Teacher's Corner

Ranked Set Sampling

Kevin Carl P. Santos
University of the Philippines Diliman

Sampling is certainly one of the major concerns in data collection and in making statistical inferences. To obtain accurate inferences, the sample should be a good representative of the target population. Several sampling schemes have already been studied and used in practice, e.g. Simple Random Sampling (SRS). SRS is very easy to implement; nevertheless, it is not always appropriate to be used when the population of interest is heterogeneous with respect to the target variable.

Furthermore, there are two important issues in sampling, namely: cost efficiency and accuracy or precision. Oftentimes, researchers want to obtain a sample keeping their expenses at low levels but achieving accurate inferences. SRS may yield unreliable conclusions when the population from which it is drawn is very heterogeneous. Moreover, transportation costs would be an issue if the sample obtained were widely spread out in the population.

McIntyre (1952) proposed a sampling scheme which produces accurate inferences and yet minimizes the costs in estimating the average yield from large plots of arable crops. The sampling procedure promises an increase in precision by using visual inspection in ranking units without measuring the actual variable of interest. The sampling scheme follows:

1. Consider a SRS of size k, where k is called the set size, from the population of size N. It is should be noted that the SRS of size k is obtained from the sampling frame only. It is not necessary to physically draw the sampled units or individuals. Then, rank the units in the SRS via visual inspection. The smallest or the first order statistic will be the first element in the sample of size k. After obtaining the first order statistic, the k-1 elements will be disregarded, returned back in the population. These k-1 elements are only used to identify the smallest in the ranked set.

2. Obtain another SRS of size k and rank these units again using visual inspection. The second order statistic or second smallest unit will be the second element in the sample of size k. Again, the k-1 units will be returned in the population.
3. Another SRS of size k will be taken from the population. The third order statistic of the ranked units will be the third element in the sample of size k. Repeat the process until the kth order statistic is obtained. The kth order statistic is the kth element in the sample of size k.
4. Performing steps 1 to 3 is equivalent to one cycle. Repeat the whole cycle m times to obtain a sample of size n = mk.

Without any mathematical proof, McIntyre (1952) claimed that the sample of n items selected in this way yield unbiased estimate of the population mean. This proposed sampling procedure did not get noticed until Halls and Dell (1966) used it in estimating the forage yields in a pine hardwood forest. The name Ranked Set Sampling (RSS) was first introduced by Halls and Dell. As mentioned earlier, McIntyre was not able to provide mathematical proofs of the optimal properties of the estimator of the population mean using RSS. Takahasi and Wakimoto (1968) provided the theoretical results proving that the RSS estimator of the mean is unbiased even with different distributional assumptions under the assumption of perfect ranking. Moreover, the RSS estimator of the mean is more efficient than that of SRS. Table 1 shows the RSS estimator and its variance as cited by Wolfe (2004).

Table 1. RSS Estimator of Population Mean, its Variance and Estimator of the Standard Error

Parameter	Estimator	Variance	Estimated Standard Error
Mean	$\bar{y} = \frac{1}{mk}\sum_{r=1}^{k}\sum_{i=1}^{m} y_{[r]i}$	$\frac{\sigma^2}{mk} - \frac{1}{m^2 r}\sum_{i=1}^{m}\left(\mu_{[i;m]} - \mu\right)^2$	$\sqrt{\frac{1}{mk-1}\sum_{r=1}^{k}\sum_{i=1}^{m}(y_{[r]i} - \bar{y})^2}$

It has been shown by Patil (1995) that the relative precision of RSS compared to SRS estimator of the population mean is:

$$1 \le RP = \frac{Var(\bar{y}_{SRS})}{Var(\bar{y}_{RSS})} \le \frac{k+1}{2},$$

where k is the set size. This implies that as k increases, the relative precision (RP) also increases. Thus, increasing the set size will yield a more reliable estimator for RSS compared to that of SRS. Despite that fact, it should be noted that taking

a large set size entails large expenses in obtaining samples. This is the reason why k is not usually large in practice.

Al-Saleh and Al-Omari (2002) stated that the measurements gathered using RSS are likely to be more regularly spaced than those drawn using SRS, and hence, the sample would most likely be a better representation of the population. Moreover, Dell and Clutter (1972) studied the possibility of imperfect ranking, i.e. the ith sample in the jth cycle may not be the ith order statistic in that sample, but rather the ith "judgment order statistic." They had similar results; that is, RSS estimator of the mean is still unbiased and more efficient than the SRS estimator.

In the advent of RSS, McIntyre (1952) used eye inspection in ranking the units or individuals in performing RSS. Stoke (1977) proposed the use of auxiliary information in conducting the sampling scheme. His idea is very similar to that of Sampling with Probability Proportional to Size (PPS). He proposed to use a concomitant variable or a frugal measurement that would be used in ranking the individuals or units. He concluded that the level of precision depends on the degree of the correlation of the two variables. The underlying assumption in his proposed method is that the frugal measurement or concomitant variable is cheap and/or very easy to obtain. In medical studies, quantitative genetics, and ecological and environmental studies, some attributes can be easily obtained or quantified. This was a breakthrough in RSS because this broadened the applications of the sampling scheme.

