ISSN 1684-8403
Journal of Statistics
Volume 16, 2009, pp. 45-58

Numerical Maximum Likelihood Estimation for the g-and-k Distribution Using Ranked Set Sample

Md. Anower Hossain[1] and Syed Shahadat Hossain[2]

Abstract

The purpose of this paper is to find the estimate of the parameters of $g-and-k$ distribution for ranked set sample by numerically maximizing the likelihood function. The estimates named as numerical maximum likelihood estimate, and corresponding mean square error and relative efficiency compared to simple random sampling are computed using a computer simulation. Ranked set sampling is seen to perform better than the usual simple random sampling method in terms of the efficiency and it is at least as precise as simple random sample. It is found that numerical maximum likelihood estimate using ranked set sample is more efficient than that of simple random sample.

Keywords

Quantile function distribution, Ranked Set Sampling (RSS), The g-and-k Distribution, Numerical Maximum Likelihood Estimation (NMLE)

1. Introduction

Ranked set sampling (RSS) (see, for example, McIntyre, 1952; Takahasi and Wakimoto, 1968) can be used in many ecological and agricultural studies where the measurement of each unit is laborious and expensive, but several units can easily be arranged in order of magnitude without requiring the actual measurement. The field of research related to the RSS has recently become increasingly important. In the last few years numerous developments have been

[1]Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka-1000, Bangladesh
Email: anower @ isrt.ac.bd
[2] Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka-1000, Bangladesh
Email: shahadat @ isrt.ac.bd

made in this field. For example, we refer to Dell and Clutter (1972), Abu-Dayyeh and Muttlak (1996), Ni Chuiv and Sinha (1998) and Hossain (1999). RSS data is seen to perform better than SRS data in estimation of population mean as well as in testing simple hypothesis (see, for example, Hossain, 1999).

By transforming the skewness and kurtosis parameter of standard normal distribution, Tukey (1977) invented the $g-and-h$ distributional family specified by its quantile function. Martinez and Iglewicz (1984) developed the notation of this family. MacGillivray (1992) proposed, based on Johnson family (formed by transforming the normal) and considering results from Tukey (1977) and Martinez and Iglewicz (1984) work, a suggestion for general transformation families with more interpretable shape parameters. Based on this suggestion, Haynes et al. (1997a, b) developed and used $g-and-k$ distribution. Rayner and MacGillivray (2002) have obtained the numerical maximum likelihood estimation for the parameters of $g-and-k$ and generalized $g-and-h$ distribution. RSS data give more power of a test than that of SRS data in hypothesis testing about parameters of $g-and-k$ distribution (see, for example, Hossain, 2007).Using these families of distributions it is possible to consider a wide variety of distribution shapes. It is convenient to use maximum likelihood estimation procedure with analytically expressed likelihoods, even if it requires computational work to maximize. Since the likelihood can be expressed only in terms of quantile function, likelihood procedure for $g-and-k$ distribution must be performed completely numerically.
In this paper, an attempt has been made to utilize RSS data instead of Simple Random Sample (SRS) data for estimating parameters of $g-and-k$ distribution. The likelihood function of $g-and-k$ distribution using RSS data is presented and maximizing this likelihood function numerically within the parameter space and the mean squared error (MLE) of the parameters are demonstrated using a simulation. The MSE is computed and the relative efficiency (RE) compared with SRS are presented.

2. Selection of Ranked Set Sample

The basic idea of the ranked set sampling is to partition the identified sample units randomly into small sets and then to rank the elements within each set according to the characteristic of interest. Then, based on the ranking, exactly one element of each ranked set is chosen as the selected sample element for quantification. To obtain a usual RSS of size pq by the usual RSS method, the following steps are required to be carried out:

1. p sets of p elements are drawn randomly;
2. the elements of each set are ordered by visual inspection or by some other means not requiring actual measurement;
3. the i^{th} smallest of the i^{th} set ($i=1,2,3,\cdots,p$) is drawn and measured; and
4. steps (1), (2) and (3) are repeated q times.

3. Quantile Distribution

Taking the inverse of the cumulative distribution function (cdf), $u = F_X(x|\vec{\theta})$, for a continuous random variable X, the quantile function is defined as

$$x = Q_X(u|\vec{\theta}) = F_X^{-1}(u|\vec{\theta})$$

where $\vec{\theta}$ is a vector of parameters. We define $u = F_X(x|\vec{\theta})$ as the depth corresponding to x, the data value.

