RANKED SET SAMPLING: AS A COST-EFFECTIVE AND MORE EFFICIENT DATA COLLECTION METHOD

Arun Kumar Sinha*
Dept. of Statistics, Central University of South Bihar
B I T Patna Campus, PO: B V College
Patna 800 014, Bihar, INDIA
(arunkrsinha@yahoo.com)

Abstract

Ranked set sampling (RSS) is a cost-effective sampling technique that induces stratification on the population through rank orders of samples. This, in turn, provides a more structured sample than a simple random sample does with the same sample size. This yields more efficient estimators of some parameters of interest. In this method a fairly large number of randomly identified sampling units are portioned into small subsets of the same size. The units of each subset are ranked separately with respect to the characteristic of interest without using their actual measurements. The measurements of the units with some specified ranks constitute a ranked set sample. In this paper we wish to discuss theory, methods and some recently reported applications of RSS to highlight its advantages. This method could be of some particular interest to those who look for a cost-effective and more efficient data collection technique for sampling and monitoring situations.

Keywords: Order statistics, Perfect and concomitant ranking, Relative precision and savings

1. Introduction

For adequately representing heterogeneous populations large sample sizes are required, but the budgetary constraints do not permit to have the desired number of measurements. One way out to deal with this impasse is to consider identification and acquisition of sampling units inexpensive as compared with their quantification. Ranked set sampling (RSS) is one such method to address the situation. It was proposed by McIntyre (1952) for estimating pasture yields In fact, the method is useful in situations in which measurements are difficult or costly to obtain, but the ranking of a subset of units with respect to the characteristics of interest is relatively economical and convenient. The method combines the control of simple random sampling (SRS) and the convenience of purposive sampling. Unlike stratified random sampling, which begins with presuming a stratification of the population, RSS accomplishes stratification through samples using easily available and low-cost outside information. It is, in fact, carried out through the ranking of randomly selected units by the field personnel. This process, in turn, exploits the experience and expertise of the field personnel for improving upon the efficiency of SRS. This also meets an objective of the total quality management (TQM), which suggests that everyone in a system should contribute to the improvement of its quality. One of the strengths of RSS is flexibility and model robustness regarding the nature of the auxiliary information needed for ranking. For achieving the cost-effectiveness the method presumes that the quantification rather than identification, acquisition and ranking of sampling units is the main component of the total sampling cost. See Mode et al. (1999) and Nahas et al. (2002) for a detailed discussion on cost-effectiveness of RSS. It has been used in many sampling and monitoring situations advantageously. Dell and Clutter (1972) discovered that the RSS estimator of a population mean remains unbiased, and is at least as efficient as the SRS estimator with the same number of quantifications even under the error in ranking. They pointed out that its performance would depend upon the characteristics of the population and also on the magnitude of the errors in ranking. Patil et al. (1994b), Wolfe (2004) and Sinha (2005 and 2014) presented a comprehensive review of the literature with some new results of interest. Sinha (2014) described its applications

in vegetation research, and Mode et al. (1999) discussed its relevance for ecological research. Apart from visual perceptions, ranking may be carried out on the basis of remotely sensed information, prior information, results of earlier sampling episodes, rank correlated covariates, expert-opinion/expert systems, etc., or some combinations of these approaches. In this paper an attempt is made to present a review of its theory, methods and applications to illustrate its cost-effectiveness and efficiency for data collection. Besides, it discusses some reported applications including estimation of multiple characteristics to highlight the advantages of RSS as a cost-effective and more efficient data collection method. In view of these scenarios the sampling method that is more a cost-efficient data collection method could play an important role for the U N Post 2015 Development Agenda being prepared for the Habitat III summit in 2016.

2. McIntyre's RSS Method

For obtaining a ranked set sample through the McIntyre's RSS procedure *m* random samples with *m* units in each sample are selected from a population with mean, μ and a finite variance, σ^2. This is equivalent to drawing m^2 units randomly and then subdividing them into *m* samples, each with m units. The *m* units of each subset are ranked with respect to the variable of interest without using their exact measurements. For this purpose some outside information like visual perception, past experience, etc., are used. Using this ranking information the unit with the smallest rank is quantified from the first subset; the unit with the second smallest rank is measured from the second subset, and this process of quantification is continued until the unit with the *m*th rank is measured from the *m*th subset. This yields *m* measurements with each of the first *m* ranks, and these constitute an MRSS of size *m*. For obtaining a larger sample of size *mr* the whole procedure is repeated *r* times. Here, *m* is referred to as the set size while *r* is called as the number of cycles. In terms of usual notations we get the sample of size, $n = mr$ from the population with its size, $N \geq m^2 r$.

