Thailand Statistician
July 2013; 11(2): 173-184
http://statassoc.or.th
Contributed paper

Application of Ranked Set Sampling Design in Environmental Investigations for Real Data Set

Neeraj Tiwari and Girja S. Pandey*

Department of Statistics, Kumaun University, S.S.J. Campus Almora, India.

*Corresponding author; e-mail: pandeygirja@yahoo.co.in

Received: 15 March 2013

Accepted: 8 July 2013

Abstract

Cost-effective sampling methods are of major concern in Statistics, especially when the measurement of the characteristic of interest is costly and/or time-consuming. In the early 1950's, for effectively estimating the yield of pasture in Australia, McIntyre (1952) proposed a sampling method which was later known as Ranked Set Sampling (RSS). Ranked set sampling can provide an efficient basis for estimating parameters of environmental variables, particularly when sampling cost is intrinsically high. The use of ranked set sampling for estimation of the mean of lognormal distribution with known coefficient of variation has also been examined by Shen (1994). Barnett (1999) suggested some specific attention to lognormal distribution encountered with environmental problems and achieved highest possible efficiency (e.g. in estimation). In the present paper we have studied in brief the application of ranked set sampling design for environmental investigations has been demonstrated with the help of a real data set. Three different estimators or forms of estimators, viz., the sample mean, the ranked set sample mean and the ranked set best linear unbiased estimators (ranked set BLUEs) which take the form of linear combinations have been considered. These three estimators illustrate the design effects for the particular case of lognormal distribution.

Keywords: Lognormal distribution, ranked set sample mean, ranked set best linear estimator, sample mean.

1. Introduction

Different type of problems arises when we seek to collect data on environmental investigation. In most of the situations it is quite difficult to obtain a simple random sample due to non-availability of sampling frame. It is also problematic to unrestrictedly choose where to take observations according to a prescribed design. Ranked set sampling can provide an efficient basis for estimating parameters of environmental variables. Ranked set sampling can provide observational economy under very special circumstances- namely, when sampling units can be easily and inexpensively gathered and ranked among themselves, but it is quite expensive to measure them accurately. This situation arises very commonly in agriculture and forestry. For example, it is easy and cheap to judge approximately by inspection about which of the several trees contains the largest volume of wood, which the next largest, and so on, whereas it is much more expensive to actually measure the amount of wood in each. Similar circumstances arise in some environmental applications. For example, consider the problem of assessing the status of a hazardous waste site (i.e., determining if a site has toxic chemicals in excess of a set standard). We often know a great deal about the sites from records, photos and physical characteristics. This knowledge will allow us to rank the areas from which we will sample in terms of high to low levels of toxic pollution. This would limit the number of expensive samples necessary to assess the status of the hazardous waste site.

The definition of ranked set sampling has also been generalized to include methods that select one order statistic from each sample of size m, but not necessarily one of each rank. This has allowed improvement in two ways. First, one can do even better in estimation of location and scale parameters by prudent selection of order statistic. For example, if the characteristic of interest were known to have a symmetric distribution in the population, the selection and averaging of all sample medians would be preferable to the averaging of all order statistics. Second, some parameters which do not enjoy improved estimation from samples selected by the classical RSS scheme, such as the correlation coefficient from a bivariate normal distribution, can be estimated more precisely using generalized RSS methods. The disadvantage of these methods is that they require assumptions about the form of the underlying distribution, and the RSS estimators suffer biases if the assumed conditions are not met. In addition, methods that require these generalized sampling schemes usually do not allow imperfect ranking.

