CADERNOS DO IME – Série Estatística

Universidade do Estado do Rio de Janeiro - UERJ
ISSN impresso 1413-9022 / ISSN on-line 2317-4536 - v.38, p.37-48, 2015
DOI: 10.12957/cadest.2015.19114

EFFICIENCY OF RANKED SET SAMPLING IN HORTICULTURAL SURVEYS

M. Iqbal Jeelani
1Division of Agricultural Statistics, SKUAST-K, India
jeelani.miqbal@gmail.com

Carlos N. Bouza
Facultad de Matemática y Computación
Universidad de La Habana, Cuba
bouza@matcom.uh.cu

Jose M. Sautto
Univeridad Atunoma de Guerrero
Acapulco, Mexico
3sautto1128@yahoo.com.mx

Abstract

In this paper, we explore the feasibility of using RSS (Ranked Set Sampling) in improving the estimates of the population mean in comparison to SRS (Simple Random Sampling) in Horticultural research. We use an experience developed with a survey of apples in India. The numerical results suggest that RSS procedure results in a substantial reduction of standard errors, and thus provides more efficient estimates than SRS, in the specific Horticultural Survey studied, using the same sample size. Then it is recommended as an easy-to-use accurate method to management of this Horticulture problem.

Key-words: *Ranked Set Sampling, Simple Random Sampling, Standard Error, Accuracy.*

1. Introduction

Horticulture investment has a growing interest. Many organizations do not have proper data recording and reporting systems to generate statistics for characterizing the main problems in decision making. Therefore, the studies must rely on sampling generated data. That is the case when economists look for data on horticultural production for management of processes. They spend much time in collecting production costs, or drawing conclusions from the results of cost-of-production surveys. For farm management surveys need to provide data useful for economic planning, as well as for related scientific and sociological research. The degree of accuracy and the sampling error that is permissible suggests using a complex inquiry but, at the same time, the sampling costs must be as small as possible. In horticulture survey sampling is commonly used for providing information for deciding on different issues as:

- Establishing the levels of trace elements and persistent organic pollutants in soils.
- Periodic monitoring the quality of different vegetables for measuring the extent of pesticide contamination.
- Examining the energy equivalents of inputs and output in greenhouse vegetable production.

See a discussion on different aspects of this kind of applications of sampling in Ozkan *et al.* (2004) and Gockowski & Ndoumbé (2004).

Commonly researches aim to obtain "good samples" and statisticians use prior information to improve the representativeness of the samples in this sense. The first attempts were to divide the population into similar subpopulations and then sampling using these structures. The groups should ensure a broader representation across the entire population. Classic models are systematic sampling, stratified sampling, probability-proportional-to-size sampling, cluster sampling, and quota sampling. The existence of whole information on some correlated auxiliary variable is considered as readily available. They use this information for improving the representativeness of the sample.

McIntyre (1952) suggested using RSS (Ranked Set Sampling). This design considers that there is some reasonable way of using the existing additional information, from each individual population unit, for ranking. In this method, a relatively large number of independent and randomly selected sampling units are partitioned into small

subsets of the same size. The units of each subset are ranked without obtaining the measurements of the interest variable. The ranking induces a stratification on the population and hence. It provides a more structured sample than SRS (Simple Random Sampling) does with the same sample size. Even in the presence of ranking errors, RSS (Ranked Set Sampling) provides unbiased and more efficient estimators of the population mean.

In section 2 we present the main issues of RSS design. Section 3 is concerned with the presentation of numerical studies using real life data. The fourth section is concerned with discussing the obtained results.

2. RSS - Ranked Set Sampling

2.1. Some basic issues of RSS

Let us consider that we deal with a set of sampling units drawn from the population which can be ranked by certain means rather cheaply without the actual measurement of the variable. The original form of RSS, conceived by McIntyre (1952,) can be described as follows.

- Step 1: randomly select k^2 sample units from the population.
- Step 2: allocate the m^2 selected units as randomly as possible into k sets, each of size k.
- Step 3: without yet knowing any values for the variable of interest, rank the units within each set based on a perception of relative values for this variable. This may be based on personal judgment or done with measurements of a covariate that is correlated with the variable of interest.
- Step 4: choose a sample for actual analysis by including the smallest ranked unit in the first set, then the second smallest ranked unit in the second set, continuing in this fashion until the largest ranked unit is selected in the last set.
- Step 5: repeat steps 1 through 4 for m cycles until the desired sample size, $n = mk$, is obtained for analysis**.**

This whole process is referred to as a cycle. The cycle then repeats m times and yields a ranked set sample of size N = mk.

