Aligarh Journal of Statistics

Vol. 34 (2014), 45-54

LOGARITHMIC SERIES DISTRIBUTION OF ORDER k

C. Satheesh Kumar and A. Riyaz

Through the present paper we develop an order k version of the logarithmic series distribution and derive its probability mass function, mean and variance. We estimate the parameters of this distribution by the method of maximum likelihood and the distribution has been fitted to certain real life data sets. Also, we discuss the generalized likelihood ratio test and Rao's score test for testing the significance of the additional parameters of the distribution.

1. INTRODUCTION

The standard logarithmic series distribution (in short, the LSD) of Fisher et. al. (1943) has been found applications in several areas of research such as biology, ecology, economics, operations research and marine sciences. Fisher et. al. (1943) obtained the LSD as the limit of a zero-truncated negative binomial distribution for investigating the distribution of butterflies in the Malayan Peninsula. Chatfield et. al. (1966) used the LSD to represent the distribution of number of items of a product purchased by a buyer in a specified time period. For a detailed account of the LSD see chapter 7 of Johnson *et. al.* (2005). Various generalized versions of the LSD have been proposed in the literature. For example see Tripathi and Gupta (1985, 1988), Ong (2000) and Khang and Ong (2007).

Through this paper we consider a generalized form of the LSD which we termed as "the logarithmic series distribution of order k (LSD_k)". In section 2 we obtain the LSD_k as the limiting form of the zero-truncated cluster negative binomial distribution and derive its probability mass function (pmf), mean and variance. In section 3, we discuss the estimation of the parameters of the LSD_k by the method of maximum likelihood and illustrate its usefulness through fitting the model to certain real life data sets. In section 4 we consider the generalized likelihood ratio test and Rao's efficient score test for testing the significance of the additional parameters of the distribution.

2. THE LOGARITHMIC SERIES DISTRIBUTION OF ORDER k AND ITS PROPERTIES

Xekalaki and Panaretos (1989) introduced the cluster negative binomial distribution through the following probability generating function (pgf).

$$G(z) = (1-\sum_{j=1}^{k}\theta_j)^r (1-\sum_{j=1}^{k}\theta_j z^j)^{-r}, \tag{2.1}$$

in which $r>0$, $\theta_j>0$ for each $j=1,2,...,k$ such that $\sum_{j=1}^{k}\theta_j<1$. The pgf of the zero-truncated cluster negative binomial distribution is

$$G_1(z) = \frac{(1-\sum_{j=1}^{k}\theta_j)^r\ (1-\sum_{j=1}^{k}\theta_j z^j)^{-r}-1}{1-(1-\sum_{j=1}^{k}\theta_j)^r}. \tag{2.2}$$

On taking the limit as $r\to 0$, we get the following from (2.2).

$$H(z) == \frac{\ln(1-\sum_{j=1}^{k}\theta_j z^j)}{\ln(1-\sum_{j=1}^{k}\theta_j)} \tag{2.3}$$

Clearly, when $k=1$ the pgf given in (2.3) reduces to that of the LSD of Fisher *et. al.* (1943).

A distribution with pgf $H(z)$ we call "the logarithmic series distribution of order k (or in short the LSD_k)". This pgf of LSD_k given in (2.3) can also be written as

$$H(z) = \frac{\sum_{j=1}^{k}\theta_j z^j}{\sum_{j=1}^{k}\theta_j}\ \frac{{}_2F_1(1\ ,1\,;2\,;\sum_{j=1}^{k}\theta_j z^j\ )}{{}_2F_1(1\ ,1\,;2\,;\sum_{j=1}^{k}\theta_j)}. \tag{2.4}$$

in which

$${}_2F_1(a\ ,b\,;c\,;z) = \sum_{r=0}^{\infty}\frac{a(a+1)...(a+r-1)b(b+1)...(b+r-1)}{c(c+1)...(c+r-1)}\frac{z^r}{r!}$$

is the Gauss hypergeometric function. For details regarding Gauss hypergeometric function see Slater (1966) or Mathai and Haubold (2008). We obtain the pmf of the LSD_k through the following result.

