RSSampling: A Pioneering Package for Ranked Set Sampling

by Busra Sevinc, Bekir Cetintav, Melek Esemen, and Selma Gurler

Abstract Ranked set sampling (RSS) is an advanced data collection method when the exact measurement of an observation is difficult and/or expensive used in a number of research areas, e.g., environment, bioinformatics, ecology, etc. In this method, random sets are drawn from a population and the units in sets are ranked with a ranking mechanism which is based on a visual inspection or a concomitant variable. Because of the importance of working with a good design and easy analysis, there is a need for a software tool which provides sampling designs and statistical inferences based on RSS and its modifications. This paper introduces an R package as a free and easy-to-use analysis tool for both sampling processes and statistical inferences based on RSS and its modified versions. For researchers, the **RSSampling** package provides a sample with RSS, extreme RSS, median RSS, percentile RSS, balanced groups RSS, double versions of RSS, L-RSS, truncation-based RSS, and robust extreme RSS when the judgment rankings are both perfect and imperfect. Researchers can also use this new package to make parametric inferences for the population mean and the variance where the sample is obtained via classical RSS. Moreover, this package includes applications of the nonparametric methods which are one sample sign test, Mann-Whitney-Wilcoxon test, and Wilcoxon signed-rank test procedures. The package is available as **RSSampling** on CRAN.

Introduction

Data collection is the crucial part in all types of scientific research. Ranked set sampling (RSS) is one of the advanced data collection methods, which provides representative sample data by using the ranking information of the sample units. It was firstly proposed by **?** and the term "ranked set sampling" was introduced in the study of **?** about the estimation of forage yields in a pine hardwood forest. **?** theoretically studied the efficiency of the mean estimator based on RSS which is unbiased for the population mean. They found that its variance is always smaller than the variance of the mean estimator based on simple random sampling (SRS) with the same sample size when the ranking is perfect. Some other results on the efficiency of RSS can be found in **?**, **?**, and **?**. **?** studied the use of concomitant variables for ranking of the sample units in the RSS procedure and found that the ranking procedure was allowed to be imperfect. In another study, she constructed the estimator for the population variance in the presence of the ranking error (**?**). For some examples and results on the regression estimation based on RSS, see, **?** and **?**. The estimation of a distribution function with various settings of RSS can be found in **?**, **?**, and **?**. Other results on distribution-free test procedures based on RSS can be found in **??**, and **?**. Additional results for inferential procedures based on RSS can be found in the recent works of **?**, **?**, and **?**. For more details on RSS, we refer the review papers by **?**, **?**, and **?**.

The RSS method and its modified versions have come into prominence recently due to its efficiency and therefore new software tools or packages for a quick evaluation is required. A free software called Visual Sample Plan (VSP) created by Pacific Northwest National Labaratory has many sampling designs including classical RSS method for developing environmental sampling plans under balanced and unbalanced cases. It provides the calculation of the required sample size and cost information with the location to be sampled. Also, a package NSM3 by **?** in R has two functions related to classical RSS method. It only provides the Monte Carlo samples and computes a statistic for a nonparametric procedure. Both the VSP and **NSM3** package include only the classical RSS method as a sampling procedure and provide limited methods for inference. Therefore, there is no extensive package for sampling and statistical inference using both classical and modified RSS methods in any available software packages. In this study, we propose a pioneering package, named RSSampling, for sampling procedures based on the classical RSS and the modified RSS methods in both perfect and imperfect ranking cases. Also, the package provides the estimation of the mean and the variance of the population and allows the use of the one sample sign, Mann-Whitney-Wilcoxon, and Wilcoxon signed-rank test procedures under classical RSS. The organization of the paper is as follows: in the following section, we give some brief information about classical RSS and modified RSS methods. Then, we introduce the details of **RSSampling** package and further, we give some illustrative examples with a real data analysis. In the last section, we give the conclusion of the study.

The classical and modified RSS methods

RSS and its modifications are advanced sampling methods using the rank information of the sample units. The ranking of the units can be done by visual inspection of a human expert or a concomitant variable. The procedure for the RSS method is as follows:

1. Select m units at random from a specified population.
2. Rank these m units by judgment without actual measurement.
3. Keep the smallest judged unit from the ranked set.
4. Select second set of m units at random from a specified population, rank these units without measuring them, keep the second smallest judged unit.
5. Continue the process until m ranked units are measured.

The first five steps are referred to as a cycle. Then, the cycle repeats r times and a ranked set sample of size $n = mr$ is obtained. Figure 1 illustrates the RSS procedure with visual inspection for the case of $r = 1$ and $m = 3$, and in the following scheme, $X_{i(j:m)}$ represents the jth ranked unit in ith set where $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, m$ and bold units represent the units which are chosen to ranked set sample.

$$\begin{pmatrix} \mathbf{X_{1(1:3)}} \leq & X_{1(2:3)} \leq & X_{1(3:3)} \\ X_{2(1:3)} \leq & \mathbf{X_{2(2:3)}} \leq & X_{2(3:3)} \\ X_{3(1:3)} \leq & X_{3(2:3)} \leq & \mathbf{X_{3(3:3)}} \end{pmatrix}$$