The procedure is a two-stage scheme. At the first stage, simple random samples are drawn and a certain ranking mechanism is employed to rank the units in each simple random sample. At the second stage, actual measurements of the variable of interest are made on the units selected based on the ranking information obtained at the first stage.

The judgment ranking relating to the latent values of the variable of interest, as originally considered by McIntyre (1952), provides one ranking mechanism.

The essence of RSS is conceptually similar to the classical stratified sampling. RSS can be considered as post-stratifying the sampling units according to their ranks in a sample. Although the mechanism is different from the stratified sampling, the effect is the same in that the population is divided into homogeneous sub-populations. In fact, we can consider any mechanism, not necessarily ranking the units according to their X values, which can post-stratify the sampling units in such a way that it does not result in a random permutation of the units. This design is of particular interest for people looking for an accurate and cost-effective survey sampling technique.

2.2. Theoretical aspects of the Ranking mechanisms

Let us start with McIntyre's (1952) original ranking mechanism, i.e., ranking with respect to the latent values of the variable of interest. If the ranking is perfect, that is, the ranks of the units tally with the numerical orders of their latent values of the variable of interest, the measured values of the variable of interest are indeed order statistics. In this case, $f_{[r]} = f_{(r)}$, the density function of the rth order statistic of a simple random sample of size k from distribution F. We have:

$$f_{(r)}(x) = \frac{k!}{(r\text{-}1)!(k\text{-}r)!} F^{r-1}(x)[1\text{-}F(x)]^{k-r} f(x)$$

It is easy to verify that $f(x) = \frac{1}{k}\sum_{r=1}^{k} f_{(r)}(x)$ for all x. This equality plays a very important role in RSS. It is this equality that gives rise to the merits of RSS. We are going to refer to equalities of this kind as fundamental equalities. A ranking mechanism is said to be consistent if the following fundamental equality holds $F(x) = \frac{1}{k}\sum_{r=1}^{k} F_{(r)}(x), \text{ for all } x.$

Obviously, perfect ranking with respect to the latent values of X is consistent. Other consistent ranking mechanisms are as follows.

When there are ranking errors, the density function of the ranked statistic with rank r is no longer $f_{(r)}$. However, we can express the corresponding cumulative distribution function $F_{[r]}$ in the form:

$$F_{[r]}(x) = \sum_{s=1}^{k} p_{sr} F_{(s)}(x),$$

where p_{sr} denotes the probability with which the *s*th (numerical) order statistic is judged as having rank *r*. If these error probabilities are the same within each cycle of a balanced RSS, we have $\sum_{s=1}^{k} p_{sr} = \sum_{r=1}^{k} p_{sr} = 1$. Hence,

$$\frac{1}{k}\sum_{r=1}^{k} F_{[r]}(x) = \frac{1}{k}\sum_{r=1}^{k}\ \sum_{s=1}^{k} p_{sr}\ F_{(s)}(x)\ = \frac{1}{k}\sum_{s=1}^{k}\ \sum_{r=1}^{k} p_{sr}\ F_{(s)}(x) = F(x).$$

There are cases, in practical problems, where the variable of interest, *X*, is hard to measure and difficult to rank as well but a concomitant variable, *Y*, can be easily measured. Then the concomitant variable can be used for the ranking of the sampling units. The RSS scheme is adapted in this situation as follows. At the first stage of RSS, the concomitant variable is measured on each unit in the simple random samples, and the units are ranked according to the numerical order of their values of the concomitant variable. Then the measured *X* values at the second stage are induced order statistics by the order of the *Y* values. Let $Y_{(r)}$ denote the *r*th order statistic of the *Y* 's and $X_{[r]}$ denote its corresponding *X*. Let $fX|Y_{(r)}(x|y)$ denote the conditional density function of *X* given $Y_{(r)} = y$ and $g_{(r)}(y)$ the marginal density function of $Y_{(r)}$. Then we have:

$$f_{[r]}(x) = \int fx\,|Y_{(r)}(x\,|\,y)g_{(r)}(y)dy.$$

It is easy to see that $f(x) = \int \sum_{r=1}^{k} \frac{1}{k}\, fx\,|Y_{(r)}(x\,|\,y)g_{(r)}(y)dy. = \frac{1}{k}\sum_{r=1}^{k} f_{(r)}(x).$