Distributions differ in the distributional shape, independent of location and scale. Using a linear transformation of the quantile function location and scale parameters are introduced. Location, scale and dependence on some arbitrary base distribution give

$$x = Q_X(u|\vec{\theta}) = A + B.R_X(z_u|\vec{\gamma}),$$

where u^{th} quantile of the base-distribution denoted by z_u, which is parameter less.

The vector γ contains all parameters in vector θ excluding the location and scale parameters A and B, respectively. To introduce the effects of the shape (not location and scale) parameters γ, the “base-distribution” can be thought of as the distribution that is “bent” by the function $R_X(z|\vec{\gamma})$. The standard normal distribution is taken as base distribution here to be the standardized normal distribution.

4. The $g-and-k$ Distribution

In terms of its quantile function, the $g-and-k$ distribution is defined as

$$Q_X\,(u|A,B,g,k)=A+B\,z_u\left(1+c\,\frac{1-e^{-g\,z_u}}{1+e^{-g\,z_u}}\right)\left(1+z_u^2\right)^k \tag{4.1}$$

where A is the location parameter and B scale parameters, g is the skewness parameter and $k>-\frac{1}{2}$ kurtosis parameter of the distribution, $z_u=\varphi^{-1}(u)$ is the u^{th} standard normal quantile, and c is a constant chosen to help produce proper distribution.

In $g-and-k$ distribution the sign of skewness parameter g indicates the direction of skewness: $g<0$ indicates the distribution is skewed to the left, and $g>0$ indicates skewness to the right, and $g=0$ indicates the symmetry of the distribution. Increasing/decreasing the absolute value of g increases/decreases the skewness in the indicated direction.

The kurtosis parameter k behaves similarly. Increasing k increases the level of kurtosis and vice versa and $k=0$ corresponds to no kurtosis added to the standard normal base distribution. However, this distribution can represents less kurtosis than the standard normal distribution when $\frac{1}{2}<k<0$. And c is the value of the 'overall asymmetry' and we use $c=0.83$ through this paper. Fig. 1 can demonstrate the change of skewness and kurtosis for change in the parameters g and k. The distribution therefore gives the opportunity to investigate any process dependent on shape of the distribution under varying degrees of asymmetry and flatness.

The density function of the $g-and-k$ distribution is given by

$$f_X\,(x|\vec{\theta})=\begin{cases}\dfrac{1}{Q_X^{/}\,((Q_X^{-1}(x\,|\vec{\gamma}))|\vec{\gamma})}\ , & x\in S\\ 0\ , & otherwise\end{cases}$$

where S is the entire real axis, Q_X is defined in (1), $Q_X^{/}$ and Q_X^{-1} are, respectively, the first derivative and inverse function of Q_X, and $\vec{\gamma}$ is the vector of parameters.

