Austrian Journal of Statistics
January 2020, Volume 49, 60–66.
http://www.ajs.or.at/
doi:10.17713/ajs.v49i1.908

Practicality of Some Variations of Ranked Set Sampling

Monjed H. Samuh
Palestine Polytechnic University

M. Hafidz Omar
King Fahd University
of Petroleum & Minerals

M. Pear Hossain
Bangabandhu Sheikh Mujibur
Rahman Science
& Technology University

Abstract

Judgement ranking in ranked set sampling (RSS) and its variations depends on the ability of an observer to rank a set of objects according to the study variable without doing any actual measurement. In practice, and in some variations of RSS, it is hard to assign these ranks. In this paper, we discuss the practicality of ranking some extensions of RSS such as median RSS, double median RSS, and double RSS. The Hellinger distance is used as a measure of practicality. Although double median RSS is the most efficient approach among the RSS variations considered, it is shown in this paper that it is the least practical.

Keywords: efficiency, Hellinger distance, median, practicality, ranked set sampling.

1. Introduction

Ranked set sampling (RSS), proposed by McIntyre (1952), is a data collection or a sampling scheme. Due to its importance for a variety of applications in statistics, it is republished in McIntyre (2005). It is proposed to estimate the mean of Australian pasture yields. McIntyre (1952, 2005) claimed that the RSS mean is an unbiased estimator of the population mean and the variance of the RSS mean is smaller than that in simple random sampling (SRS) with equal measurement elements. This sampling scheme is useful when it is difficult to measure large number of elements but visually (without inspection) ranking some of them is easier. The scheme involves randomly selecting m sets (each of size m elements) from the study population. The elements of each set are ordered with regards to the study variable by any negligible cost method or visually without measurements. Finally, the i^{th} minimum from the i^{th} set, $i = 1, 2, \ldots, m$, are identified for measurement. The obtained sample is called a ranked set sample of set size m. Takahasi and Wakimoto (1968) provided the mathematical theory behind the claims of McIntyre (1952, 2005).

As claimed by McIntyre (1952, 2005) it is later shown in the literature that estimators calculated based on RSS are more efficient than their counterpart in SRS. For example, Stokes and Sager (1988) showed that the empirical distribution function based on RSS is more efficient than its counterpart in SRS. Some authors estimate the parameters of a specific distribution

using RSS, see for example Al-Saleh and Diab (2009) and Sarikavanij, Kasala, Sinha, and Tiensuwan (2014).

To better improve the efficiency of the estimators some variations of RSS were proposed. Al-Saleh and Al-Kadiri (2000) suggested double RSS (DRSS), as a method that improves efficiency of the RSS estimators while keeping m fixed. They reported that the RSS estimator is less efficient than when using DRSS. Muttlak (1997) proposed median RSS (MRSS) as a modification of RSS to improve the efficiency of the estimators of the population mean for symmetric distributions and of the population median. The procedure of MRSS is similar to RSS but in lieu of identifying the i^{th} minimum from the i^{th} set only the median of each set is identified. Given odd set size m, the $\left(\frac{m+1}{2}\right)^{th}$ smallest element is identified from each set for measurement. When m is even, from the first $\frac{m}{2}$ sets the $\left(\frac{m}{2}\right)^{th}$ smallest element is identified for measurement and from the second $\frac{m}{2}$ sets the $\left(\frac{m}{2}+1\right)^{th}$ smallest element is identified for measurement. Samawi and Tawalbeh (2002) suggested a double MRSS (DMRSS) as an alternative procedure to improve the efficiency of the sample mean. They compared the DMRSS with SRS, RSS, DRSS, and some other sampling schemes and found that DMRSS is the most efficient scheme.

Recently, there have been work on multi-stage sampling. Amro and Samuh (2017) constructed a permutation test based on multistage ranked set sampling (MSRSS) that has more statistical power. Mahdizadeh and Zamanzade (2017a) discussed the estimation of a symmetric distribution function under MSRSS. In addition, Mahdizadeh and Zamanzade (2017b) discussed the estimation of reliability under MSRSS. More recently, Mahdizadeh and Zamanzade (2019) explored the estimation of body fat by an efficient method through multistage pair RSS. Samuh (2018) discussed the estimation of distribution function under multistage median RSS.