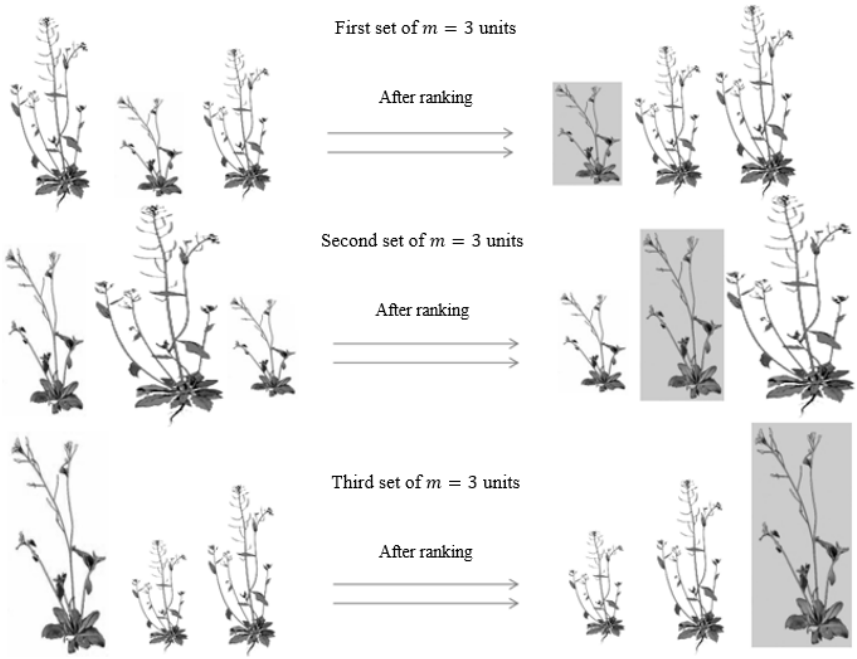

Figure 1: Ranking with visual inspection for one cycle, **?**

RSS design obtains more representative samples and gives more precise estimates of the population parameters relative to SRS (**?**). The main difference between the RSS method and the other modified methods is the selection procedure of the sample units from the ranked sets. For example, **?** suggested extreme RSS using the minimum or maximum units from each ranked set. **?** introduced median RSS using only median units of the random sets. **?** suggested balanced groups RSS which is defined as the combination of extreme RSS and median RSS. For additional examples of modified methods, see **?**, **?**, and for robust methods see, **?**, **?**, and **?**. In literature, the studies for modified RSS methods are generally interested in obtaining a sample more easily or making a more robust estimation for a population parameter. Such studies are made for the investigation of properties (for example, bias and mean squared error) of a proposed estimator and they have generally focused on the comparisons of SRS and RSS methods. Note that the true comparisons of the modified RSS methods to the others are difficult to present in general terms. Because the advantages of the sampling methods, when compared to each other, may vary according to the situations such as the parameter to be estimated, underlying distribution, the presence of ranking error, etc. For more detailed information on the modifications of

RSS, see **?** and references therein. In the following, the modified RSS methods which are considered in **RSSampling** are introduced.

Extreme RSS

Extreme RSS (ERSS) is the first modification of RSS suggested by **?** to estimate the population mean only using the minimum or maximum ranked units from each set. The procedure for ERSS can be described as follows: select m random sets each of size m units from the population and rank the units within each set by a human expert or a concomitant variable. If the set size m is even, the lowest ranked units of each set are chosen from the first $m/2$ sets, and the largest ranked units of each set are chosen from the other $m/2$ sets. If the set size is odd, the lowest ranked units from the first $(m-1)/2$ sets, the largest ranked units from the other $(m-1)/2$ sets and median unit from the remaining last set are chosen. If we repeat the procedure r times, we have a sample of size $n = mr$. An example of the procedure for $r = 1$ and $m = 4$ is shown below.

$$\begin{pmatrix} \mathbf{X_{1(1:4)}} \leq & X_{1(2:4)} \leq & X_{1(3:4)} \leq & X_{1(4:4)} \\ \mathbf{X_{2(1:4)}} \leq & X_{2(2:4)} \leq & X_{2(3:4)} \leq & X_{2(4:4)} \\ X_{3(1:4)} \leq & X_{3(2:4)} \leq & X_{3(3:4)} \leq & \mathbf{X_{3(4:4)}} \\ X_{4(1:4)} \leq & X_{4(2:4)} \leq & X_{4(3:4)} \leq & \mathbf{X_{4(4:4)}} \end{pmatrix}$$

Median RSS

Median RSS (MRSS) was suggested by **?**. In this method, only median units of the random sets are chosen as the sample for estimation of population mean. For the odd set sizes, the $((m+1)/2)$th ranked units are chosen as the median of each set. For even set sizes, the $(m/2)$th ranked units are chosen from the first $m/2$ sets and the $((m+2)/2)$th ranked units are chosen from the remaining $m/2$ sets. If necessary, procedure can be repeated r times and we have $n = mr$ sample of size. An example of the procedure for $r = 1$ and $m = 3$ is shown below.

$$\begin{pmatrix} X_{1(1:3)} \leq & \mathbf{X_{1(2:3)}} \leq & X_{1(3:3)} \\ X_{2(1:3)} \leq & \mathbf{X_{2(2:3)}} \leq & X_{2(3:3)} \\ X_{3(1:3)} \leq & \mathbf{X_{3(2:3)}} \leq & X_{3(3:3)} \end{pmatrix}$$