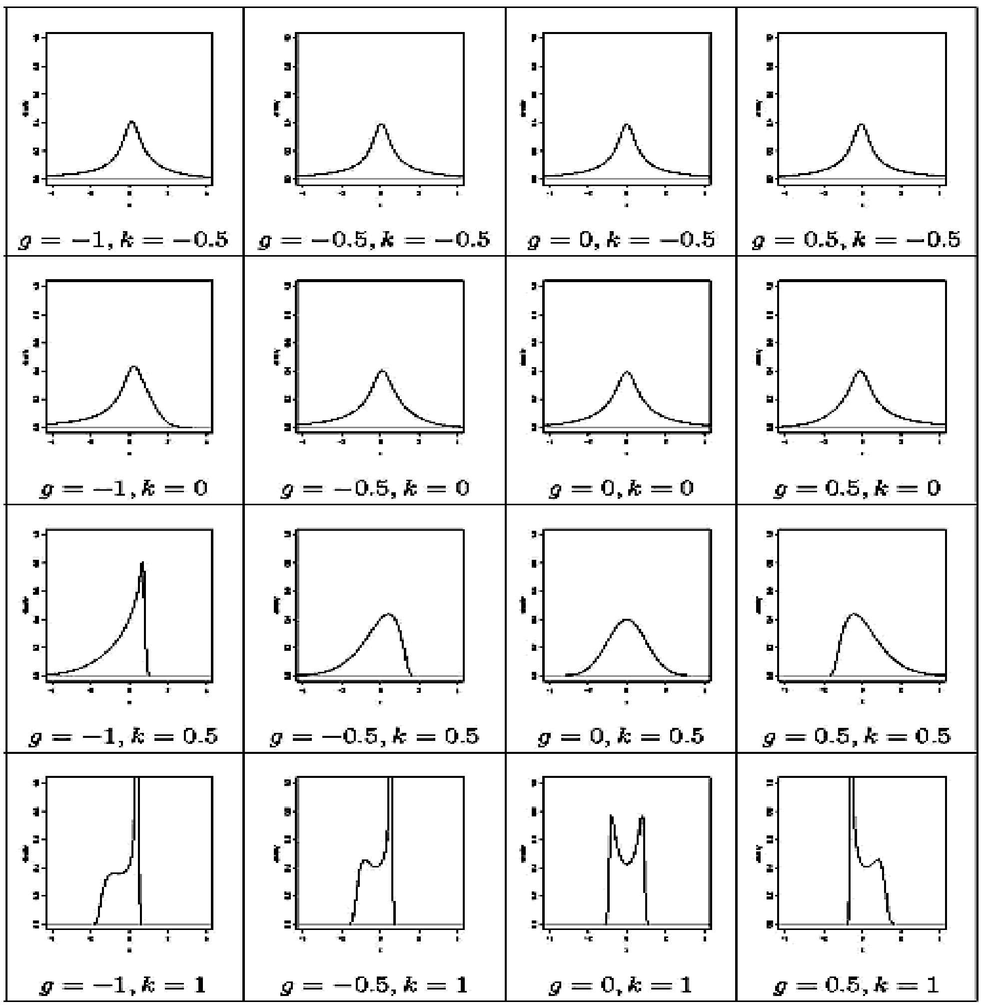

Fig. 1: Density curves of $g-and-k$ distribution for different combinations of g and k

5. Maximum Likelihood Estimation

The most popular method of parameter estimation, maximum likelihood estimation (MLE), must be performed completely numerically for $g-and-k$ distribution and indeed all quantile distributions which do not have a closed form inverse.

Let $X_1, X_2, \cdots, X_n$ be a ranked set sample of size n from a $g-and-k$ distribution with parameter vector $\gamma=(A, B, g, k)$. The observations in the ranked set sample are independent, but not identically distributed. That is, each observation in a ranked set sample has distinct distribution.

6. Computing the Likelihood

Define Ω as the space of parameter vectors corresponding to quantile distributions that are theoretically completely proper. For improper vectors $\vec{\gamma} \notin \Omega$ the likelihood is taken as zero. We will work with the log-likelihood rather than the likelihood, that is

$$l(\vec{\gamma}|x_1, x_2, \cdots, x_n)=\log L(\vec{\gamma})$$

$$=\sum_{i=1}^{n}(i-1)\log F(x_i|\vec{\gamma})+\sum_{i=1}^{n}(n-i)\log[1-F(x_i|\vec{\gamma})]+\sum_{i=1}^{n}\log f(x_i|\vec{\gamma}), \quad \vec{\gamma}\in\Omega$$

To calculate this log likelihood $l(\vec{\gamma}|x_1, x_2, \cdots, x_n)$ for a given ranked set sample $x_1, x_2, \cdots, x_n$ at a point $\vec{\gamma}\in\Omega$ requires the following steps:

1. For each data point x_i , obtain the sample quantile u. Also find z_u, the u^{th} quantile of the standard normal distribution. Then equating the first four sample quantile and corresponding population quantile , we get a system of nonlinear equations of four unknowns A, B, g and k:

$$A+Bz_{u_1}\left(1+c\frac{1-e^{-g\,z_{u_1}}}{1+e^{-g\,z_{u_1}}}\right)\left(1+z_{u_1}^2\right)^k=x_1$$

$$A+Bz_{u_2}\left(1+c\frac{1-e^{-g\,z_{u_2}}}{1+e^{-g\,z_{u_2}}}\right)\left(1+z_{u_2}^2\right)^k=x_2$$

$$A+Bz_{u_3}\left(1+c\,\frac{1-e^{-g\,z_{u3}}}{1+e^{-g\,z_{u3}}}\right)\left(1+z_{u_3}^2\right)^k=x_3$$

$$A+Bz_{u_4}\left(1+c\,\frac{1-e^{-g\,z_{u4}}}{1+e^{-g\,z_{u4}}}\right)\left(1+z_{u_4}^2\right)^k=x_4$$

where z_{u_i} is the u_i^{th} quantile of the standard normal distribution corresponding to the i^{th} observation x_i. The solution of this system of nonlinear equations for A, B, g and k are used as the initial estimates.