Ranked set sampling can provide an efficient basis for estimating parameters of environmental variables, particularly when sampling cost are intrinsically high. Various ranked set sampling estimators were considered by different researchers for the population mean and contrasted in terms of their efficiencies and usefulness, with special concern for the sample design considerations. In many environmental situations, it has been observed that the data obtained from the site generally follows a distribution with heavy right tail, such as a lognormal distribution. The average concentration of an air pollutant such as Sulfur-di-Oxide, Carbon monoxide, Nitrous Oxides, etc., is approximately lognormally distributed (see, Larsen [1]). When the parameters of such distributions are completely unknown, the researchers face a great challenge to estimate these parameters with minimum cost and maximum precision. The use of ranked set sampling for estimation of the mean of lognormal distribution with known coefficient of variation has also been examined by Shen [2]. Survey data from different sources can give a basic idea about the shape and scale parameters of the distribution from which the data has been taken. Barnett [3] suggested some specific attention to lognormal distribution encountered with environmental problems and achieved highest possible efficiency (e.g. in estimation). He assumed that a random variable X with distribution function in the form $F\{(x-\mu/\sigma)\}$ and considered the data level Y of potassium contamination at 40 locations over a given region (given by Walter & Oliver), where $X=Y^2$ and $\log\left(\frac{X-\mu}{\sigma}\right)\sim N(0,1)$, for $(X>\mu)$. He found that a lognormal distribution fitted well the squares of these values by normal probability curve. Finally he has given five different estimators of the mean of a random variable X. The general RSS scheme proposed by Wang et al. [4] takes more than one units in a ranked set with select pre-specified ranks for the full measurement. Sengupta and Mukhuti [5] proposed some unbiased estimators of the variance of exponential distribution. Jemain et al. [6] suggested multistage median ranked set sampling (MMRSS) method for estimating the population mean. Frey [7] demonstrated a new imperfect ranking model for ranked set sampling. Some variations of ranked set sampling studied by Jamain et al. [8]. Ozturk [9] had proposed a inference in the presence of ranking error in ranked set sampling. Baklizi

[10] described empirical likelihood intervals for the population mean and quantiles based on balanced RSS. Liu and Lin [11] studied the problem of empirical likelihood for balanced ranked set sampled data. Estimation of population variance using ranked set sampling with auxiliary variable was studied by Hadhrami [12].

In this paper we assume a random variable X which follows a lognormal distribution with distribution function $F(x)$. We have considered a real data set of annual rainfall at 36 metrological sub-divisions of India, taken from Central Statistical Organisation, Ministry of Statistics and Programme Implementation, Government of India in year 2006 and obtained the values of the three estimators given below :

(1) The sample mean.

(2) The ranked set sample mean.

(3) The ranked set best linear unbiased estimators (ranked set BLUEs).

In Section 2, we discuss in brief the expressions of variance of three estimators of the mean of a random variable X. Section 3, describes the design effects for the particular case of lognormal distribution with the help of consider a real example and demonstrates the utility of the ranked set sampling designs with the help of normal probability curve and comparison of the variance of three estimators for sample size n = 2, 3,…,10. Section 4, concludes the finding of the present paper.

2. Variances of three different estimators according to weight

Let X is a random variable having distribution function $F\{(x-\mu)/\sigma\}$. Let X is symmetric with mean μ and $E(X)$ is a linear combination of μ and σ. For a random sample $x_1, x_2, ..., x_n$, the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is unbiased for $\mu_X = E(X)$ with variance σ_X^2/n where $\sigma_X^2 = Var(X)$.

Considering $X_{(1)}, X_{(2)}, ..., X_{(n),}$, the order statistics of a sample of size n and let $U_{(i)} = (X_{(i)} - \mu)/\sigma$. Then we define $\alpha_i = E(U_{(i)})$ and $\beta_i = Var(U_{(i)})$. Suppose the ranked set sample $x_{(1)1}, x_{(2)2}, ..., x_{(n)n}$, is obtained in a usual manner as the set of smallest, second smallest, up to largest observed values in n conceptual samples $x_{i1}, x_{i2}, ..., x_{in}$, (i =1, 2,….n) under the assumption that correct ordering has taken place.