Theorem 2.1: For $x = 1, 2, ...$ the *pmf* $p_x = P(X=x)$ of the LSD_k with *pgf* (2.3) is the following

$$p_x = C \sum_{I_x} \frac{n!}{x_1! x_2! ... x_k!} \theta_1^{x_1} \theta_2^{x_2} ... \theta_k^{x_k}, \tag{2.5}$$

in which $C = \left[-\ln(1 - \sum_{j=1}^{k} \theta_j) \right]^{-1}$,

$\sum_{I_x}$ denote the summation over all $k-$tupples $(x_1, x_2, ..., x_k)$ of non-negative integers in the set

$$I_x = \{(x_1, x_2, ..., x_k) : \sum_{j=1}^{k} j\, x_j = x\}$$

and

$$n = (\sum_{j=1}^{k} x_j) - 1 \ .$$

Proof: From (2.3) we have

$$H(z) = \sum_{x=0}^{\infty} p_x z^x \tag{2.6}$$

$$= C \left[-\ln(1 - \sum_{j=1}^{k} \theta_j z^j) \right] \tag{2.7}$$

On expanding the logarithmic function in (2.7), we get

$$H(z) = C \sum_{n=1}^{\infty} \frac{(\sum_{j=1}^{k} \theta_j z^j)^n}{n}$$

$$= C \sum_{n=0}^{\infty} \frac{(\sum_{j=1}^{k} \theta_j z^j)^{n+1}}{n+1} \ . \tag{2.8}$$

Now, by applying the multinomial expansion in (2.8) we have

$$H(z) = C \sum_{n=0}^{\infty} \sum_{J_n} \frac{(n+1)!}{x_1! x_2! ... x_k!} \frac{\theta_1^{x_1} \theta_2^{x_2} ... \theta_k^{x_k}}{(n+1)} z^{\delta} \tag{2.9}$$

in which $\delta = \sum_{j=1}^{k} j\, x_j$ and $\sum_{J_n}$ denote the summation over all $k-$tupples $(x_1, x_2, ..., x_k)$ of non-negative integers in the set

$$J_n = \{(x_1, x_2, ..., x_k) : \sum_{j=1}^{k} x_j = n+1\}.$$

On equating the coefficient of z^x on the right hand side expressions of (2.6) and (2.9) we get (2.5).

Next we obtain the mean and variance of the LSD_k through the following result.

Theorem 2.2: The mean and variance of the LSD_k are

$$\text{Mean} = C\alpha\beta$$

and

$$\text{Variance} = C\alpha\left[\sum_{j=1}^{k} j(j-1)\theta_j + \beta + C\alpha\beta^2\right],$$

in which

$$\alpha = (1 - \sum_{j=1}^{k} \theta_j)^{-1}$$

and

$$\beta = \sum_{j=1}^{k} j\theta_j .$$

Proof: follows from the fact that

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

and

$$\text{Variance} = H^{(2)}(1) + H^{(1)}(1) - (H^{(1)}(1))^2,$$

where for $r = 1, 2$,

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

3. ESTIMATION

Here we discuss the estimation of the parameters of the LSD_k by the method of maximum likelihood and have illustrated the usefulness of the model with the help of certain real life data sets. Let $a(x)$ be the observed frequency of x events and let y be the highest value of x observed. Then the likelihood function of the sample is

$$L = \prod_{x=1}^{y} [p_x]^{a(x)}, \tag{3.1}$$

where p_x is the *pmf* of the LSD_k as given in (2.5). Now taking logarithm on both sides of (3.1), we have

$$\log L = \sum_{x=1}^{y} a(x) \log(p_x)$$

$$= \sum_{x=1}^{y} a(x) \left[\log C + \log \phi(x; \theta_1; \theta_2, ..., \theta_k) \right], \tag{3.2}$$

in which

$$\phi(x; \theta_1; \theta_2, ..., \theta_k) = \sum_{I_x} \frac{n!}{x_1! x_2! ... x_k!} \theta_1^{x_1} \theta_2^{x_2} ... \theta_k^{x_k}.$$