Although recent works in the literature address the usefulness of MSRSS schemes, the current paper explores the comparison of RSS schemes up to double stage sampling only as the paper is an initial work on practicality of these schemes. Intuitive insights can be drawn into MSRSS schemes from this work but more elaborate studies will be needed to address this comparison for the MSRSS schemes and will use up more space. In the interest of conserving space, the paper is focused mainly on the RSS schemes up to double stage sampling.

In the process of DMRSS, the data points are identified based on the data points of MRSS. For example, if m is odd, the data points of the DMRSS are just the medians of the data points of MRSS; that is, the data points of DMRSS are the medians of the medians of the SRS. It is clear that identifying median of the medians is a hard process, and this contradict the nature of RSS schemes which require visual comparison without inspection (a rationale originally mentioned by McIntyre (1952)). In the process of DRSS, the data points are identified based on the data points of the RSS. For example, the first data point of DRSS is the minimum of the RSS data points, which is easy to be identified visually without inspection. Al-Saleh and Al-Kadiri (2000) have shown by the degree of distinguishability and the probability of perfect ranking that ranking an iid data points is harder than ranking ordered (but independent) data points. Thus, ranking observations in a DMRSS is harder than in a DRSS. In other words, DRSS is more practical than DMRSS. In this paper, since observations that are closer to each other are more difficult to rank, we suggest to use the Hellinger distance (defined in Eq. (4) later in this paper) as a measure of ranking practicality.

To our knowledge, practicality of RSS schemes have not been compared in the literature with regards to Hellinger distance. The rest of the paper is organized as follows. A general setup and some basic results are given in Sec. 2. Hellinger distance is defined and applied to RSS schemes in Sec. 3. Finally, Sec. 4 concludes the paper.

2. Some basic properties of the sampling schemes

Let X be a continuous random variable with cumulative distribution function (cdf) $F(x)$, and probability density function (pdf) $f(x)$.

Simple random sampling Let $X_1, X_2, \ldots, X_m$ indicate a SRS from $f(x)$, then X_i are independent and identically distributed as $f(x)$. Note that when $f(x)$ is infinite, SRS and random sample are used synonymly.

Ranked set sampling Let $Y_1^{(1)}, Y_2^{(1)}, \ldots, Y_m^{(1)}$ be a RSS; that is $Y_i^{(1)}$ is the i^{th} order statistic of the random sample $X_1, X_2, \ldots, X_m$, where the superscript (1) represents stage 1. The cdf of $Y_i^{(1)}$ is

$$F_{Y_i^{(1)}}(y) = F_{X_{(i)}}(y) = \sum_{k=i}^{m} \binom{m}{k} F^k(y)\,(1-F(y))^{m-k}\,,\ i = 1, 2, \ldots, m.$$

The pdf of $Y_i^{(1)}$ is

$$f_{Y_i^{(1)}}(y) = m\binom{m-1}{i-1} F^{i-1}(y)\,(1-F(y))^{m-i}\,f(y),\ i = 1, 2, \ldots, m.$$

Double RSS Let $Y_1^{(2)}, Y_2^{(2)}, \ldots, Y_m^{(2)}$ be a DRSS; that is $Y_i^{(2)}$ is the i^{th} order statistic of the RSS $Y_1^{(1)}, Y_2^{(1)}, \ldots, Y_m^{(1)}$ (which are independent but not identical random variables). Hence, the cdf of $Y_i^{(2)}$ is

$$F_{Y_i^{(2)}}(y) = \sum_{l=i}^{m} \sum_{S_l} \left(\prod_{k=1}^{l} F_{Y_{j_k}^{(1)}}(y) \prod_{k=l+1}^{m} \left(1 - F_{Y_{j_k}^{(1)}}(y)\right) \right), \tag{1}$$

where S_l is the set of the entire permutations $(j_1, j_2, \ldots, j_m)$, of the integers $(1, 2, \ldots, m)$ for which $j_1 < j_2 < \cdots < j_l$, and $j_{l+1} < j_{l+2} < \cdots < j_m$ (David and Nagaraja (2003)). The pdf of $Y_i^{(2)}$ is the derivative of $F_{Y_i^{(2)}}(y)$.

