943). *The numbers of publications written by biologists*. Annals of Eugenics, 12, no. 1, pp. 143–146.

E. Xekalaki (1983). *A property of the Yule distribution and its applications*. Communications in Statistics-Theory and Methods, 12, no. 10, pp. 1181–1189.

G. U. Yule (1925). *A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS*. Philosophical Transactions of the Royal Society of London, Series B, 213, pp. 21–87.

SUMMARY

A modified version of Yule distribution is introduced here and discuss some of its properties by deriving expressions for its probability generating function, raw moments, factorial moments etc. Certain recursion formulae for its probabilities, raw moments and factorial moments are also developed. Various methods of estimation are employed for estimating the parameters of the distribution and certain test procedures are suggested for testing the significance of the additional parameters of the distribution. The distribution has been fitted to certain real-life data sets for illustrating its usefulness, compared with certain existing models available in the literature. Further, a simulation study is conducted for assessing the performance of the maximum likelihood estimators.

Keywords: Generalized likelihood ratio test; Maximum likelihood estimation; Model Selection; Probability generating function; Simulation.