Let $\hat{\theta}_j$ denote the maximum likelihood estimator of the parameter θ_j of the LSD_k, for $j = 1, 2, ..., k$. On differentiating (3.2) partially with respect to the parameter θ_j, for $j = 1, 2, ..., k$ and equating to zero, we get the following system of likelihood equations.

$$\sum_{x=1}^{y} a(x) \left[\frac{-C}{(1 - \sum_{J=1}^{k} \theta_J} + \frac{\sum_{I_x} \frac{n!}{x_1! x_2! ... (x_j - 1)! x_{j+1}! ... x_k!} \theta_1^{x_1} \theta_2^{x_2} ... \theta_j^{x_j - 1} \theta_j^{x_j + 1} ... \theta_k^{x_k}}{\phi(x; \theta_1, \theta_2, ..., \theta_k)} \right] = 0 \tag{3.3}$$

The likelihood equations do not always have a solution because the LSD_k is not a regular model. Therefore, when likelihood equations do not always have a solution, the maximum of the likelihood function attained at the border of the domain of parameters. We obtained the second order partial derivatives of $\log p_x$ with respect to parameter θ_j and using *MATHCAD* softwares we observed that these equations give negative values for all $\theta_j > 0$ such that

$\sum_{j=i}^{k}\theta_j < 1$. Thus the density of the LSD_k is log-concave and have maximum likelihood estimators of the parameter θ_j are unique (cf. Puig. 2003). Now on solving these likelihood equations by using mathematical softwares such as *MATHLAB*, *MATHCAD*, *MATHEMATICA* etc., one can obtain the maximum likelihood estimators of the parameters of the LSD_k.

For numerical illustration, we have considered three real life data sets of which the first and second data sets are from family epidemics of common cold obtained by Heasman and Reid (1961) and the third data set is a zero-truncated data set on the counts of the number of European red mites on apple leaves, used earlier by Jani and Shah (1979). We have fitted both the LSD and the LSD_k to these data sets and the results obtained along with the corresponding values of the expected frequencies, chi-square values, degrees of freedom $(d.f.)$ and $P-$values for each of the models are presented in the Table 1, Table 2 and Table 3. Based on the chi-square values and $P-$values given in the tables, it can be observed that the LSD_k gives a better fit to the given data sets compared to the existing model.

Table 1: Observed frequencies and computed values of the expected frequencies of the LSD and the LSD_k by method of maximum likelihood for the first data set.

Number of cases	Observation	LSD	LSD_k	
			$k=2$	$k=3$
1	156	179.080	159.478	156.332
2	55	42.108	53.724	50.578
3	19	13.310	17.182	19.36
4	10	4.598	6.776	7.968
5	2	2.904	4.840	7.774
Total	242	242	242	242
Estimates of the parameters		$\hat{\theta}_1 = 0.47$	$\hat{\theta}_1 = 0.40$ $\hat{\theta}_1 = 0.055$	$\hat{\theta}_1 = 0.41$ $\hat{\theta}_2 = 0.035$ $\hat{\theta}_3 = 0.001$
Chi-square values		12.051	0.311	5.173
d.f.		2	1	1
P- values		0.04	0.568	0.023

Table 2: Observed frequencies and computed values of the expected frequencies of the *LSD* and the LSD_k by method of maximum likelihood for the second data set.

Number of cases	Observ-ation	*LSD*	LSD_k	
			$k=2$	$k=3$
1	112	129.415	110.591	116.202
2	35	32.942	39.639	38.372
3	17	11.222	16.109	14.661
4	11	4.344	7.24	6.335
5	6	3.077	7.421	5.43
Total	181	181	181	181
Estimates of the parameters		$\hat{\theta}_1 = 0.51$	$\hat{\theta}_1 = 0.53$ $\hat{\theta}_2 = 0.05$	$\hat{\theta}_1 = 0.50$ $\hat{\theta}_2 = 0.04$ $\hat{\theta}_3 = 0.001$
Chi-square values		17.812	2.835	4.317
d.f.		2	2	1
P- values		< .00001	0.242	0.038

Table 3: Observed frequencies and computed values of the expected frequencies of the *LSD* and the LSD_k by the method of maximum likelihood for the third data set.