In a general set up $X_{(i:m)j}$ denotes the *i*th order statistic based on perfect ranking in the *j*th cycle, for $i = 1, \ldots, m$ and $j = 1, \ldots, r$. Note that these are not independent and identically distributed (*iid)* in general, but for a given value of *i* these are so with $E\left(X_{(i:m)j}\right) = \mu_{(i:m)}$, and $\text{var}(X_{(i:m)j}) = \sigma^2_{(i:m)}$. The McIntyre's estimator, $\hat{\mu}_{MRSS}$, of the population mean, μ, is defined as follows:

$$\hat{\mu}_{MRSS} = \frac{1}{mr}\sum_{i=1}^{m}\sum_{j=1}^{r} X_{(i:m)j} \ . \tag{1}$$

Also, if $\hat{\mu}_{(i:m)} = \frac{1}{r}\sum_{j=1}^{r} X_{(i:m)j}$ then $\hat{\mu}_{MRSS} = \frac{1}{m}\sum_{i=1}^{m}\hat{\mu}_{(i:m)}$. (1a)

Here $\text{E}\left(\hat{\mu}_{(i:m)}\right) = \mu_{(i:m)}$; $\text{E}\left(\hat{\mu}_{MRSS}\right) = \mu$ and $\text{var}\left(\hat{\mu}_{(i:m)}\right) = \frac{\sigma^2_{(i:m)}}{r}$. Thus we get

$$\text{var}\left(\hat{\mu}_{MRSS}\right) = \frac{1}{m^2 r}\sum_{i=1}^{m}\sigma^2_{(i:m)}. \tag{2}$$

This expression is also expressed as

$$\text{var}(\hat{\mu}_{\text{MRSS}}) = \frac{1}{mr}\left[\sigma^2 - \frac{1}{m}\sum_{i=1}^{m}\left(\mu_{(i:m)} - \mu\right)^2\right]. \tag{3}$$

The relative precision (RP) of the MRSS estimator, $\hat{\mu}_{MRSS}$ as compared with simple random sample (SRS) estimator, $\hat{\mu}_{SRS}$ with the same sample size, *n*, is computed as follows:

$$\text{RP} = \frac{\text{var}(\hat{\mu}_{SRS})}{\text{var}(\hat{\mu}_{MRSS})}.$$

As $\text{var}(\hat{\mu}_{SRS}) = \frac{\sigma^2}{mr}$, this leads to

$$\text{RP} = \frac{\sigma^2}{\overline{\sigma^2}}; \text{ where } \overline{\sigma^2} = \frac{\sum_{i=1}^{m} \sigma^2_{(i:m)}}{m}. \quad (4)$$

The variance of $\hat{\mu}_{MRSS}$ given in equation (3) yields the expression for RP as given below:

$$\text{RP} = \frac{1}{1 - \frac{1}{m\,\sigma^2} \sum_{i=1}^{m} \left(\mu_{(i:m)} - \mu\right)^2} \quad . \quad (5)$$

2.1 Some More Efficient Estimators

(a) Skewed Distributions

McIntyre (1952) suggested taking the sample size of each rank order proportional to its standard deviation while sampling an asymmetrical population. This leads to

$$r_i = \frac{n\sigma_{(i:m)}}{\sum_{i=1}^{m} \sigma_{(i:m)}}; \quad i = 1, \cdots, m \quad . \quad (6)$$

The RSS estimator, $\hat{\mu}_{MRSS_{UA}}$ of the population mean, μ based on the unequal allocation of samples (also called unbalanced allocation) is given by

$$\hat{\mu}_{MRSSUA} = \frac{1}{m} \sum_{i=1}^{m} \frac{T_i}{r_i} \text{ and } \text{var}(\hat{\mu}_{MRSSUA}) = \frac{1}{m^2} \sum_{i=1}^{m} \frac{\sigma^2_{(i:m)}}{r_i}$$

where $T_{i \text{ shows}}$ the sum of the quantification of the r_i units having *i*th rank order.
On putting the value of $r_{i \text{ from}}$ equation (6) into the expression for $\text{var}(\hat{\mu}_{MRSSUA})$ we have

$$\text{var}(\hat{\mu}_{MRSSUA}) = \frac{(\overline{\sigma})^2}{n} \quad (7)$$

where $\overline{\sigma} = \frac{\sum_{i=1}^{m} \sigma_{(i:m)}}{m}$.