2. For each data point x_i, numerically solve the equation $x_i = Q_X(u_i|\vec{\gamma})$ to find the depths of each data point, $u_i = Q_i^{-1}(x_i|\vec{\gamma})$.
3. For each obtained depth u_i, calculate $Q_X^{/}(u|\vec{\gamma})$, where $Q_X^{/}(u|\vec{\gamma}) = \dfrac{\partial Q_X(u|\vec{\gamma})}{\partial u}$.
4. Finally, since $u_i = Q_i^{-1}(x_i|\vec{\gamma}) = F(x_i|\vec{\gamma})$ and $f(x_i) = \dfrac{1}{Q_X^{/}(u_i|\vec{\gamma})}$, calculate the log-likelihood function.

Since the second step requires z_{u_i} rather than u_i , the $g-and-k$ distribution expressed in z_u rather than u. Computation of step two and three is very straightforward. This is a robust numerical procedure because of the properties of a proper quantile distribution function.

7. Maximizing the Likelihood

In this paper numerical likelihood maximization was carried out on the log-likelihood using the MATLAB implementation of the Nelder-Mead simplex algorithm (see Press et al., 1993). It has the great advantage of not requiring derivative information about the log-likelihood.

8. Simulation Study for NMLE

Using Nelder-Mead simplex method, we obtain the NMLE of the parameters A, B, g and k of the $g-and-k$ distribution. Then we simulate these results for $n=5000$ times, i.e., we draw ranked set sample for different size in 5000 times and get the estimates of the parameters of our desired parameters using the method described earlier. Average of these 5000 estimates for each parameter is then calculated and named it by numerical maximum likelihood estimate (NMLE) of the parameters. In this paper, for fixed $A=0$ and $B=1$, we consider the combinations of (g, k) as (0, 0), (0, 0.5), (0.5, 0.5), (1, 0) and (-0.5, 0). The results of the simulations are presented in Table 1 through Table 5.

Table 1: NMLE of the parameters of $g-and-k$ distribution for $A=0$, $B=1$, $g=0$ and $k=0$

Sample size (n)	Method		A	B	g	k
			0	**1**	**0**	**0**
4	RSS	NMLE	0.20	1.32	0.24	-0.11
		√(MSE)	0.0532	0.289	0.0178	0.0216
	SRS	NMLE	0.27	1.21	0.28	-0.12
		√(MSE)	0.0851	0.423	0.0261	0.034
	RE		**2.55**	**2.23**	**2.15**	**2.47**
5	RSS	NMLE	0.24	1.30	0.21	-0.15
		√(MSE)	0.0491	0.2075	0.0161	0.0209
	SRS	NMLE	0.22	1.24	0.25	-0.19
		√(MSE)	0.0812	0.330	0.0250	0.033
	RE		**2.73**	**2.52**	**2.41**	**2.49**
6	RSS	NMLE	0.21	1.29	0.25	-0.17
		√(MSE)	0.0480	0.1898	0.0150	0.0195
	SRS	NMLE	0.22	1.21	0.19	-0.22
		√(MSE)	0.0798	0.310	0.0234	0.031
	RE		**2.76**	**2.96**	**2.43**	**2.52**
7	RSS	NMLE	0.27	1.28	0.18	-0.10
		√(MSE)	0.0455	0.1801	0.0140	0.0179
	SRS	NMLE	0.29	1.34	0.18	-0.18
		√(MSE)	0.0759	0.301	0.0224	0.0299
	RE		**2.78**	**2.79**	**2.56**	**2.79**

Table 2: NMLE of the parameters of $g-and-k$ distribution for $A=0$, $B=1$, $g=0$ and $k=0.5$

Sample size (n)	Method		A	B	g	k
			0	**1**	**0**	**0.5**
4	RSS	NMLE	0.25	1.38	0.19	0.35
		√(MSE)	0.0512	0.279	0.0176	0.0211
	SRS	NMLE	0.24	1.25	0.31	0.41
		√(MSE)	0.0794	0.398	0.0257	0.0310
	RE		**2.40**	**2.03**	**2.13**	**2.21**
5	RSS	NMLE	0.24	1.34	0.27	0.39
		√(MSE)	0.0482	0.201	0.0154	0.020
	SRS	NMLE	0.22	1.31	0.29	0.42
		√(MSE)	0.0776	0.312	0.0232	0.030
	RE		**2.59**	**2.20**	**2.26**	**2.25**
6	RSS	NMLE	0.26	1.27	0.42	0.44
		√(MSE)	0.0476	0.180	0.0130	0.018
	SRS	NMLE	0.21	1.25	0.31	0.37
		√(MSE)	0.0781	0.279	0.0201	0.028
	RE		**2.69**	**2.40**	**2.39**	**2.41**
7	RSS	NMLE	0.35	1.38	0.28	0.43
		√(MSE)	0.0464	0.155	0.0131	0.0164
	SRS	NMLE	0.35	1.34	0.21	0.42
		√(MSE)	0.0770	0.252	0.02031	0.0261
	RE		**2.75**	**2.64**	**2.40**	**2.53**