No. of mites per leaves	Leaves observed	*LSD*	LSD_k	
			$k=2$	$k=3$
1	38	52.939	40.40	38.80
2	17	15.617	16.8	16.48
3	10	6.143	8.56	8.80
4	9	2.715	4.96	5.20
5	3	1.283	3.12	3.36
6	2	0.631	2.00	2.08
7	1	0.315	0.504	0.577
8	0	0.353	3.656	4.703
Total	80	80	80	80
Estimates of the parameters		$\hat{\theta}_1 = 0.59$	$\hat{\theta}_1 = 0.72$ $\hat{\theta}_2 = 0.04$	$\hat{\theta}_1 = 0.4$ $\hat{\theta}_2 = 0.04$ $\hat{\theta}_3 = 0.002$
Chi-square values		24.506	0.428	5.052
d.f.		2	1	1
P- values		< .00001	0.513	0.025

4. TESTING OF THE HYPOTHESIS

In this section we discuss the testing of the hypothesis

$$H_0 : \theta_{i_1} = \theta_{i_2} = ... = \theta_{i_m} = 0,$$

for any particular subset $\{i_1, i_2, ..., i_m\}$ of the set $\{1,2,\cdots, k\}$, by using generalized likelihood ratio test and Rao's efficient score test.

In case of generalized likelihood ratio test, the test statistic is

$$-2\log\lambda = 2(l_1 - l_2), \qquad (4.1)$$

where

$$l_1 = \log L(\underline{\hat{\theta}}; x),$$

in which $\underline{\hat{\theta}}$ is the maximum likelihood estimator of $\underline{\theta} = (\theta_1, \theta_2, ..., \theta_k)$ with no restrictions

and

$$l_2 = \log L(\underline{\hat{\theta}}^*; x),$$

in which $\underline{\hat{\theta}}^*$ is the maximum likelihood estimate of $\underline{\theta}$ under H_0. The log likelihood function $L = L(\underline{\theta}, \underline{x})$ is as defined in (3.2), and the test statistic $-2\log\lambda$ given in (4.1) is asymptotically distributed as χ^2 with m degree of freedom (for details see Rao, 1973).

Next we consider the testing of significance of the parameter θ_2 in case of the fitting of the LSD_k to the data sets considered in section 3. Here $H_0 : \theta_2 = 0$ against $H_1 : \theta_2 > 0$. We have computed the values of $\log L(\underline{\hat{\theta}}; x)$, $\log L(\underline{\hat{\theta}}^*; x)$ and the test statistic given in (4.1) in case of all the three data sets and presented in Table 4.

Table 4: The computed the values of $\log L(\underline{\hat{\theta}}; x)$, $\log L(\underline{\hat{\theta}}^*; x)$ and the generalized likelihood ratio test statistic.

	$\log L(\underline{\hat{\theta}}^*; x)$	$\log L(\underline{\hat{\theta}}; x)$	Test statistic
Data set 1	-107.694	-105.435	4.518
Data set 2	-92.712	-90.534	4.356
Data set 3	-55.156	-53.037	4.238

Since the critical value for the test with $\alpha = 0.05$ and degrees of freedom one is 3.84, the null hypothesis is rejected in all the cases. Hence we conclude that the additional parameter θ_2 in the model is significant in all the three data sets considered in the paper.

In case of Rao's score test, the statistic is

$$M = V'\phi^{-1}V, \tag{4.2}$$

where ϕ is the Fisher information matrix under H_0 and

$$V' = \left(\frac{1}{\sqrt{n}} \frac{\partial \log L}{\partial \theta_1}, \frac{1}{\sqrt{n}} \frac{\partial \log L}{\partial \theta_2}, \ldots, \frac{1}{\sqrt{n}} \frac{\partial \log L}{\partial \theta_k} \right)$$

under H_0, in which L is the likelihood function as defined in (3.1). The test statistic given in (4.2) follows chi-square distribution with m degrees of freedom (for details see Rao, 1973).