The relative precision (RP_{ua}) of $\hat{\mu}_{MRSSUA}$ relative to $\hat{\mu}_{SRS}$ with the same number of quantifications is given below:

$$RP_{ua} = \frac{\text{var}(\hat{\mu}_{SRS})}{\text{var}(\hat{\mu}_{\text{MRSSUA}})} .$$

This yields that

$$RP_{ua} = \frac{\sigma^2/n}{\sum_{i=1}^{m} \frac{\sigma^2_{(i:m)}}{r_i} \Big/ m^2} . \tag{8}$$

This could also be expressed as

$$RP_{ua} = \left(\frac{\sigma}{\overline{\sigma}}\right)^2 . \tag{9}$$

See Patil et al. (1994b) for these results. Further, it is interesting to note that

$$\text{var}(\hat{\mu}_{MRSS}) - \text{var}(\hat{\mu}_{MRSSUA}) = \frac{\sum_{i=1}^{m} \left(\sigma_{(i:m)} - \overline{\sigma}\right)^2}{mn} .$$

Here $RP_{ua} \geq RP$ and $0 \leq RP_{ua} \leq m$.

For obtaining an estimator of the RP_{ua} we use the estimator of the population variance and that of the population variance of the ith rank order based on unequal sample sizes as given below.

$$\hat{\sigma}^2_{MRSSUA} = \sum_{i=1}^{m} \left[\frac{m(r_i - 1) + 1}{m^2 r_i (r_i - 1)}\right] \sum_{j=1}^{r_i} \left(X_{(i:m)j} - \overline{X}_{(i:m)}\right)^2 + \left(\frac{1}{m}\right) \sum_{i=1}^{m} \left(\overline{X}_{(i:m)} - \overline{X}_{(m)r}\right)^2 . \tag{10}$$

$$\hat{\sigma}^2_{(i:m)UA} = \frac{\sum_{j=1}^{r_i} \left(X_{(i:m)j} - \hat{\mu}_{(i:m)_{UA}}\right)^2}{r_i - 1} \text{ when } \hat{\mu}_{(i:m)UA} = \frac{1}{m} \sum_{i=1}^{m} \frac{T_i}{r_i} . \tag{11}$$

See Norris et al. (1995).

(b) Symmetric Distributions

One could notice two patterns of variances of symmetric distributions. Under one family of symmetric distributions the variances of the order statistics increase with the rank order until the middle position, and then they decrease to the end (for example, Uniform (0,1), Unfolded Weibull (2,0,1), Symmetric Beta (2) etc.) while for the other family the variances decrease with the rank order until the middle and then they increase to have a symmetric pattern (for example Normal (0,1), Logistic (0,1), Laplace (0,1), etc.). Kurtosis seems to discriminate between the two families of symmetric distributions. For the first family of distributions we need to quantify the extreme order statistics to obtain the minimum variance of the estimator of the population mean. For getting the minimum variance in the other family of we need to measure either the middle order statistic or the two closest middle order statistics in 1:1 proportion depending on whether the set size is odd or even.

3. RSS Methods with Concomitant Ranking

RSS presumes that the sampling units are correctly ranked with respect to the variable of interest. This is a perfect ranking (PR) scenario, but this may not always be a rational approach in real life

occasions. In these situations one could take help of some other characteristic for ranking, which is inexpensive, easily available and highly correlated with the main characteristic of interest. Unlike perfect ranking scenario the ranking so obtained may be referred to as concomitant ranking because of its dependence on a concomitant variable. In this situation the relative precision of the estimator will be less than that of the estimator under perfect ranking scenario. But the reduction depends on the magnitude of the correlation between the main variable and the concomitant variable. For a more detailed discussion on concomitant ranking see Sinha (2005) and Patil et al. (1994a and b).

3.1 Estimation of Multiple Characteristics

Using the experience and expertise of the field personnel the RSS methods could be employed for estimating multiple characteristics cost-effectively in a single investigation. Of course, the level of cost-effectiveness depends on the magnitude of the correlation between the main variable of interest and other associated variables. Patil et al.(1994a and b) initiated the work in this direction considering a sampling situation referred to by Sengupta et al. (1951), and discussed by further by Stokes (1980).

Using the data set of of 399 trees provided by Platt et al. (1988) the means of diameter, height and age were estimated employing the RSS methods. With the equal allocation of RSS the estimates of the RPs were obtained as 3.40, 2.36 and 1.76 for height, diameter and age respectively. The corresponding values in the case the unequal allocation are 4.21, 3.05 and 2.60 respectively. These results suggest that the unequal allocation is overall the most efficient method of RSS. See Norris et al. (1995) for more details.