Percentile RSS

? suggested another modification for the RSS, percentile RSS (PRSS), where only the upper and lower percentiles of the random sets are chosen as the sample for selected value of p, where $0 \leq p \leq 1$. Suppose that m random sets with the size m are chosen from a specific population to sample m units and ranked visually or with a concomitant variable. If the set size is even, the $(p(m+1))$th smallest units from the first $m/2$ sets and the $((1-p)(m+1))$th smallest units from the other $m/2$ sets are chosen. If m is odd, the $(p(m+1))$th smallest units are chosen from the first $(m-1)/2$ sets, the $((1-p)(m+1))$th smallest units are chosen from the other $(m-1)/2$ sets and the median unit is chosen as the mth unit from the last set. An example of the procedure for $r = 1$, $m = 5$ and $p = 0.3$ is as below.

$$\begin{pmatrix} X_{1(1:5)} \leq & \mathbf{X_{1(2:5)}} \leq & X_{1(3:5)} \leq & X_{1(4:5)} \leq & X_{1(5:5)} \\ X_{2(1:5)} \leq & \mathbf{X_{2(2:5)}} \leq & X_{2(3:5)} \leq & X_{2(4:5)} \leq & X_{2(5:5)} \\ X_{3(1:5)} \leq & X_{3(2:5)} \leq & X_{3(3:5)} \leq & \mathbf{X_{3(4:5)}} \leq & X_{3(5:5)} \\ X_{4(1:5)} \leq & X_{4(2:5)} \leq & X_{4(3:5)} \leq & \mathbf{X_{4(4:5)}} \leq & X_{4(5:5)} \\ X_{5(1:5)} \leq & X_{5(2:5)} \leq & \mathbf{X_{5(3:5)}} \leq & X_{5(4:5)} \leq & X_{5(5:5)} \end{pmatrix}$$

Balanced groups RSS

Balanced groups RSS (BGRSS) can be defined as the combination of ERSS and MRSS. **?** suggested to use BGRSS for estimating the population mean with a special sample size $m = 3k$. In their study, BGRSS procedure can be described as follows: $m = 3k$ (where $k = 1, 2, 3, \dots$) sets each size of m are selected randomly from a specific population. The sets are randomly allocated into three groups and units in each set are ranked. The smallest units from the first group, median units from the second group and the largest units from the third group of ranked sets are chosen. When the set size is odd, the median unit in the second group is defined as the $((m+1)/2)$th ranked unit in the set and when

the set size is even, the median unit is defined as the mean of the $(m/2)$th and the $((m+2)/2)$th ranked units. BGRSS process for one cycle and $k = 2$ can be described as below.

$$\begin{pmatrix} \mathbf{X_{1(1:6)}} \leq & X_{1(2:6)} \leq & X_{1(3:6)} \leq & X_{1(4:6)} \leq & X_{1(5:6)} \leq & X_{1(6:6)} \\ \mathbf{X_{2(1:6)}} \leq & X_{2(2:6)} \leq & X_{2(3:6)} \leq & X_{2(4:6)} \leq & X_{2(5:6)} \leq & X_{2(6:6)} \\ X_{3(1:6)} \leq & X_{3(2:6)} \leq & \mathbf{X_{3(3:6)}} \leq & \mathbf{X_{3(4:6)}} \leq & X_{3(5:6)} \leq & X_{3(6:6)} \\ X_{4(1:6)} \leq & X_{4(2:6)} \leq & \mathbf{X_{4(3:6)}} \leq & \mathbf{X_{4(4:6)}} \leq & X_{4(5:6)} \leq & X_{4(6:6)} \\ X_{5(1:6)} \leq & X_{5(2:6)} \leq & X_{5(3:6)} \leq & X_{5(4:6)} \leq & X_{5(5:6)} \leq & \mathbf{X_{5(6:6)}} \\ X_{6(1:6)} \leq & X_{6(2:6)} \leq & X_{6(3:6)} \leq & X_{6(4:6)} \leq & X_{6(5:6)} \leq & \mathbf{X_{6(6:6)}} \end{pmatrix}$$

Double RSS

? introduced another modification of RSS, that is double RSS (DRSS) as a beginning of multistage procedure. Several researchers also extended the DRSS method to modified versions such as double extreme RSS (DERSS) by **?**, double median RSS (DMRSS) by **?**, and double percentile RSS (DPRSS) by **?**. The DRSS procedure is described as follows: m^3 units are identified from the target population and divided randomly into m groups, the size of each is m^2. Then, the usual RSS procedure is used on each group to obtain m ranked set samples each of size m. Finally, RSS procedure is applied again on the obtained ranked set samples in the previous step to get a double ranked set sample of size m.