Median RSS Let $W_1^{(1)}, W_2^{(1)}, \ldots, W_m^{(1)}$ be a MRSS; that is

$$W_i^{(1)} = \begin{cases} X_{(\frac{m+1}{2})} & \text{if } m \text{ odd } \&\, i = 1, \ldots, m \\ X_{(\frac{m}{2})} & \text{if } m \text{ even } \&\, i = 1, \ldots, \frac{m}{2} \\ X_{(\frac{m+2}{2})} & \text{if } m \text{ even } \&\, i = \frac{m+2}{2}, \ldots, m \end{cases} \tag{2}$$

The pdf of $W_i^{(1)}$ is

$$f_{W_i^{(1)}}(x) = \begin{cases} f_{X_{(\frac{m+1}{2})}}(x) & \text{if } m \text{ odd } \&\, i = 1, \ldots, m \\ f_{X_{(\frac{m}{2})}}(x) & \text{if } m \text{ even } \&\, i = 1, \ldots, \frac{m}{2} \\ f_{X_{(\frac{m+2}{2})}}(x) & \text{if } m \text{ even } \&\, i = \frac{m+2}{2}, \ldots, m \end{cases} \tag{3}$$

The cdf of $W_i^{(1)}$ is obtained by integrating $f_{W_i^{(1)}}(x)$.

Double median RSS Let $W_1^{(2)}, W_2^{(2)}, \ldots, W_m^{(2)}$ be a DMRSS; that is

$$W_i^{(2)} = \begin{cases} W^{(1)}_{(\frac{m+1}{2})} & \text{if } m \text{ odd } \& \, i = 1, \ldots, m \\ W^{(1)}_{(\frac{m}{2})} & \text{if } m \text{ even } \& \, i = 1, \ldots, \frac{m}{2} \\ W^{(1)}_{(\frac{m+2}{2})} & \text{if } m \text{ even } \& \, i = \frac{m+2}{2}, \ldots, m \end{cases}$$

The pdf of $W_i^{(2)}$ is

$$f_{W_i^{(2)}}(x) = \begin{cases} f_{W^{(1)}_{(\frac{m+1}{2})}}(x) & \text{if } m \text{ odd } \& \, i = 1, \ldots, m \\ f_{W^{(1)}_{(\frac{m}{2})}}(x) & \text{if } m \text{ even } \& \, i = 1, \ldots, \frac{m}{2} \\ f_{W^{(1)}_{(\frac{m+2}{2})}}(x) & \text{if } m \text{ even } \& \, i = \frac{m+2}{2}, \ldots, m \end{cases}$$

The cdf of $W_i^{(2)}$ is obtained by integrating $f_{W_i^{(2)}}(x)$.

3. Hellinger distance

Suppose Y and X are two random variables with density functions $f_Y(x)$ and $f_X(x)$, respectively. The Hellinger distance (See for example Nikulin (2001)) between Y and X is defined by

$$H(X,Y) = \left(1 - \int_{-\infty}^{\infty} \sqrt{f_Y(x) f_X(x)} dx\right)^{\frac{1}{2}}. \tag{4}$$

Obviously, for independent and identical random variables, $H(X,Y) = 0$. So the Hellinger distance between any two data points of the SRS $X_1, X_2, \ldots, X_m$ is zero. Therefore, identifying the ordered data points (for getting either RSS or MRSS) based on the SRS is difficult.

Now, given the data points of the RSS $(Y_1^{(1)}, Y_2^{(1)}, \ldots, Y_m^{(1)})$, then for $k, l = 1, 2, \ldots, m$,

$$\begin{aligned} H^2\left(Y_k^{(1)}, Y_l^{(1)}\right) &= 1 - \int_{-\infty}^{\infty} \sqrt{f_{Y_k}(y) f_{Y_l}(y)} dy \\ &= 1 - \int_{-\infty}^{\infty} \sqrt{m^2 f^2(y) \binom{m-1}{k-1} \binom{m-1}{l-1} F^{k+l-2}(y) \left(1 - F(y)\right)^{2m-k-l}} dy. \end{aligned}$$