Consider the use of the ranked set sample for estimation of μ and σ. The usual estimator of E(X) is the ranked set sample mean

$$\overline{X}_{RSS} = \frac{1}{n}\sum_{i=1}^{n} X_{(i)i} , \tag{1}$$

which is an unbiased estimator with variance $\sigma^2 \sum \beta_i / n^2$. The variance of $\overline{X}_{RSS}$ for symmetrical X can be re-expressed as $\left(1-\sum \alpha_i^2 / n\right)\sigma^2 / n$. In general we have

$$\begin{aligned} Var\left(\overline{X}_{RSS}\right) &= Var(\overline{X}_{SRS}) - \frac{1}{n^2}\sigma^2 \sum_{i=1}^{n}\left(\alpha_i - \overline{\alpha}\right)^2 \\ &\leq Var(\overline{X}_{SRS}) \\ &= \frac{\sigma_X^2}{n} - \frac{1}{n^2}\sigma^2 \sum_{i=1}^{n}\left(\alpha_i - \overline{\alpha}\right)^2 \end{aligned} \tag{2}$$

where $\overline{\alpha} = \frac{1}{n}\sum_{i=1}^{n} \alpha_i$.

This shows that RSS will always result in at least as precise estimate as obtained from SRS. The Relative Efficiency of RSS with respect to SRS is given by :

$$e\left(\overline{X}_{RSS,} \overline{X}_{SRS}\right) = \frac{Var\left(\overline{X}_{SRS}\right)}{Var\left(\overline{X}_{RSS}\right)} = \left\{1 - \sum_{i=1}^{n}\left(\alpha_i - \overline{\alpha}\right)^2 \sigma^2 / n\sigma_X^2\right\}^{-1} \tag{3}$$

The optimally chosen weights provide a gain in relative precision of estimation of E(X). Barnett and Moore [13] obtained the ranked set best linear unbiased estimator (ranked set BLUEs) of μ and σ in $F\{(x-\mu)/\sigma\}$ and hence of E(X) in the case of lognormal distribution. The BLUEs of μ and σ take the forms,

$$\mu^* = \sum_{i=1}^{n} V_i X_{(i)i} \tag{4}$$

$$\sigma^* = \sum_{i=1}^{n} W_i X_{(i)i} \qquad (5)$$

where

$$V_i = \frac{1}{\beta_i}\left\{\sum_{j=1}^{n}\left(\frac{\alpha_j^2}{\beta_j}\right) - \alpha_i \sum_{j=1}^{n}\left(\frac{\alpha_i}{\beta_j}\right)\right\} \Big/ \Delta \qquad (6)$$

and

$$W_i = \frac{1}{\beta_i}\left\{\alpha_i \sum_{j=1}^{n}\left(\frac{1}{\beta_j}\right) - \sum_{j=1}^{n}\left(\frac{\alpha_i}{\beta_j}\right)\right\} \Big/ \Delta \qquad (7)$$

with

$$\Delta = \left(\sum_{i=1}^{n}\frac{\alpha_i^2}{\beta_i}\right)\left(\sum_{i=1}^{n}\frac{1}{\beta_i}\right) - \left[\left(\sum_{i=1}^{n}\frac{\alpha_i}{\beta_i}\right)\right]^2 \qquad (8)$$

The variances of μ^* and σ^* and covariance between μ^* & σ^* are :

$$Var(\mu^*) = \frac{\sigma^2 \sum_{i=1}^{n}\frac{\alpha_i^2}{\beta_i}}{\Delta} \qquad (9)$$

and

$$Var(\sigma^*) = \frac{\sigma^2 \sum_{i=1}^{n}\frac{1}{\beta_i}}{\Delta} \qquad (10)$$

Also the simplified form of $Cov(\mu^*, \sigma^*)$ is

$$Cov(\mu^*, \sigma^*) = \frac{-\sigma^2 \sum_{i=1}^{n}\frac{\alpha_i}{\beta_i}}{\Delta} \qquad (11)$$