L-RSS

L-RSS, which is a robust RSS procedure, is based on the idea of L statistic and it was introduced by **?** as a generalization of different type of RSS methods. The first step for L-RSS procedure is selecting m random sets with m units and ranking the units in each set. Let k be the L-RSS coefficient, where $k = \lfloor m\alpha \rfloor$ for $0 \leq \alpha < 0.5$ and $\lfloor m\alpha \rfloor$ is the largest integer value less than or equal to $m\alpha$. Then, the $(k+1)$th ranked units from the first $k+1$ sets, $(m-k)$th ranked units from the last $k+1$ sets and ith ranked units from the remaining sets which are numbered with i, where $i = k+2, \ldots, m-k-1$ are selected. The L-RSS procedure for the case of $m = 6$ and $k = 1$ ($\alpha = 0.20$) in a cycle can be shown as below:

$$\begin{pmatrix} X_{1(1:6)} \leq & \mathbf{X_{1(2:6)}} \leq & X_{1(3:6)} \leq & X_{1(4:6)} \leq & X_{1(5:6)} \leq & X_{1(6:6)} \\ X_{2(1:6)} \leq & \mathbf{X_{2(2:6)}} \leq & X_{2(3:6)} \leq & X_{2(4:6)} \leq & X_{2(5:6)} \leq & X_{2(6:6)} \\ X_{3(1:6)} \leq & X_{3(2:6)} \leq & \mathbf{X_{3(3:6)}} \leq & X_{3(4:6)} \leq & X_{3(5:6)} \leq & X_{3(6:6)} \\ X_{4(1:6)} \leq & X_{4(2:6)} \leq & X_{4(3:6)} \leq & \mathbf{X_{4(4:6)}} \leq & X_{4(5:6)} \leq & X_{4(6:6)} \\ X_{5(1:6)} \leq & X_{5(2:6)} \leq & X_{5(3:6)} \leq & X_{5(4:6)} \leq & \mathbf{X_{5(5:6)}} \leq & X_{5(6:6)} \\ X_{6(1:6)} \leq & X_{6(2:6)} \leq & X_{6(3:6)} \leq & X_{6(4:6)} \leq & \mathbf{X_{6(5:6)}} \leq & X_{6(6:6)} \end{pmatrix}$$

When $k = 0$, then this procedure leads to the classical RSS and when $k = \lfloor (m-1)/2 \rfloor$, then it leads to the MRSS method.

Truncation-based RSS

The truncation-based RSS (TBRSS) was presented by **?**. This procedure can be summarized as follows: select randomly m sets each of size m units from the population and rank the units in each set. Then, determine TBRSS coefficient k as in the L-RSS method and select the minimums of the first k sets and the maximums of the last k sets. From the remaining $m - 2k$ samples, select the ith ranked unit of the ith sample $(k+1 \leq i \leq m-k)$. The one cycled TBRSS method for the case of $m = 8$ and $k = 2$ ($\alpha = 0.35$) is shown below.

$$\begin{pmatrix} \mathbf{X_{1(1:8)}} \leq & X_{1(2:8)} \leq & X_{1(3:8)} \leq & X_{1(4:8)} \leq & X_{1(5:8)} \leq & X_{1(6:8)} \leq & X_{1(7:8)} \leq & X_{1(8:8)} \\ \mathbf{X_{2(1:8)}} \leq & X_{2(2:8)} \leq & X_{2(3:8)} \leq & X_{2(4:8)} \leq & X_{2(5:8)} \leq & X_{2(6:8)} \leq & X_{2(7:8)} \leq & X_{2(8:8)} \\ X_{3(1:8)} \leq & X_{3(2:8)} \leq & \mathbf{X_{3(3:8)}} \leq & X_{3(4:8)} \leq & X_{3(5:8)} \leq & X_{3(6:8)} \leq & X_{3(7:8)} \leq & X_{3(8:8)} \\ X_{4(1:8)} \leq & X_{4(2:8)} \leq & X_{4(3:8)} \leq & \mathbf{X_{4(4:8)}} \leq & X_{4(5:8)} \leq & X_{4(6:8)} \leq & X_{4(7:8)} \leq & X_{4(8:8)} \\ X_{5(1:8)} \leq & X_{5(2:8)} \leq & X_{5(3:8)} \leq & X_{5(4:8)} \leq & \mathbf{X_{5(5:8)}} \leq & X_{5(6:8)} \leq & X_{5(7:8)} \leq & X_{5(8:8)} \\ X_{6(1:8)} \leq & X_{6(2:8)} \leq & X_{6(3:8)} \leq & X_{6(4:8)} \leq & X_{6(5:8)} \leq & \mathbf{X_{6(6:8)}} \leq & X_{6(7:8)} \leq & X_{6(8:8)} \\ X_{7(1:8)} \leq & X_{7(2:8)} \leq & X_{7(3:8)} \leq & X_{7(4:8)} \leq & X_{7(5:8)} \leq & X_{7(6:8)} \leq & X_{7(7:8)} \leq & \mathbf{X_{7(8:8)}} \\ X_{8(1:8)} \leq & X_{8(2:8)} \leq & X_{8(3:8)} \leq & X_{8(4:8)} \leq & X_{8(5:8)} \leq & X_{8(6:8)} \leq & X_{8(7:8)} \leq & \mathbf{X_{8(8:8)}} \end{pmatrix}$$

Note that when $k = 0$ or $k = 1$, TBRSS scheme is equivalent to the classical RSS scheme.