Table 3: NMLE of the parameters of $g-and-k$ distribution for $A=0$, $B=1$, $g=0.5$ and $k=0.5$

Sample size (n)	Method		A	B	g	k
			0	1	0.5	0.5
4	RSS	NMLE	0.20	1.20	0.50	0.68
		√(MSE)	0.0612	0.312	0.0175	0.0215
	SRS	NMLE	0.24	1.26	0.65	0.45
		√(MSE)	0.0824	0.457	0.0264	0.0313
	RE		**1.81**	**2.41**	**2.20**	**2.11**
5	RSS	NMLE	0.27	1.21	0.54	0.69
		√(MSE)	0.0542	0.279	0.0168	0.020
	SRS	NMLE	0.25	1.54	0.39	0.57
		√(MSE)	0.0789	0.421	0.0255	0.03
	RE		**2.11**	**2.27**	**2.30**	**2.25**
6	RSS	NMLE	0.29	1.09	0.70	0.48
		√(MSE)	0.0510	0.245	0.0145	0.017
	SRS	NMLE	0.29	1.25	0.57	0.55
		√(MSE)	0.0770	0.386	0.0221	0.027
	RE		**2.27**	**2.48**	**2.32**	**2.52**
7	RSS	NMLE	0.41	1.21	0.35	0.39
		√(MSE)	0.0469	0.190	0.0135	0.016
	SRS	NMLE	0.31	1.24	0.65	0.68
		√(MSE)	0.0758	0.310	0.0201	0.0256
	RE		**2.61**	**2.66**	**2.41**	**2.56**

Table 4: NMLE of the parameters of $g-and-k$ distribution for $A=0,\ B=1,\ g=1$ and $k=0$.

Sample size (n)	Method		A	B	g	k
			0	**1**	**1**	**0**
4	RSS	NMLE	0.24	1.24	1.22	-0.25
		√(MSE)	0.0612	0.312	0.0175	0.0215
	SRS	NMLE	0.28	1.19	1.38	-0.30
		√(MSE)	0.0824	0.457	0.0264	0.0313
	RE		**1.81**	**2.14**	**2.27**	**2.11**
5	RSS	NMLE	0.29	1.26	1.24	-0.11
		√(MSE)	0.0542	0.28	0.0155	0.000253
	SRS	NMLE	0.20	1.26	1.27	-0.14
		√(MSE)	0.0789	0.421	0.0245	0.029
	RE		**2.12**	**2.26**	**2.49**	**2.32**
6	RSS	NMLE	0.21	1.29	1.15	1.21
		√(MSE)	0.0510	0.235	0.0135	0.017
	SRS	NMLE	0.21	1.13	1.09	-0.12
		√(MSE)	0.0750	0.354	0.0221	0.026
	RE		**2.16**	**2.26**	**2.67**	**2.33**
7	RSS	NMLE	0.35	1.31	1.25	-0.16
		√(MSE)	0.0452	0.197	0.0111	0.014
	SRS	NMLE	0.38	1.19	1.26	-0.15
		√(MSE)	0.069	0.298	0.0182	0.0218
	RE		**2.33**	**2.28**	**2.68**	**2.42**

Table 5: NMLE of the parameters of $g-and-k$ distribution for $A=0,\ B=1,\ g=-0.5$ and $k=0$.