After putting so much attention in estimating the mean, Stokes (1980) studied the estimation of the variance while Chen et al. (2005) studied the estimation of the population proportion using RSS. Since then, the theoretical foundation of RSS has developed more extensively. Many studies have been made both on parametric and nonparametric RSS. Different people modified the original procedure of RSS. Some of the modifications are Double RSS (Al-Saleh and Al-Kadiri, 2000) which was later on generalized to Multi-Stage RSS (Al-Saleh and Al-Omari, 2002), Median RSS (Muttlak, 1997) which was also generalized to Multistage Median RSS (Jemain et al., 2007) and Extreme RSS (Samawi et al., 1996) which is now being used in genetics for quantitative trait loci (QTL) mapping as cited by Chen et al. (2005). Ratio (Kadilar, 2007) and regression-type (Chen, 2001) estimators using RSS have already been developed.

In addition to that, a recent modification of RSS was implemented by Chen et al. (2008) in the application of RSS in treatment comparisons in clinical trials. Each time two set of experimental units are randomly taken and ranked separately according to some concomitant variable. The units with odd ranks are assigned to the first treatment while the units with even ranks are assigned to the second treatment for the first ranked set. For the second ranked set, units with even ranks will be designated to the first treatment while those with odd ranks to the second

treatment. They were able to show that this method of treatment assignment is much more efficient than the usual random assignment of treatment.

Numerous studies in the literature focus on RSS in the context of an infinite population. However, in practice, the population is finite. Patil et al. (1995) proposed finite population corrections for RSS. They concluded that the relative precision of the RSS estimator of the population mean depends upon the number of replications or cycles of the set size k. Jozani and Johnson (2009) explored the design-based estimation for RSS in finite populations. Using theoretical and simulation results, it was shown that RSS design can yield significant improvement in efficiency over SRS design in finite populations. Bouza (2001) investigated the use of model-assisted ranked set sampling. He proposed the use of a ratio estimator and a simple linear regression superpopulation model as a counterpart to the design-based estimation approach.

Evidently, ranked set sampling has regained interest among researchers in sampling designs. Several disciplines have already been using this sampling scheme due to its promising results. More researchers and statisticians should look into the applications of RSS in their respective fields where it promises efficiency at lower cost.

References

AL-SALEH, M. and M. AL-KADIRI, 2000, Double-ranked set sampling, *Statistics & Probability Letters*, 48, 205-212.

AL-SALEH, M. and A. AL-OMARI, 2002, Multistage ranked set sampling, *Journal of Statistical Planning and Inference*, 102, 273–286.

BOUZA, C., 2001, Model-Assisted Ranked Survey Sampling, *Biometrical Journal*, 43 (2), 249-259.

CHEN, Z., 2001, Ranked Set Sampling with Regression-type Estimators, *Journal of Statistical Planning and Inference*, 92, 181-192.

CHEN, H., E. STASNY and D. WOLFE, 2005, Ranked Set Sampling for Efficient Estimation of a Population Proportion, *Statistics in Medicine*, 24 (21), 3319-3329.

CHEN Z., G. ZHENG, K. GHOSH, and Z. LI, 2005, Linkage disequilibrium mapping of quantitative trait loci by selective Genotyping, *American Journal of Human Genetics*. 77, 661–669

CHEN, Z., J. LIU, L. SHEN, and Y. WANG, 2008, General Ranked Set Sampling for Efficient Treatment Comparisons, *Statistica Sinica*. 18, 91-104.

DELL, T. and J. CLUTTER, 1972, Ranked set sampling theory with order statistics background, *Biometrika*, 28, 545–555.

HALLS, L. and T. DELL, 1966, Trial of ranked set sampling for forage yields, *Forest Science*, 12, 22–26.

JEMAIN, A., A. AL-OMARI, and K. IBRAHIM, 2007, Multistage Median Ranked Set Sampling for Estimating the Population Median, *Journal of Mathematics and Statistics*, 3 (2), 58-64.

JOZANI, M. and B. JOHNSON, 2010, Design based estimation for ranked set sampling in finite populations, *Environmental and Ecological Statistics*.

KADILAR, C., Y. UNYAZICI, and H. CINGI, 2007, Ratio Estimator for the Population Mean using Ranked Sampling, *Statistical Papers*, 50 (2), 301-309.

MCINTYRE, G.A., 1952, A method of unbiased selective sampling using ranked sets, *Australian Journal of Agricultural Research*, 3, 385–390.

MUTTLAK, H., 1997, Median ranked set sampling, *Journal of Applied Statistical Science*, 6.

PATIL, G.P., 1995, Editorial: Ranked set sampling, *Environmental and Ecological Statistics*, 2, 271–285.

PATIL, G., A. SINHA and C. TAILLIE, 1995, Finite population corrections for ranked set sampling, *Annals of the Institute of Statistical Mathematics*, 47 (4), 621-636.

SAMAWI, H., M. AHMEND and W. ABU-DAYYEH, 1996, Estimating the Population Mean using Extreme Ranked Set Sampling. *Biometrical Journal*, 38 (5), 577-586.

STOKES, S., 1977, Ranked set sampling with concomitant variables, *Communications in Statistics - Theory and Methods*, A6 (12), 1207–1211.

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

TAKAHASI, K. and K. WAKIMOTO, 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.

WOLFE, D., 2004, Ranked Set Sampling: An Approach to More Efficient Data Collection. *Statistical Science*, 19(4), 636-643.