Robust extreme RSS

Robust extreme RSS (RERSS) scheme was introduced by **?**. This method can be described as follows: identify m random sets with m units and rank the units within each set. Select the $(k+1)$th ranked units from the first $m/2$ sets where $k = \lfloor m\alpha \rfloor$ for $0 < \alpha < 0.5$ and $\lfloor m\alpha \rfloor$ is the largest integer value less than or equal to $m\alpha$. Then, select the $(m-k)$th ranked units from the other $m/2$ sets. If the set size m is odd, $((m+1)/2)$th ranked unit is selected additionally from the last remaining set. The procedure for one cycle and the case of $m = 6$ and $k = 1$ ($\alpha = 0.20$) can be shown as below.

$$\begin{pmatrix} X_{1(1:6)} \leq & \mathbf{X_{1(2:6)}} \leq & X_{1(3:6)} \leq & X_{1(4:6)} \leq & X_{1(5:6)} \leq & X_{1(6:6)} \\ X_{2(1:6)} \leq & \mathbf{X_{2(2:6)}} \leq & X_{2(3:6)} \leq & X_{2(4:6)} \leq & X_{2(5:6)} \leq & X_{2(6:6)} \\ X_{3(1:6)} \leq & \mathbf{X_{3(2:6)}} \leq & X_{3(3:6)} \leq & X_{3(4:6)} \leq & X_{3(5:6)} \leq & X_{3(6:6)} \\ X_{4(1:6)} \leq & X_{4(2:6)} \leq & X_{4(3:6)} \leq & X_{4(4:6)} \leq & \mathbf{X_{4(5:6)}} \leq & X_{4(6:6)} \\ X_{5(1:6)} \leq & X_{5(2:6)} \leq & X_{5(3:6)} \leq & X_{5(4:6)} \leq & \mathbf{X_{5(5:6)}} \leq & X_{5(6:6)} \\ X_{6(1:6)} \leq & X_{6(2:6)} \leq & X_{6(3:6)} \leq & X_{6(4:6)} \leq & \mathbf{X_{6(5:6)}} \leq & X_{6(6:6)} \end{pmatrix}$$

If $k = 0$ and $k = (m/2)$, then this sampling procedure corresponds to ERSS and MRSS methods, respectively.

RSSampling package

The package **RSSampling** is available on CRAN and can be installed and loaded via the following commands:

```
> install.packages("RSSampling")
> library("RSSampling")
```

The package depends on the **stats** package and uses a function from the non-standard package **LearnBayes** (**?**) for random data generation in the Examples section. The proposed package consists of two main parts which are the functions for sampling methods described in Table 1 and the functions for inference procedures described in Table 2 based on RSS. The sampling part of the package includes perfect and imperfect rankings with a concomitant variable allowing researchers to sample with classical RSS and the modified versions. The functions for inference procedures provide estimation for parameters and some hypothesis testing procedures based on RSS.

Sampling with RSSampling

In this part, we introduce a core function, which is called `rankedsets`, to obtain `s` ranked sets consisting of randomly chosen sample units with the set size m. By using this function, we developed the functions given in Table 1 which provide researchers means to obtain a sample under different sampling schemes. One can also use `rankedsets` function for the studies based on other modified RSS methods which are not mentioned in this paper.

Function	Description
`rss`	Performs classical RSS method
`Mrss`	Performs modified RSS methods (MRSS, ERSS, PRSS,BGRSS)
`Rrss`	Performs robust RSS methods (L-RSS, TBRSS, RERSS)
`Drss`	Performs double RSS methods (DRSS, DMRSS, DERSS, DPRSS)
`con.rss`	Performs classical RSS method by using a concomitant variable
`con.Mrss`	Performs modified RSS methods (MRSS, ERSS, PRSS,BGRSS) by using a concomitant variable
`con.Rrss`	Performs robust RSS methods (L-RSS, TBRSS, RERSS) by using a concomitant variable
`obsno.Mrss`	Determines the observation numbers of the units which will be chosen to the sample for classical and modified RSS methods by using a concomitant variable

Table 1: The functions for the sampling methods in **RSSampling** package

The function `rss` provides the ranked set sample with perfect ranking from a specific data set, X, provided in matrix form where the columns and rows represent the sets and cycles, respectively. One

can see the randomly chosen ranked sets by defining `sets = TRUE` (default `sets = FALSE`) with the set size m and the cycle size r. For the modified RSS methods, the function `Mrss` provides a sample from MRSS, ERSS, PRSS, and BGRSS which are represented by `"m"`,`"e"`, `"p"`, and `"bg"`, respectively. The `type = "r"`, defined as the default, represents the classical RSS. For the sampling procedure PRSS, there is an additional parameter `p` which defines the percentile. We note that, when `p = 0.25` in PRSS, one can obtain a sample with quartile RSS given by **?**. `Rrss` provides samples from L-RSS, TBRSS, and RERSS methods which are represented by `"l"`, `"tb"`, and `"re"`, respectively. The parameter `alpha` is the common parameter for these methods and defines the cutting value. `Drss` function is for double versions of RSS, MRSS, ERSS, and PRSS under perfect ranking. `type = "d"` is defined as the default which represents the double RSS. Values `"dm"`, `"de"`, and `"dp"` are defined for DMRSS, DERSS, and DPRSS methods, respectively.