For asymmetric X, we have $E(X) = \mu_X = \mu + \bar{\alpha}\sigma,$ since $E(U) = \bar{\alpha}$. Thus it is the linear combination $\mu + \bar{\alpha}\sigma$ which must be estimated if we are interested in μ_X , the mean of X. Its ranked set BLUE will be

$$\mu_X^* = \mu^* + \overline{\alpha}\sigma^* \tag{12}$$

Using the results of Barnett and Moore [13], we can express its variance in the form

$$Var\left(\mu_X^*\right) = Var\left(\mu^*\right) + \overline{\alpha}^2 Var\left(\sigma^*\right) + 2\overline{\alpha}Cov\left(\mu^*, \sigma^*\right)$$

$$= \frac{\sigma^2}{\Delta}\sum_{i=1}^{n}\frac{\left(\alpha_i - \overline{\alpha}\right)^2}{\beta_i} \tag{13}$$

By definition, this optimal linear estimator μ_X^* cannot be less than $\overline{X}_{RSS}$. The relative efficiency is

$$e\left(\mu_X^*, \overline{X}_{RSS}\right) = \Delta\sum_{i=1}^{n}\beta_i \Bigg/ \left\{n^2\sum_{i=1}^{n}\left[\left(\alpha_i - \overline{\alpha}\right)^2 / \beta_i\right]\right\}. \tag{14}$$

Thus we have considered three forms of the estimator of μ_X, namely those which give equal weights to unordered observations (the sample mean), equal weights to ordered observations (the ranked set mean) and optimal weights to ordered observations (the ranked set BLUE).

3. Performance of three different estimators in a real data set following a lognormal distribution

As discussed earlier, in environmental studies we commonly encounter positively skew data and in such cases the lognormal distribution often provides a reasonable model. We shall now examine whether the various distinctions reported above have important efficiency gain implications in these situations, especially since sampling costs can be high and we need to design the sampling and estimation procedures to keep these to a minimum.

We consider a real rainfall data of Table 1 on levels X of 36 Metrological Sub-divisions of India, taken from Central Statistical Organisation, Ministry of Statistics and Programme Implementation, Government of India in the year 2006.

Table 1. Levels X (Annual Rainfall) of 36 Metrological Sub-divisions of India in Millimeter.

S.No.	Meteorological sub-divisions	Actual rain fall (in millimeter)
1	Andaman and Nicobar Iselands	2,448
2	Arunachal Pradesh	2,108
3	Assam & Meghalaya	1,777
4	Nagaland, Manipur, Mizoram & Tripura	1,562
5	Sub Himalayan, West Bengal & Sikkim	2,305
6	Gengetic West Bengal	1,587
7	Orissa	1,810
8	Jharkhand	1,356
9	Bihar	1,001
10	East Uttar Pradesh	772
11	West Uttar Pradesh	510
12	Uttaranchal	1,265
13	Haryana, Delhi, and Chandigarh	377
14	Punjab	545
15	Himanchal Pradesh	896
16	Jammu & Kashmir	1,477
17	West Rajasthan	362
18	East Rajasthan	712
19	West Madhya Pradesh	1,141
20	East Madhya Pradesh	1,008
21	Gujarat Region	1,458
22	Saurashtra and Kutch	703
23	Konkan & Goa	3,379
24	Madhya Maharashtra	1,181
25	Marathwada	819
26	Vidarbha	1,277
27	Chhattisgarh	1,232
28	Coastal Andhra Pradesh	1,067
29	Telangana	1,045
30	Rayalaseema	609
31	Tamil Nadu and Pondicherry	912
32	Coastal Karnataka	3,866
33	North Interior Karnataka	628
34	South Interior Karnataka	951
35	Kerala	3,298
36	Lakshdweep	1,695

Source: Central Statistical Organisation, Ministry of Statistics and Programme Implementation, Government of India in year 2006.

We find that a Normal distribution fits well the log of these values as shown in the probability plot of Figure 1. The lognormal distribution takes various forms [14] but the one which seems particularly suited to the data of Table 1, and which we shall examine in detail, is that in which

$$\log(X) \sim N(0,\ 1),$$

i.e. after appropriate change of scale and location the logarithm of the variable of interest has a standard normal distribution (Figure 1).