2.3 Estimation of means using ranked set sampling

Let $h_{(x)}$ be any function of *x*. Denote by μ_h the expectation of $h(X)$, i.e., $\mu h = Eh(X)$. We consider in this section the estimation of μ_h by using a ranked set sample. Examples of $h(x)$ include:

(a) $h(x) = x^l$, $l = 1, 2, \cdots$, corresponding to the estimation of population moments,

(b) $h(x) = I\{x \le c\}$ where $I\{\cdot\}$ is the usual indicator function, corresponding to the estimation of distribution function,

(c) $h(x) = \frac{1}{\lambda}K\left(\frac{t-x}{\lambda}\right)$, where K is a given function and λ is a given constant, corresponding to the estimation of density function.

We assume that the variance of $h(X)$ exists, then $\hat{\mu}_{h.RSS} = \frac{1}{mk}\sum_{r=1}^{k}\sum_{i=1}^{m} h(X_{[r]i})$.

We consider first the statistical properties of and then the relative efficiency of RSS with respect to SRS in the estimation of means. They are based on the following result.

Theorem 1. Suppose that the ranking mechanism in RSS is consistent. Then,

i) The estimator $\hat{\mu}_{h.RSS}$ is unbiased, i.e., $E\hat{\mu}_{h.RSS} = \mu_h$

ii) $\mathrm{Var}(\hat{\mu}_{h.RSS}) \le \frac{\sigma_h^2}{mk}$, where σ_h^2 denotes the variance of $h(X)$, and the inequality is strict unless the ranking mechanism is purely random.

iii) As $m \to \infty$,

$$\sqrt{mk}\,(\hat{\mu}_{h.RSS} - \mu_h) \to N(0, \sigma^2_{h.RSS})$$

in distribution, where,

$$\sigma^2_{h.RSS} = \frac{1}{k}\sum_{r=1}^{k}\sigma^2_{h[r].}$$

Here $\sigma^2_{h[r]}$ denotes the variance of $h(X_{[r]i})$

Proof :

i) It follows from the fundamental equality that

$$E\hat{\mu}_{h.RSS} = \tfrac{1}{mk}\sum_{r=1}^{k}\sum_{i=1}^{m} Eh(X_{[r]i}) = \frac{1}{k}\sum_{r=1}^{k} Eh(X_{[r]i})$$

$$= \frac{1}{k}\sum_{r=1}^{k}\int h(x)\,dF_{[r]}(x) = \int h(x)\,d\,\frac{1}{k}\sum_{r=1}^{k}F_{(r)}(x)$$

$$= \int h(x)\,dF(x)\mu_h$$

ii)

$$\mathrm{Var}(\hat{\mu}_{h.RSS}) = \frac{1}{(mk)^2}\sum_{r=1}^{k}\sum_{i=1}^{m}\mathrm{Var}(h(X_{[r]}i)) = \frac{1}{mk^2}\sum_{r=1}^{k}\mathrm{Var}\,(h(X_{[r]}))$$

$$= \frac{1}{mk}\left(\frac{1}{k}\sum_{r=1}^{k}(E[h(X_{[r]})]^2 - [Eh(X_{[r]})]^2\right)$$

$$= \frac{1}{mk}\left(m_{h2} - \frac{1}{k}\sum_{r=1}^{k}[Eh(X_{[r]})]^2 \right),$$

Where m_{h2} denotes the second moment of $h(X)$. It follows from the Cauchy-Schwarz inequality that

$$\frac{1}{k}\sum_{r=1}^{k}[Eh(X_{[r]})]^2 \geq \left(\frac{1}{k}\sum_{r=1}^{k} Eh(X_{[r]}) \right)^2 = \mu_h^2,$$

where the equality holds only when $Eh(X_{[1]}) = \cdots = Eh(X_{[r]})$ in which case the ranking mechanism is purely random.