In the literature, most of the theoretical inferences and numerical studies are conducted based on perfect ranking. However, in real life applications, the ranking process is done with an expert judgment or a concomitant variable. Let us consider RSS with a concomitant variable Y. A set of m units is drawn from the population, then the units are ranked by the order of Y. The concomitant variable $Y_{i(j:m)}$ represents the jth ranked unit in ith set and the variable of interest $X_{(i,j)}$ represents the jth unit in ith set, where $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, m$. In the following example, the procedure of RSS using Y is given for $m = 3$.

$$\begin{array}{llll} (\mathbf{Y_{1(1:3)}}, \mathbf{X_{(1,1)}}) \leq & (Y_{1(2:3)}, X_{(1,2)}) \leq & (Y_{1(3:3)}, X_{(1,3)}) & \longrightarrow \quad \mathbf{X_{(1,1)}} \\ (Y_{2(1:3)}, X_{(2,1)}) \leq & (\mathbf{Y_{2(2:3)}}, \mathbf{X_{(2,2)}}) \leq & (Y_{2(3:3)}, X_{(2,3)}) & \longrightarrow \quad \mathbf{X_{(2,2)}} \\ (Y_{3(1:3)}, X_{(3,1)}) \leq & (Y_{3(2:3)}, X_{(3,2)}) \leq & (\mathbf{Y_{3(3:3)}}, \mathbf{X_{(3,3)}}) & \longrightarrow \quad \mathbf{X_{(3,3)}} \end{array}$$

The functions `con.rss`, `con.Mrss`, and `con.Rrss` provide methods to obtain a sample under imperfect ranking. With the `con.rss` function, a researcher can obtain a classical ranked set sample from a specific data set using a concomitant variable Y with the set size m and cycle size r to make inference about the variable of interest X. The functions `con.Mrss` and `con.Rrss` have similar usage with `con.rss` function except the selection method which is defined by `type` parameter. Also, these functions are simply extensions of the `Mrss` and `Rrss` for concomitant variable cases.

In a real-world research, the values of the variable of interest X are unknown and the researchers measure X values of the sample units after choosing them from the population with a specific sampling method. The function `obsno.Mrss` provides the code for this kind of application, when the researchers prefer to use RSS methods. After determining the sample frame and the concomitant variable to be used for ranking, the code provides the number of the units to be selected according the values of the concomitant variable. Then, the researcher obtain easily the observation numbers of the units which will be chosen to the sample. `type = "r"` is defined as the default which represents the classical RSS. MRSS, ERSS, PRSS, and BGRSS are represented by `"m"` , `"e"`, `"p"`, and `"bg"`, respectively.

Inference with RSSampling

Statistical inference refers to the process of drawing conclusions and having an information about the interested population. Researchers are generally interested in fundamental inferences for the parameters such as mean and variance. Using the **RSSampling** package, we provide an easy way to estimate the parameters about the interested population and to use some distribution-free tests; namely the sign, Mann-Whitney-Wilcoxon, and Wilcoxon signed-rank tests for nonparametric inference when the sampling procedure is RSS.

Function	Description
`meanRSS`	Performs mean estimation and hypothesis testing with classical RSS method
`varRSS`	Performs variance estimation with classical RSS method
`regRSS`	Performs regression estimation for mean of interested population with classical RSS method
`sign1testrss`	Performs one sample sign test with classical RSS method
`mwwutestrss`	Performs Mann-Whitney-Wilcoxon test with classical RSS method
`wsrtestrss`	Performs Wilcoxon signed-rank test with classical RSS method

Table 2: The functions for inference in **RSSampling** package

The `meanRSS` function provides point estimation, confidence interval estimation, and asymptotic hypothesis testing for the population mean based on RSS see, (**?**). For the variance estimation based on RSS, we define `varRSS` function which has two `type` parameters; `"Stokes"` and `"Montip"`. **?** proved that estimator of variance is asymptotically unbiased regardless of presence of ranking error. For

the "Montip" type estimation, **?** showed that there is no unbiased estimator of variance for one cycle but they proposed unbiased estimator of variance for more than one cycle. With regRSS function, regression estimator for mean of interested population can be obtained based on RSS. The β coefficient ("B" in regRSS function) is calculated under the assumption of known population mean for concomitant Y. Note that, the ranked set samples for interested variable X and for concomitant variable Y must be the same length. One can find the detailed information about regression estimator based on RSS in **?**.

Finally, for nonparametric inference, sign1testrss, mwwutestrss, and wsrtestrss functions implement, respectively, the sign test, the Mann-Whitney-Wilcoxon test, and the Wilcoxon signed-rank test depending on RSS. The normal approximation is used to construct the test statistics and an approximate confidence intervals. For detailed information on these test methods, see the book of **?**.

Examples

In this section, we present examples illustrating the **RSSampling** package.