This has distribution function of the form $F(x)$. It is readily confirm that, if

$$U = \left(\frac{X-\mu}{\sigma}\right),$$

then $E(U) = \sqrt{e}$ and $Var(U) = e(e-1)$.

Thus

$$E(X) = \mu_X = \mu + \sigma\sqrt{e}, \qquad Var(X) = \sigma_X^2 = \sigma^2 e(e-1) \tag{15}$$

The ranked set sample mean $\overline{X}_{RSS}$ is unbiased for $E(X) = \mu + \sigma\sqrt{e}$ with variance $\sigma^2 \sum \beta_i / n^2$ or $\left(e^2 - \sum \alpha_i^2 / n\right)\sigma^2 / n$, (from equation (2) with $\overline{\alpha} = \sqrt{e}$) where $\alpha_i = E\left(U_{(i)}\right)$ and $\beta_i = Var\left(U_{(i)}\right)$ for the lognormal distribution. Mean and variances of order statistics for this distribution have been studied by Gupta et al. [15] and can be used for efficiency comparison between $\overline{X}_{SRS}$, $\overline{X}_{RSS}$ and the optimal rank set BLUE, $\mu_X^* = \mu^* + \sigma^* \sqrt{e}$. The explicit form μ_X^* in terms of the α_i and β_i can be obtained from equation (4) to (8). From equation (13) we have

$$Var\left(\mu_X^*\right) = \frac{\sigma^2}{\Delta} \sum_{i=1}^{n} \frac{\left(\alpha_i - \sqrt{e}\right)^2}{\beta_i} \tag{16}$$

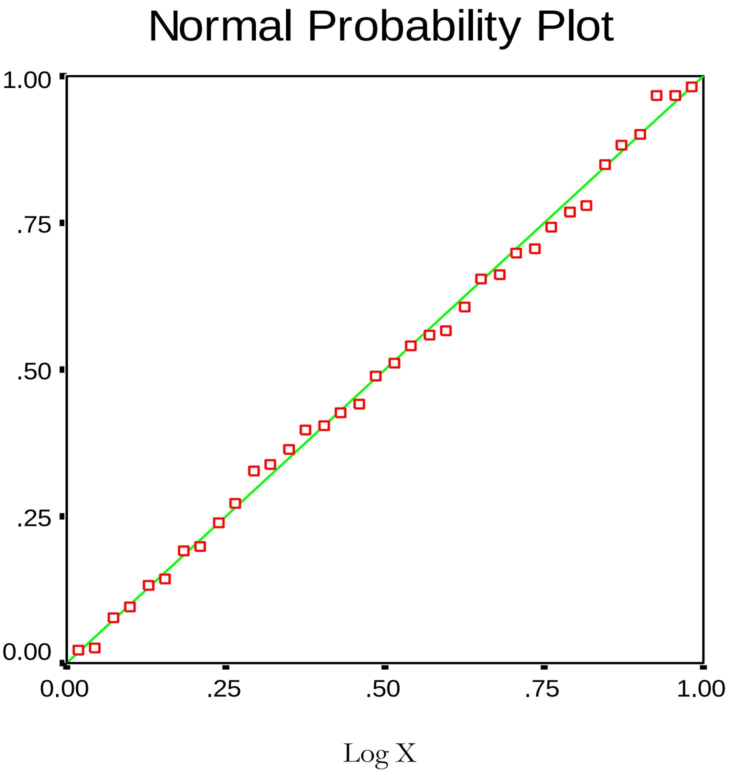

Figure 1. Normal Probability curve for log of Annual Rainfall in Millimeter (Mean = 3.0649406, Standard deviation = .24994622).