4. Estimation of Urban Populations

RSS may be used to estimate the population of urban populations of a state quite effectively. The technique is illustrated using the urban population of Bihar of 1991 Census. There were 145 urban places according to the Census. We have first randomly selected 81 places using SRSWR and put them in 27 rows with each row consisting of the three urban places with their populations. RSS is used with the set size, m as 3 and the number of replications r as 9 to estimate the sample mean and then we obtained an estimate of the population of 145 places. The actual total population was 9905706 while the male and female populations were 5372380 and 4533326 respectively. Both RSS with equal and unequal methods were used to estimate the total population of 145 places. For unequal RSS method we used $r_1=3$, $r_2=10$ and $r_3=14$. Note that these values are obtained using the Neyman's criterion. Also, because of some considerable differences among the sex ratios of different places the total population of 145 places are obtained directly using the total populations of 27 places as well as estimating the populations of male and female population separately first and then the total populations of 145 places were obtained by adding the two estimates. Table 1 provides the relative precision and savings, the estimated population, difference between the actual and the estimated values with the percentage difference for each allocation. The separate estimations of the populations of male and female yielded an estimate of the total population as 9882234. This shows the difference between the actual and the new estimated total population as 23472, which is 0.234% with respect to the actual population. This approach appears better because of the variations in the sex ratios, and the smaller difference between the actual total population and the new estimate of the total population supports this contention.

Table1**.** Relative precision and savings, estimated population, actual and percentage difference for respective allocation types

Allocation	Relative Precision	Relative Savings	Estimated Population, and Actual Difference, with % Difference
Equal Allocation (m = 3 and r = 9, n=27)	1.493	33%	11204016 and 1298299 (13%)
Unequal Allocation (Total) (m=3; r_1=3, r_2=10, r_3=14)	2.897	65%	9872224 and 33482 (0.338%)
Unequal Allocation (Male) (m=3; r_1=3, r_2=10, r_3=14)	2.861	65%	5352192 and 20188 (0.376%)
Unequal Allocation (Female) (m=3; r_1=3, r_2=10, r_3=14)	2.303	66%	4530042 and 3284 (0.072%)
Total (as the sum of the estimates of male and female populations)	-	-	9882234 and 23472 (0.237%)

References

Dell, T. R., and Clutter, J. L. (1972). Ranked set sampling theory with order statistics background. *Biometrics*, 28, 545-553.

McIntyre, G. A. (1952). A method of unbiased selective sampling using ranked sets. *Australian Journal of Agricultural Research*, 3, 385-390.

Mode, N., Conquest, L. and Marker, D. (1999). Ranked set sampling for ecological research: accounting for the total cost of sampling. *Environmetrics*, 10, 179-194.

Nahas, Ramzi W., Wolfe, Douglas, A. and Chen, Haiing (2002). Ranked set sampling: Cost and optimal set size. *Biometrics* 58, 964-971.

Norris, R. C., Patil, G. P. and Sinha, A. K. (1995). Estimation of multiple characteristics by ranked set sampling methods. *Coenoses* (The journal of the International Center for Theoretical and Applied Ecology, Italy), 10 (2-3), 95-111.

Patil, G. P., Sinha, A. K. and Taillie, C. (1994a). Ranked set sampling: A novel method to accomplish observational economy in environmental studies. *Center for Statistical Ecology and Environmental Statistics, Department of Statistics, Pennsylvania State University, University Park, PA, USA*. Technical Report No. 94-0411.

Patil, G. P., Sinha, A. K. and Taillie, C. (1994b). Ranked set sampling. In: *Handbook of Statistics*, 12 (G. P. Patil and C. R. Rao eds.), 167-200. North-Holland, Elsevier Science B. V., Amsterdam.

Platt, W. J., Evans, G. W., and Rathbun, S. L. (1988). The population dynamics of a long-lived conifer. *The American Naturalist* 131(4), 491-515.

Sengupta, J. M., Chakravarti, I. M. And Sarkar, D. (1951). Experimental survey for the estimation of cinchona yield. *Bulletin of the International Statistical Institute*, 33, 313-331.

Sinha, Arun K. (2014). Ranked set sampling methods for vegetation research. *Int. J. Mendel*, 31 (1-2), 13-22.

Sinha, Arun K. (2005). On some recent developments in ranked set sampling *Bulletin of Informatics and Cybernetics*, **37**, 137-160.

Stokes, S. L. (1980). Estimation of variance using judgment ordered ranked set samples. *Biometrics*, 36, 35-42.

Wolfe, Douglas A. (2004). Ranked set sampling: an approach to more efficient data collection. *Statistical Science*, 19, No.4, 636-643.