Let $F(y) = u$ and $du = f(y)dy$, then

$$\begin{aligned} H^2\left(Y_k^{(1)}, Y_l^{(1)}\right) &= 1 - m\sqrt{\binom{m-1}{k-1}\binom{m-1}{l-1}} \int_0^1 u^{\frac{k+l}{2}-1} (1-u)^{m-\frac{k+l}{2}} du \\ &= 1 - m\sqrt{\binom{m-1}{k-1}\binom{m-1}{l-1}} \frac{\Gamma\left(\frac{k+l}{2}\right)\Gamma\left(m - \frac{k+l}{2} + 1\right)}{\Gamma(m+1)} \\ &= 1 - \frac{\left(\frac{k+l}{2} - 1\right)!\left(m - \frac{k+l}{2}\right)!}{\sqrt{(m-k)!(m-l)!(k-1)!(l-1)!}}. \end{aligned}$$

The results for particular values of m, k, and l are shown in the third column of Table 1. Note that the Hellinger distances in this case are not zeros; that is, the further work of identifying the ordered data points of DRSS (i.e., for stage 2) based on the RSS data points (stage 1) is easier now than using SRS data points. It is simple to verify that when $|k - l| = 2$,

$$H^2\left(Y_k^{(1)}, Y_l^{(1)}\right) = 1 - \sqrt{\frac{k(m-k-1)}{(k+1)(m-k)}}.$$

Now, given the data points of the MRSS $(W_1^{(1)}, W_2^{(1)}, \ldots, W_m^{(1)})$, and suppose m is odd. According to Eq. (2) and Eq. (3) due to the iid case, $H\left(W_k^{(1)}, W_l^{(1)}\right) = 0$ for each $k, l = 1, 2, \ldots, m$. Therefore, getting a DMRSS based on the MRSS practically is the same as obtaining a MRSS based on the SRS. But if m is even, according to Eq. (2) and Eq. (3), the Hellinger distance is given by

$$H\left(W_k^{(1)}, W_l^{(1)}\right) = \begin{cases} H(W_{\frac{m}{2}}^{(1)}, W_{\frac{m+2}{2}}^{(1)}) > 0 & \text{if } k \leq \frac{m}{2} \,\&\, l > \frac{m}{2} \\ 0 & \text{otherwise} \end{cases}$$

Now suppose $Y_1^{(2)}, Y_2^{(2)}, \ldots, Y_m^{(2)}$ be a DRSS, then it can be seen from Eq. (1) that $F_{Y_i^{(2)}}(x)$ is a function of $F(x)$; that is $F_{Y_i^{(2)}}(x) = G_i(F(x))$ and $f_{Y_i^{(2)}}(x) = f(x) g_i(F(x))$, where $g_i(\cdot) = \frac{d}{dx} G_i(\cdot)$. Hence

$$\begin{aligned} H^2\left(Y_k^{(2)}, Y_l^{(2)}\right) &= 1 - \int_{-\infty}^{\infty} \sqrt{f_{Z_k}(x) f_{Z_l}(x)} dx \\ &= 1 - \int_{-\infty}^{\infty} \sqrt{g_k(F(x))\, g_l(F(x))} f(x) dx. \end{aligned}$$

Let $F(x) = u$ and $du/dx = f(x)$, then

$$H^2\left(Y_k^{(2)}, Y_l^{(2)}\right) = 1 - \int_0^1 \sqrt{g_k(u) g_l(u)} du$$

In this case, it is clear from the last column of Table 1 that Hellinger distances are getting higher than in stage 1 (Column 3).

Table 1: Hellinger Distances, $m = 2, 3, 4$; 1^{st} and 2^{nd} stage

m	(k, l)	stage 1	stage 2
2	$(1, 2)$	0.4633	0.5920
3	$(1, 2)$	0.4086	0.5473
	$(1, 3)$	0.7071	0.8625
	$(2, 3)$	0.4086	0.5473
4	$(1, 2)$	0.3870	0.5306
	$(1, 3)$	0.6501	0.8304
	$(1, 4)$	0.8399	0.9628
	$(2, 3)$	0.3412	0.4889
	$(2, 4)$	0.6501	0.8304
	$(3, 4)$	0.3870	0.5306

Due to the properties of order statistics $Z_1, \ldots, Z_m$, it can be seen that $H(Z_1, Z_m)$ is the largest distance and $H(Z_{\frac{m}{2}}, Z_{\frac{m+2}{2}})$ is the minimum distance. Also note that $H(Z_1, Z_{1+r}) = H(Z_{m-r}, Z_m)$, $1 < r < m$. Apparently increasing m decreases the Hellinger distances for the same pair of order statistics; which is reasonable in the sense that identifying the ordered data point from a small m is easier than in a large m. It can also be concluded from Table 1 that identifying the ordered data points for stage 2 (DRSS) based on the ordered data points of stage 1 (RSS) is consistently easier than identifying the ordered data points for stage 1 (RSS) based on the identical data points of SRS. This result is consistent with findings in Al-Saleh and Al-Kadiri (2000).