Next we consider the testing of significance of the parameter θ_2 in case of the fitting of the LSD_k to all the three data sets, as considered above. Let $H_0 : \theta_2 = 0$ against $H_1 : \theta_2 > 0$. We have computed the values of M for (i) the LSD_k with $k = 2$ in case of first data set as M_1 (ii) the LSD_k with $k = 2$ in case of second data set as M_2 and (iii) the LSD_k with $k = 2$ in case of third data set as M_3 as given below.

$$M_1 = \begin{pmatrix} -0.247 & 25.771 \end{pmatrix} \begin{bmatrix} 0.045 & -0.05 \\ -0.05 & 0.0097 \end{bmatrix} \begin{pmatrix} -0.247 \\ 25.771 \end{pmatrix}$$

$$= 6.674$$

$$M_2 = \begin{pmatrix} -0.584 & 14.205 \end{pmatrix} \begin{bmatrix} 0.422 & -0.236 \\ -0.236 & 0.135 \end{bmatrix} \begin{pmatrix} -0.584 \\ 14.205 \end{pmatrix}$$

$$= 31.374$$

$$M_3 = \begin{pmatrix} -4.028 & 9.359 \end{pmatrix} \begin{bmatrix} 0.14 & -0.072 \\ -0.072 & 0.044 \end{bmatrix} \begin{pmatrix} -4.028 \\ 9.359 \end{pmatrix}$$

$$= 11.543$$

Since the critical value for the test with $\alpha = 0.05$ and degrees of freedom one is 3.84, the null hypothesis is rejected in all the three cases.

REFERENCES

Chatfield C., Ehrenberg, A. S. C. and Goodhardt G. J. (1966): Progress on a simplified model of stationary purchasing behavior (with discussion). *J. Roy. Statist. Soc. Ser. A,* **129**, 317-367.

Fisher R. A., Corbet A. S. and Williams C. B. (1943): The relation between the number of individuals in a random sample of an animal population. *Journal of Animal Ecology,* **12**, 42-58.

Heasman M.A. and Reid D. D. (1961): Theory and observation in family epidemics of the common cold. *British Journal Preventive and Social Medicine,* **15,** 12-16.

Jani P.N. and Shah S.M. (1979): On fitting of the generalized logarithmic series distribution. *J. Indian Soc. Agricultural Statist.,* **30**, 1-10.

Johnson N. L., Kemp A. W. and Kotz S. (2005): *Univariate Discrete Distributions*, Wiley, New York.

Khang T. F. and Ong S. H. (2007): A new generalization of the logarithmic distribution arising from the inverse trinomial distribution. *Comm. Statist. Theory Methods,* **36**, 3 – 21.

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

Ong S. H. (2000): On a generalization of the log-series distribution. *J. Appl. Statist. Sci.*, **10**(**1**), 77- 88.

Puig P. (2003): Characterizing additively closed discrete models by a property of their MLEs, with an application to generalized Hermite distribution. *J. Amer. Statist. Assoc.,* **98**, 687-692.

Rao C. R. (1973): *Linear Statistical Inference and its Applications,* John Wiley, New York.

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

Thripathi R. C. and Gupta, R.C. (1985): A generalization of the log-series distribution. *Communications in Statistics-Theory and Methods,* **14**, 1779-1799.

Tripathi R. C. and Gupta R. C. (1988): Another generalization of the logarithmic series and the geometric distribution. *Comm. Statist. Theory Methods*, **17,** 1541-1547.

Xekalaki E., and Panaretos J. (1989): On some distribution arising in inverse cluster sampling. *Comm. Statist. Theory Methods,* **18**, 355-366.

Received : 19.12.2012

Revised : 12.08.2013

C. Satheesh Kumar and A. Riyaz

Department of Statistics
University of Kerala
Trivandrum-India
e-mail:drcsatheeshkumar@gmail.com
e-mail: riyazstatoyour@gmail.com