iii) By the fundamental equality, $\mu_h = \frac{1}{k}\sum_{r=1}^{k}\mu_{h[r]}$, where $\mu_{h[r]}$ is the expectation of $h(X_{[r]i})$. Then, we can write

$$\sqrt{mk}\,(\hat{\mu}_{h.RSS} - \mu_h) = \frac{1}{\sqrt{k}}\sum_{r=1}^{k}\sqrt{m}\,[\frac{1}{m}\sum_{i=1}^{m}h(X_{[r]}i) - \mu_{h[r]}]$$

$$\frac{1}{\sqrt{k}}\sum_{r=1}^{k} Z_{mr}, \text{say.}$$

By the multivariate central limit theorem, (Z_{m1}, $\cdots$, Z_{mk}) converges to a multivariate normal distribution with mean vector zero and covariance matrix given by Diag $(\sigma^2_{h[1],\cdots,}\,\sigma^2_{h[k]})$. Part (iii) then follows.

We know that $\sigma_h^2/(mk)$ is the variance of the moment estimator of μ_h based on a simple random sample of size *mk*. Theorem 1 implies that the moment estimator of μ_h based on an RSS sample always has a smaller variance than its counterpart based on an SRS sample of the same size. In the context of RSS, we have tacitly assumed that the cost or effort for drawing sampling units from the population and then ranking them is negligible. When we compare the efficiency of a statistical procedure based on an RSS sample with that based on an SRS sample, we assume that the two samples have the same size. Let $\hat{\mu}_{h.SRS}$ denote the sample mean of a simple random sample of size *mk*. We define the relative efficiency of RSS with respect to SRS in the estimation of μ_h as follows:

$$RE(\hat{\mu}_{h.RSS}, \hat{\mu}_{h.SRS})\frac{Var(\hat{\mu}_{h.SRS})}{Var(\hat{\mu}_{h.RSS})}$$

Then, Theorem 1 implies that RE $(\hat{\mu}_{h.RSS}, \hat{\mu}_{h.SRS}) \geq 1$. In order to investigate the relative efficiency in more detail, we derive the following:

$$\sigma^2_{h.RSS} = \frac{1}{k}\sum_{r=1}^{k}\sigma^2_{h[r]}$$

$$= \frac{1}{k}\sum_{r=1}^{k}(E[h(X_{[r]})]^2 - [Eh(X_{[r]})]^2)$$

$$= \frac{1}{k}\sum_{r=1}^{k}(E[h(X_{[r]})]^2 - \mu_h^2 + \mu_h^2 - \frac{1}{k}\sum_{r=1}^{k}[Eh(X_{[r]})]^2$$

$$= \sigma_h^2 - \frac{1}{k}\sum_{r=1}^{k}(\mu_{h[r]} - \mu_h)^2.$$

Thus, we can express the relative efficiency as:

$$RE(\hat{\mu}_{h.RSS}, \hat{\mu}_{h.SRS}) = \frac{\sigma_h^2}{\sigma^2_{h.RSS}} = \left[1 - \frac{\frac{1}{k}\sum_{r=1}^{k}(\mu_{h[r]} - \mu_h)^2}{\sigma_h^2}\right]^{-1}$$

It is clear from the above expression that, as long as there is at least one *r* such that $\mu_{h[r]} \neq \mu_h$, the relative efficiency is greater than 1. For a given underlying distribution and a given function h, the relative efficiency can be computed, at least, in principle.

2.4 Estimation of the variance using an RSS sample

The natural estimates of σ^2 using an SRS sample and an RSS sample are given, respectively, by

$$S^2{}_{SRS} = \frac{1}{mk-1}\sum_{r=1}^{k}\sum_{i=1}^{m}(X_{ri} - \overline{X}_{SRS})^2,$$

where $\overline{X}_{SRS} = \frac{1}{mk}\sum_{r=1}^{k}\sum_{i=1}^{m}X_{ri}$, and $S^2{}_{SRS} = \frac{1}{mk}\sum_{r=1}^{k}\sum_{i=1}^{m}(X_{[r]i} - \overline{X}_{SRS})^2$,

where $\overline{X}_{RSS} = \frac{1}{mk}\sum_{r=1}^{k}\sum_{i=1}^{m}X_{[r]i.}$

The properties of S^2_{SRS} were studied by Stokes (1980) Unlike the SRS version S^2_{SRS} the RSS version S^2_{RSS} is biased. It can be derived, that:

$$E_{S^2_{SRS}} = \sigma^2 + \frac{1}{k(mk-1)}\sum_{r=1}^{k}(\mu_{[r]} - \mu)^2$$