4. Conclusion

For a single stage sampling, MRSS and RSS have the same practicality, and since it is shown in the literature that MRSS is more efficient than RSS we recommend to use MRSS. For a second stage sampling, although it is shown in the literature that DMRSS is more efficient than DRSS, we recommend to use DRSS because it is more practical than DMRSS. This will speed up the visual ranking process and reduce the ranking error, and therefore identify the data points quickly.

Acknowledgments The authors wish to thank the associated editor and referees for their constructive comments which improve the final version of this paper. The authors also would like to acknowledge excellent research support from KFUPM grant number SR151009.

References

Al-Saleh MF, Al-Kadiri MA (2000). "Double-Ranked Set Sampling." *Statistics & Probability Letters*, **48**(2), 205–212. doi:10.1016/S0167-7152(99)00206-0.

Al-Saleh MF, Diab YA (2009). "Estimation of the Parameters of Downton's Bivariate Exponential Distribution Using Ranked Set Sampling Scheme." *Journal of Statistical Planning and Inference*, **139**(2), 277–286. doi:10.1016/j.jspi.2008.04.021.

Amro L, Samuh MH (2017). "More Powerful Permutation Test Based on Multistage Ranked Set Sampling." *Communications in Statistics - Simulation and Computation*, **46**(7), 5271–5284. doi:10.1080/03610918.2016.1152364.

David HA, Nagaraja HN (2003). *Order Statistics, 3rd Edition*. Wiley, New York.

Mahdizadeh M, Zamanzade E (2017a). "Estimation of a Symmetric Distribution Function in Multistage Ranked Set Sampling." *Statistical Papers*. doi:10.1007/s00362-017-0965-x.

Mahdizadeh M, Zamanzade E (2017b). "Reliability Estimation in Multistage Ranked Set Sampling." *REVSTAT: A Statistical Journal*, **15**(4), 565–581.

McIntyre GA (1952). "A Method for Unbiased Selective Sampling, Using Ranked Sets." *Australian Journal of Agricultural Research*, **3**(4), 385–390. doi:10.1071/AR9520385.

McIntyre GA (2005). "A Method for Unbiased Selective Sampling, Using Ranked Sets." *The American Statistician*, **59**(3), 230–232. doi:10.1198/000313005X54180.

Muttlak HA (1997). "Median Ranked Set Sampling." *Journal of Applied Statistical Science*, **6**, 245–255.

Nikulin MS (2001). "Hellinger Distance." *Encyclopedia of Mathematics*.

Samawi HM, Tawalbeh EM (2002). "Double Median Ranked Set Sample: Comparing to Other Double Ranked Samples for Mean and Ratio Estimators." *Journal of Modern Applied Statistical Methods*, **1**(2), 428–442. doi:10.22237/jmasm/1036109460.

Samuh MH (2018). "Estimation of Distribution Function Under Multistage Median Ranked Set Sampling." *Journal of Applied Probability and Statistics*, **13**(2), 41–59.

Sarikavanij S, Kasala S, Sinha BK, Tiensuwan M (2014). "Estimation of Location and Scale Parameters in Two-Parameter Exponential Distribution Based on Ranked Set Sample." *Communications in Statistics-Simulation and Computation*, **43**(1), 132–141. doi:10.1080/03610918.2012.698776.

Stokes SL, Sager TW (1988). "Characterization of a Ranked-Set Sample With Application to Estimating Distribution Functions." *Journal of the American Statistical Association*, **83**(402), 374–381. doi:10.2307/2288852.

Takahasi K, 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), 1–31. doi:10.1007/BF02911622.

Affiliation:

Monjed H. Samuh
Department of Applied Mathematics & Physics
Palestine Polytechnic University
Hebron, Palestine
E-mail: monjedsamuh@ppu.edu
URL: http://staff.ppu.edu/monjedsamuh

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 49 *Submitted:* 2018-11-20
January 2020 *Accepted:* 2019-04-10