Sample size (n)	Method		A **0**	B **1**	g **-0.5**	k **0**
4	RSS	NMLE	0.28	1.29	-0.21	-0.19
		√(MSE)	0.0514	0.326	0.0213	0.0222
	SRS	NMLE	0.29	1.25	-0.25	-0.21
		√(MSE)	0.0750	0.482	0.0314	0.0340
	RE		**2.12**	**2.18**	**2.17**	**2.34**
5	RSS	NMLE	0.24	1.31	1.27	1.40
		√(MSE)	0.0514	0.277	0.0177	0.019
	SRS	NMLE	0.31	1.19	-0.18	-0.12
		√(MSE)	0.0741	0.424	0.0276	0.0297
	RE		**2.18**	**2.34**	**2.43**	**2.44**
6	RSS	NMLE	0.29	1.25	-0.17	-0.11
		√(MSE)	0.0480	0.246	0.0145	0.016
	SRS	NMLE	0.28	1.35	-0.19	-0.14
		√(MSE)	0.0724	0.379	0.0234	0.0252
	RE		**2.27**	**2.37**	**2.60**	**2.48**
7	RSS	NMLE	0.34	1.25	-0.21	-0.16
		√(MSE)	0.0447	0.216	0.0131	0.014
	SRS	NMLE	0.29	1.21	-0.22	-0.15
		√(MSE)	0.0687	0.343	0.022	0.0221
	RE		**2.36**	**2.52**	**2.80**	**2.49**

9. Discussion and Conclusion

From Table 1, it can be observed that as sample size increases the relative efficiency (RE) of RSS estimates increases compared to SRS estimates. Specifically, for sample size four, the RE of the parameter A is 2.55, and it is 2.73, 2.76 and 2.78 for sample size 5, 6 and 7, respectively (Table 1). The same result is observed in the other Tables. From Table 1 and Table 2, we observed that for fixed A, B, g and sample size n, as the value of kurtosis parameter k increases the RE of RSS estimates decreases but RSS estimate still remain more efficient than SRS estimates. Comparing Table 1 and Table 3, it can be observed that for fixed A, B and sample size n, as we increase skewness parameter g and kurtosis parameter k simultaneously the RE of RSS estimates decreases but still remain above unity .

In practice, most of the analysis is based on the normality assumption, however the data may not be always from normal population. In this situation by considering the data from $g-and-k$ distribution, we can easily estimate the amount of departure from normality.

References

1. Abu-Dayyeh, W. and Muttlak, H. A. (1996). Using ranked set sampling for hypothesis tests on the scale parameters of exponential and uniform distributions. *Pakistan Journal of Statistics*, **12**, 131-138.
2. Dell, T. R. and Clutter, J.L. (1972). Ranked set sampling theory with order statistics background. *Biometrics*, 28,545-553.
3. Haynes, M. A., Gatton, M.L. and Mengersen, K.L. (1997a). Generalized control charts for nonnormal data. School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia, Technical Report Number 97/4.
4. Haynes, M. A., MacGillivray H. L. and Mengersen, K. L. (1997b). Robustness of ranking and selection rules using generalized g-and-k distributions. *Journal of Statistical Planning and Inference*, **65**, 45-66.
5. Hossain, M. A. (2007). *Numerical maximum likelihood estimation and numerical likelihood ratio test for ranked set sample data: A g-and-k distribution approach.* Unpublished M.S. thesis, Institute of Statistical Research and Training, University of Dhaka.
6. Hossain, S. S. (1999). *The Efficiency of Improved Ranked Set Sampling Methods*. Ph. D. Thesis, Deakin University, Victoria, Australia.
7. MacGillivray, H. L. (1992). Shape properties of the generalized $g-and-h$ and Johnson families, *Communications in Statistics- Theory and Methods*, **21(5)**, 1233-1250.
8. Martinez, J. and Iglewicz, B. (1984). Some properties of the Tukey g-and-h family of distribution. *Communications in Statistics-Theory and Methods*, **13(3)**, 353-369.
9. McIntyre, G. A. (1952). A method for unbiased selective sampling, using ranked sets, *Australian Journal of Agricultural Rsesearch*, **3**, 385-390.

10. NI Chuiv, N. and Sinha, B. K. (1998). On some aspect of ranked set sampling. *Handbook of Statistics*, **17**, N. Balakrishnan, and C.R. Rao, eds., Elsevier Science Publisher.
11. Press, W.H., Teukolsky S. A., Vetterling W. T. and Flannery, B. P. (1993). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge.
12. Rayner, G. D. and MacGillivray, H. L. (2002). Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distribution. *Statistics and Computing*, **12**, 57-75.
13. Takahasi, K. and Wakimoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematics*, **20**, 1-31.
14. Tukey J. W. (1977). Modern techniques in data analysis. In: *NSF-Sponsored Regional Research Conference*, Southeastern Massachusetts University, North Dartmouth, Massachusetts.