An appropriate measure of relative efficiency of S^2_{SRS} with respect to S^2_{SRS} is then given by

$$RE(S^2_{SRS}, S^2_{SRS}) = \frac{Var(S^2_{SRS})}{MSE(S^2_{RSS})} = \frac{Var(S^2_{SRS})}{Var(S^2_{RSS}) + [\frac{1}{k(mk-1)} \sum_{r=1}^{k} (\mu_{[r]} - \mu)^2]^2}$$

It can be easily seen that $RE(S^2_{RSS}, S^2_{SRS}) < ARE(S^2_{RSS}, S^2_{RSS})$.

Since $\frac{1}{k}\sum_{r=1}^{k}(\mu_{[r]} - \mu)^2 < \sigma^2$, it is clear that $\frac{1}{k(mk-1)}\sum_{r=1}^{k}(\mu_{[r]} - \mu)^2$ will decrease as either *k* or *m* increases.

3. Numerical studies

A study provided what we call “Apple data”. This data are utilized in the present paper. The block Ganderbal was selected for the present study in the District Ganderbal. District Ganderbal being inseparable part of the state, naturally inherits the same characteristics which predominately exist in the economy of the state. Agriculture is the main source of income and employment in the district. More than half of the population, directly and indirectly derive their livelihood from it. Paddy, maize and horticulture are the principle crops grown in the district. There is a good network of agricultural infrastructure available throughout the length and breadth of the district. Total area sown under different food and non-food crops is about 27735 hectares, out of which 15828 hectares constituting 57 per cent was under cereal food crops. At present 8738 hectares are under major horticulture crops with 3866 hectares constituting 44 per cent are under apple cultivation and out of 47916 MT of production of horticulture crops, apple production is 34873 MT which is 72 per cent of the total production. A survey was conducted for estimation of average yield of apple in the district Ganderbal at block level. Since at present 8738 hectares are under major horticulture crops with 3866 hectares constituting 44 percent of the area is under apple cultivation in district Ganderbal. A total of 420 orchards were reported in the block Ganderbal covering an area of 772.8 hectares with 73,496 total number of trees. Total production of apple in the block was found out to be 6758.52 metric tons (Mt) with the productivity of 8.74 Mt/ha. American, Delicious and Maharaji were the main varieties of apple cultivated in the block.

The data was collected on Apple production from district Ganderbal of Kashmir valley from 420 orchards in 30 villages. The variables choosen for the study where Yield (MT), Bearing trees, Total number of trees, Area (ha). We take equal sample size from each sampling design and estimate the standard error in each sampling design. The sample sizes considered were 15, 25, 45, 65 and the set sizes considered were 2,4,10 shown in Table: 1 along with correlation coefficients ρ ranging from 0.80 to 0.65.

Three distinct simulations based on three combinations of sample sizes and set sizes for each sampling design; each simulation uses a combination of variables for ranking and quantification.

Sampling procedure	**Variable combinations**	**STANDARD ERRORS**				
		No of sets	**Sample sizes**			
			15	**25**	**45**	**65**
Simple random sampling	**Yield *vs* Area**	**2**	177.13	171.16	163.52	155.71
		4	174.43	1697.27	158.04	152.38
		10	162.27	156.04	150.43	141.63
Ranked set sampling		**2**	174.43	167.32	157.63	149.43
		4	167.78	161.42	153.43	144.27
		10	159.43	152.32	145.01	141.63
Simple random sampling	**Bearing trees *vs* Area**	**2**	1753.43	1726.39	1704.58	1677.54
		4	1740.52	1721.32	1695.04	1671.41
		10	1725.65	1714.42	1687.58	1651.32
Ranked set sampling		**2**	1712.38	1696.43	1683.43	1665.52
		4	1706.12	1688.18	1667.53	1648.37
		10	1687.35	1677.53	1664.43	1632.52
Simple random sampling	**Total trees *vs* Area**	**2**	2268.52	2257.25	2237.63	2226.13
		4	2260.48	2254.1	2233.57	2215.08
		10	2256.33	2236.09	2225.01	2209.54
Ranked set sampling		**2**	2249.11	2237.51	2227.54	2215.09
		4	2245.27	2226.62	2220.63	2205.52
		10	2227.52	2217.11	2213.57	2201.54

Table.1: Variable combinations along with standard errors.