Sampling with TBRSS using a concomitant variable

This example shows the process to obtain a sample by using TBRSS method for the variable of interest, X, ranked by using the concomitant variable Y assuming that they are distributed as multivariate normal. We determined the set size m is 4 and the cycle size r is 2. The ranked sets of Y and the sets of X are obtained using the function con.Rrss. Thus, the resultant sample for X is given as below.

```
##Loading packages
library("RSSampling")
library("LearnBayes")

## Imperfect ranking example for interested (X) and concomitant (Y) variables
## from multivariate normal dist.
set.seed(1)
mu <- c(10, 8)
variance <- c(5, 3)
a <- matrix(c(1, 0.9, 0.9, 1), 2, 2)
v <- diag(variance)
Sigma <- v%*%a%*%v
x <- rmnorm(10000, mu, Sigma)
xx <- as.numeric(x[,1])
xy <- as.numeric(x[,2])

## Selecting a truncation-based ranked set sample
con.Rrss(xx, xy, m = 4, r = 2, type = "tb", sets = TRUE, concomitant = FALSE,
         alpha = 0.25)

$corr.coef
[1] 0.9040095

$var.of.interest
         [,1]      [,2]      [,3]     [,4]
[1,] 12.332134 13.116611 15.675967 21.72312
[2,] 11.350275  8.846237 10.164005 17.07950
[3,]  4.143757  9.608573  8.708221 11.57671
[4,]  2.284106  9.535388 12.709489 14.11595
[5,]  3.212739  8.089833 11.430411 14.53190
[6,]  6.556222 12.759335 13.210037 11.02219
[7,]  3.337564 -0.864634 12.800243 13.47315
[8,]  5.988893  8.850680 13.208956 15.82731

$concomitant.var.
         [,1]      [,2]      [,3]      [,4]
[1,] 8.034720 10.398398 11.800919 13.754743
[2,] 8.003575  8.118947 11.136804 12.149531
[3,] 4.733177  7.377396  8.866563 11.658837
```

```
[4,] 4.027061  8.008146  9.977435 10.912382
[5,] 3.909958  6.220087  7.564130  8.739562
[6,] 5.893001  8.760754 10.067927 10.244593
[7,] 2.119661  2.813413 10.651769 10.775596
[8,] 5.406154  7.722866  8.602551 10.874853

$sample.x
         m = 1     m = 2     m = 3    m = 4
r = 1 12.332134  8.846237  8.708221 14.11595
r = 2  3.212739 12.759335 12.800243 15.82731
```

Obtaining observation number in MRSS method

Random determination of the sample units is an important task for practitioners. The function `obsno.Mrss` is for the practitioners who have the frame of the population with unknown variable X and known concomitant variable Y. In the following example, the observation numbers for median ranked set sample units are obtained in order to take the measurement of the interested variable X.

```
## Loading packages
library("RSSampling")

## Generating concomitant variable (Y) from exponential dist.
set.seed(5)
y = rexp(10000)

## Determining the observation numbers of the units which are chosen to sample
obsno.Mrss(y, m = 3, r = 5, type = "m")

              m = 1        m = 2        m = 3
      r = 1 "Obs.  2452" "Obs.  6417" "Obs.  3227"
      r = 2 "Obs.  9094" "Obs.  1805" "Obs.  9877"
      r = 3 "Obs.  1333" "Obs.  9252" "Obs.  3219"
      r = 4 "Obs.  6397" "Obs.  7038" "Obs.  5019"
      r = 5 "Obs.  446"  "Obs.  9663" "Obs.  10"
```

A simulation study based on RSS using a concomitant variable

In order to illustrate the usage of the package, we give a simulation study with 10,000 repetitions for mean estimation of X based on RSS method using a concomitant variable. It demonstrates the effect of the correlation level between X and Y on the mean squared error (MSE) of estimation. Samples are obtained when `m = 5` and `r = 10` assuming that X and Y are distributed as multivariate normal. Figure 2 as an output of the simulation study indicates that when the correlation level is increasing, MSE values are decreasing.

```
## Loading packages
library("RSSampling")
library("LearnBayes")

## Imperfect ranking example for interested (X) and concomitant (Y) variables
## from multivariate normal dist.
mu <- c(10, 8)
variance <- c(5, 3)
rho = seq(0, 0.9, 0.1)
se.x = mse.x = numeric()
repeatsize = 10000
for (i in 1:length(rho)) {
    set.seed(1)
    a <- matrix(c(1, rho[i], rho[i], 1), 2, 2)
    v <- diag(variance)
    Sigma <- v%*%a%*%v
    x <- rmnorm(10000, mu, Sigma)
```

```
    xx <- as.numeric(x[,1])
    xy <- as.numeric(x[,2])
    for (j in 1:repeatsize) {
        set.seed(j)
        samplex = con.Mrss(xx, xy, m = 5, r = 10, type = "r", sets = FALSE,
                           concomitant = FALSE)$sample.x
        se.x[j] = (mean(samplex)-mu[1])^2
    }
    mse.x[i] = sum(se.x)/repeatsize
}
plot(rho[-1], mse.x[-1], type = "o", lwd = 2,
     main = "MSE values based on increasing correlation levels",
     xlab = "corr.coef.", ylab = "MSE", cex = 1.5, xaxt = "n")
axis(1, at = seq(0.1, 0.9, by = 0.1))
```