4. Conclusion

We compare the variances of three estimators (for lognormal distribution) $\overline{X}_{SRS}$, $\overline{X}_{RSS}$ and μ_X^* for different sample size $n = 2, 3, \ldots, 10$. $\overline{X}_{RSS}$ is more efficient than $\overline{X}_{SRS}$, the ranked set BLUE estimator μ_X^* performs best among these three estimators considered by us. It is evident from the last two columns of Table 2 that the relative efficiency depends on sample size and increases with the increase of sample size. Thus we conclude, for lognormal distribution, that the obvious design rule to follow is to choose the largest practicable ranked set samples within the overall sampling effort and use the ranked set BLUE.

Table 2. Variances and relative efficiencies of $\overline{X}_{SRS}$, $\overline{X}_{SRS}$ and μ_X^* for the lognormal distribution.

n	$Var(\overline{X}_{SRS})$	$Var\left(\overline{X}_{RSS}\right)$	$Var\left(\mu_X^*\right)$	$e\left(\overline{X}_{RSS,}\overline{X}_{SRS}\right)$	$e\left(\mu_X^*,\overline{X}_{RSS}\right)$
2	.030369	.027495	.007677	1.104537	3.581164
3	.020246	.017937	.003732	1.12875	4.805597
4	.015185	.013239	.00214	1.14987	6.185253
5	.012148	.010458	.001371	1.161552	7.627975
6	.010123	.008625	.000948	1.173628	9.099089
7	.008677	.007329	.000692	1.183907	10.59244
8	.007592	.006365	.000526	1.192828	12.09235
9	.006749	.005621	.000413	1.200688	13.59711
10	.006074	.005029	.000333	1.207679	15.10431

References

[1] Larsen, R.I. A new mathematical model of air pollution concentration averaging time and frequency, Journal of the air pollution central assoc. 1969; 19: 24-30.

[2] Shen, W.H. On estimation of a lognormal mean using a ranked set sample, Sankhya, 1994; 56: 323-333.

[3] Barnett V. Ranked set sample design for environmental investigations. Environmental and Ecological Statistics, University of Nottingham, UK, 1999; Vol. 6: 59-74.

[4] Wang, Y.G., Chen, Z., and Liu, J. General ranked set sampling with cost considerations, Biometrics, 2004; 60: 556-561.

[5] Sengupta, S., and Mukhuti S. Unbiased variance estimation in a simple exponential population using ranked set samples. Journal of Statistical Planning and Inference, 2006; 136: 1526-1553.

[6] Jemain, A., A.A. Amer, A., Ibrahim, K. Multisatage median ranked set sampling (MMRSS). Journal of Mathematics and Statistics, 2006.

[7] Frey J. New imperfect ranking models for ranked set sampling. J. Stat Plan Inf, 2007; 137: 1433-1445.

[8] Jemain, A., A.A. Amer, A., Ibrahim, K. Some variations of ranked set sampling. Electronic Journal of Applied Statistics Analysis, 2008; 1: 1-18.

[9] Ozturk, O. Inference in the presence of ranking error in ranked set sampling. Canadian Journal of Statistics, 2008.

[10] Baklizi, A. Empirical likelihood intervals for the population mean and quantiles based on balanced ranked set sample. Stat Methods Appl., 2009; 18: 483-505.

[11] Liu, T.Q. and Lin N. Empirical likelihood for balanced ranked-set sampled data. Science in China Series A: Mathematics, 2009.

[12] Hadhrami, S.A.A. Estimation of the population variance using ranked set sampling with auxiliary variable. Int. J. Contemp. Math. Sciences, 2010; Vol.-5: No.-52; 2567-2576.

[13] Barnett, V. and Moore, K. Best linear unbiased estimates in ranked set sampling with particular reference to imperfect ordering. Journal of Applied Statistics, 1997; 24: 697-710.

[14] Johnson, N.L., Kotz, S. and Balakrishnan, N. Continuous Univariate Distribution. Wiley, New York 1995; Vol.-2.

[15] Gupta, S.S., McDonald, G.C. and Galarneau, D.I. Moments product moments and percentage points of the order statistics from the lognormal distribution for samples of size twenty and less, Sankhya B, 1974; 36: 230-260.