4. Conclusions

From the above results it is concluded, as theoretically expected, that RSS, when used in place of SRS provided estimates of population mean that are more accurate. The results of Table .1 reveals this fact. There is also a considerably reduction in the standard errors as we increase the sample size. Obtaining a sample in this manner maintains the unbiasedness of SRS; however, by incorporating 'outside' information about the sample units, we are able to contribute a structure to the sample that increases its representativeness of the true underlying population. If we quantified the same number of sample units, by a simple random sample, then we have no control over which units

entering the sample. Perhaps all the units would come from the lower end of the range, or perhaps most would be clustered at the low end while one or two units would come from the middle or upper range. With SRS, the only way to increase the prospect of covering the full range of possible values is to increase the sample size. RSS has a balanced nature in the sense that equal number of observations will be obtained from each rank. It can be easily shown that the sample mean using RSS has a smaller standard errors than the sample mean using the traditional simple random sampling (SRS) when the number of observations are same. Therefore, the costs of sampling may be reduced as, if we fix the optimal sample size n for SRS, with RSS we may use a smaller value of n for attaining the same accuracy.

Referências

AL-OMARI, A. I.; BOUZA, C. N. (2014). Review of Ranked Set Sampling: Modifications and Applications. **Revista Investigación Operacional**, 35, 215-240.

BAI, Z. D.; CHEN, Z. (2003). On the theory of ranked set sampling and its ramifications. **Journal of Statistical Planning and Inference**109: 81-99.

BOUZA, C. N. (2010). Ranked set sampling for estimating of population under non-response. **Revista Investigacion Operacional** 31 : 140-150.

BOUZA, C. N. (2013). Handling Missing Data in Ranked Set Sampling, **Springer Briefs in Statistics**, Springer

CHEN, Z. (2001). Ranked-set sampling with regression type estimators. **Journal of Statistical Planning and Inference**, 92 : 181-192.

CHEN, Z.; BAI, Z.D. (2000).The optimal ranked-set sampling scheme for parametric families. **Sankhya Ser**. A. 62: 178-192.

COCHRAN, W. G. (1977). **Sampling Techniques**. John Wiley and Sons, New York.

GAAJENDRA, K. A.; BOUZA, C. (2012). Double sampling with rank set selection in the second phase with non-response: Analytical results and Monte Carlo experiments. **Journal of Probability and Statistics**, 23 : 45-53.

GOCKOWSKI J.; NDOUMBÉ, M. (2004): The adoption of intensive monocrop horticulture in southern Cameroon. **Agricultural Economics**.30, 195–202

JEELANI, M. I., MIR, S. A., KHAN,I., NAZIR,N.; JEELANI, F. (2014). Non-response problems in ranked set sampling. Pakistan Journal of Statistics. 30(4), 555-562.

KAUR, A., PATIL, G.P.; TAILLIE, C. (1997). Unequal allocation models for ranked set sampling with skew distributions. **Biometrics**, 53 : 123-130.

MARTIN, W. L., SHANK, T. L., ODERWALD, R. G.; SMITH, D. W. (1980). Evaluation of ranked set sampling for estimating shrub phytomass in Appalachian Oak forest.**Technical Report** No.FWS-4-80, School of Forestry and Wildlife Resources VPI & SU Blacksburg, VA.

MCINTYRE, G. A. (1952). A Method for unbiased selective sampling, using ranked sets. **Australia Journal of Agric. Res**. 3: 385-390.

OZKAN, B., KURKLU, A.; AKCAOZ, H. (2004): An input–output energy analysis in greenhouse vegetable production: a case studyfor Antalya region of Turkey. **Biomass and Bioenergy**, 26, 89–95.

OZTURK, O.; WOLFE, D. A. (1998). Optimal ranked set sampling protocol for the signed rank test. **Technical Report** TR 630, Ohio State University Department of Statistics.

RISCH, N.; ZHANG, H. (1995). Extreme discordant sib pairs for mapping quantitative trait loci in humans. **Science**. 268 : 1584-1589.