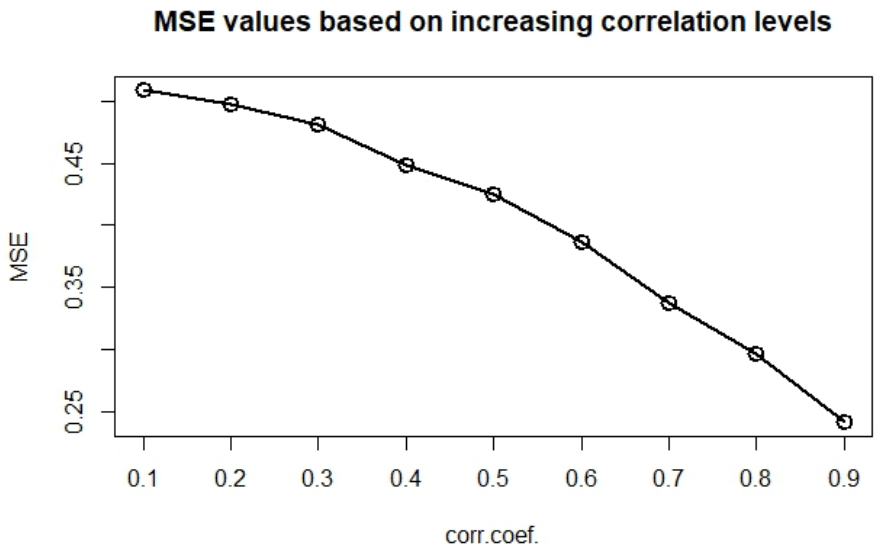

Figure 2: MSE values based on increasing correlation levels

A real data example

In this real data example, we used the abolone data set which is freely available at https://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.data. The data consists of 9 variables of 4177 units and the variables are; sex (Male/Female/Infant), length (mm), diameter (mm), height (mm), whole weight (grams), shucked weight (grams), viscera weight (grams), shell weight (grams), and rings (+1.5 gives the age of abalone in years), respectively. The data comes from an original study of the population biology of abalone by **?**. Also, **?** and **?** used the abalone data set for application of the fuzzy based modification of RSS and partial groups RSS methods, respectively. The data set can be obtained easily by using the following R command.

```
abaloneData <- read.csv(url("https://archive.ics.uci.edu/ml/machine-learning-databases
               /abalone/abalone.data"), header = FALSE, col.names = c("sex", "length",
              "diameter", "height", "whole.weight", "shucked.weight", "viscera.weight",
               "shell.weight", "rings"))
```

Suppose that we aimed to estimate the mean of *viscera weight* and confidence interval and also test the hypothesis claiming that the mean of the *viscera weight* is equal to 0.18. The measurement of *viscera weight* which is the gut weight of abalone after bleeding is an expensive and time-consuming process. Because the measurement of *whole weight* is easy and highly correlated with *viscera weight* (the correlation coefficient is 0.966), we used *whole weight* as the concomitant variable to obtain a sample of size 25 in RSS method. We have the following results for *viscera weight*.

```
cor(abaloneData$viscera.weight, abaloneData$whole.weight)
[1] 0.9663751

set.seed(50)
sampleRSS = con.rss(abaloneData$viscera.weight, abaloneData$whole.weight, m = 5, r = 5,
                    sets = TRUE, concomitant = FALSE)$sample.x

meanRSS(sampleRSS, m = 5, r = 5, alpha = 0.05, alternative = "two.sided", mu_0 = 0.18)
$mean
[1] 0.17826

$CI
[1] 0.1293705 0.2271495

$z.test
[1] -0.06975604

$p.value
[1] 0.9443878

varRSS(sampleRSS, m = 5, r = 5, type = "Stokes")
[1] 0.0135364
```

The results from our sample data indicate that the estimated mean and the variance are 0.17826 and 0.01354, respectively. According to the hypothesis testing result, we conclude that there is no strong evidence against the null hypothesis (`p.value`> 0.05).

Conclusion

RSS is an efficient data collection method compared to SRS especially in situations where the measurement of a unit is expensive but the ranking is less costly. In this study, we propose a package which obtains sample from RSS and its modifications and provide functions to allow some inferential procedures by RSS. We create a set of functions for sampling under both perfect and imperfect rankings with a concomitant variable. For the inferential procedures, we consider mean, variance, and regression estimator and sign, Mann-Whitney-Wilcoxon, and Wilcoxon signed-rank tests for the distribution free tests. Proposed functions in the package are illustrated with the examples and analysis of a real data is given. Future improvements of the package may be provided by adding new inference procedures based on RSS methods.

Acknowledgments

The authors thank two anonymous referees and the associate editor for their helpful comments and suggestions which improved the presentation of the paper. This study is supported by the Scientific and Technological Research Council of Turkey (TUBITAK-COST Grant No. 115F300) under ISCH COST Action IS1304.

Busra Sevinc
The Graduate School of Natural and Applied Sciences, Dokuz Eylul University
Izmir, Turkey
busra.sevincc@gmail.com

Bekir Cetintav
Department of Statistics, Mehmet Akif Ersoy University
Burdur, Turkey
bekircetintav@mehmetakif.edu.tr

Melek Esemen
The Graduate School of Natural and Applied Sciences, Dokuz Eylul University
Izmir, Turkey
melek.esemen@gmail.com

Selma Gurler
Department of Statistics, Dokuz Eylul University
Izmir, Turkey
selma.erdogan@deu.edu.tr