Austrian Journal of Statistics
MMMMMM YYYY, Volume VV, Issue II, 1–xx.
http://www.ajs.or.at/

An efficient variant of Ranked Set Sampling, Probability Proportional to Size with Application to Economic Data

Saeid Amiri,
Montreal Neurological Institute, McGill, CANADA

Hossein Hassani
Institute for International Energy Studies, Tehran, Iran

Saeed Hervi
Cardiff Business School, Cardiff, Wales, UK

Abstract

In this paper, we apply the Ranked Set Sampling (RSS) technique to economic data in the form of a homescan market research data set for the meat food group. The RSS method is then extended to select sampling units based on the Probability-Proportional-to-Size (PPS) approach. The new proposed ranked set sampling, using the PPS-derived method, RPPS, is assessed via Monte Carlo investigations and an extensive homescan data set to evaluate its performances. The results are promising and in line with theoretical and simulation studies, showing that the RPPS technique is more reliable and has a smaller variance than the PPS route.

Keywords: Bootstrap method; Monte Carlo simulation; Probability proportional to size; Ranked set sample.
MSC[2020]: 62G30, 62F07, 62F40..

1. Introduction

Ranked Set Sampling, hereafter referred to as RSS, is a sampling approach whose basic structure could lead to improved statistical inference in a range of situations where the actual measurement of the variable of interest is difficult or expensive to obtain, while sampling units can be easily and cheaply ordered by certain means, including visual inspection, without actual quantification. In fact, it is an intriguing development in data collection techniques that enables one to gather a more informative sample than can be garnered through simple random sampling. RSS is a two-stage sampling technique where a number of sampling units are first ranked with respect to the variable of interest, and second, measurements are taken from a fraction of these ranked units. Rank-based sampling designs are powerful alternatives to simple random sampling (SRS), often offer notable improvements in precision, and have been used in diverse applications, including the applications for RSS designs in ecological and environmental studies (e.g., Al-Saleh and Zheng (2002) and Kvam (2003)), forestry (Halls and Dell (1966)) medical studies (Samawi and Al-Sagheer (2001) and Chen, Stasny, and Wolfe (2005)), and reliability (Mahdizadeh and Zamanzade (2018)),

among others. Such applications have attracted widespread attention; in this paper, we consider a possible application of RSS using prices data.

Heravi and Morgan (2014) evaluated various sampling methods for meat prices, stratifying by kind of meat and other attributes, such as brand, method of storage/preservation, and region, to estimate the Consumer Prices Index (CPI). This index is an important macroeconomic indicator that attempts to summarize the changes in the price of a 'typical' basket of goods and is widely used for formulating economic policy and indexing pensions and welfare benefits. Therefore, its accurate measurement is critical, and it is clearly of interest to know how various sampling schemes perform in the context of such a price index construction. The authors mentioned above suggested that Probability Proportional to Size sampling (PPS) is an accurate method to estimate CPI.

It is well-known that both RSS is superior to the SRS, and Heravi and Morgan (2014) show using PPS can get an accurate estimate of CPI, so it is of interest to study a hybrid of RSS and PPS. In this paper, we extend the RSS method, using PPS (denoted as RPPS) and evaluate this new proposed sampling technique's performance. In fact, RSS is not so much a sampling technique as a data management method; accordingly, the combination of RSS and sampling methods like PPS would be of great interest. In this study, the performance measures considered are the bias and standard deviation of the mean estimate. We considered PPS with replacement to keep the probability constant. PPS deals with finite populations. The inference under RSS with a finite population is considered for different designs, see Deshpande, Frey, and Ozturk (2006), Abbasi and Yousaf Shad (2022), Zamanzade and Mahdizadeh (2020), Abbasi and Shad (2021) and more recently PPS is considered in Ozturk (2019) and Ozturk (2020) that considered the stratified populations, here we focus deeply on RPPS and show its applications in macroeconomic data.

Next, we will provide an overview of the data structure of an RSS and present a summary of this RSS method. Section 3 discusses PPS, Section 4 explores the performance of RPPS relative to PPS, and Section 5 describes two numerical studies to explore the finite sampling properties of the proposed method. We also present an application of the proposed technique to the homescan data set. Section 6 provides some concluding remarks.

2. Foundation of Ranked Set Sampling

Perfect RSS has two stages. In the first stage, units are identified and ranked perfectly. In the second stage, measurements are taken from a fraction of the ranked elements. To obtain an RSS of size k, one should choose an SRS of k units, $\{Y_{11},\ldots,Y_{1k}\}$, rank them without measurement on the variable of interest $Y_{(11)} \leq \ldots \leq Y_{(1k)}$, and select the smallest one, i.e., $Y_{(11)}$. Next we select the second smallest on the second SRS sample of k units, $Y_{(22)}$. This procedure is then repeated until k observations have been collected. Let us consider a cycle of RSS sample and denote as $\mathbf{Y} = \{Y_{(1)},\ldots,Y_{(k)}\}$. $Y \sim F(.)$ and $\sigma^2 < \infty$, the estimate of the population mean μ and its variance are

$$\begin{aligned} \bar{Y}_{RSS} &= \sum_{i=1}^{k} \frac{Y_{(i)}}{k}, \\ V(\bar{Y}_{RSS}) &= \frac{\sigma^2}{k} - \sum_{i=1}^{k} \frac{(\mu_{(i)} - \mu)^2}{k^2}, \end{aligned}$$

where μ is the mean of population and $\mu_{(i)}$ denotes the mean of ith order statistic in an SRS of size k. Takahasi and Wakimoto (1968) consider the relative precision comparing RSS estimation of population mean to SRS and showed that relative precision is bounded by 1 and by $(k+1)/2$ for any distribution with finite variance, the upper bound is achieved when sampling from the uniform distribution.

Dealing with finite population and following Arnold, Balakrishnan, and Nagaraja (2008), the probability mass function can be expressed.

$$P_{(r)}(y) = P_{(r)}(Y \leq y) - P_{(r)}(Y < y) = F_{(r)}(y) - F_{(r)}(y^-), \tag{1}$$

where $P_{(r)}$ and $F_{(r)}$ are the probability mass function and the cumulative distribution function of a ranked statistic with rank r. This presentation helps us to work with the discrete distribution. Using this expression

$$P_{(r)}(y) = \sum_{j=r}^{k} \binom{k}{j} \Big((1-F(y))^{k-j} (F(y))^j - (F(y^-))^j (1-F(y^-))^{k-j} \Big).$$

It can be shown that

$$P(y) = \frac{1}{k} \sum_{r=1}^{k} P_{(r)}(y). \tag{2}$$

An easier way to show statement (2) holds is to use the following density function

$$P_{(r)}(y) = C(r,k) \int_{F(y^-)}^{F(y)} u^{r-1} (1-u)^{k-r} du,$$

where $C(r,k) = \frac{k!}{(r-1)!(k-r)!}$. The use of this formula here is straightforward, for details, see Arnold *et al.* (2008). We have

$$\begin{aligned} \sum_{r=1}^{k} P_{(r)}(y) &= \sum_{r=1}^{k} C(r,k) \int_{F(y^-)}^{F(y)} u^{r-1}(1-u)^{k-r} du = \int_{F(y^-)}^{F(y)} \sum_{r=1}^{k} C(r,k) u^{r-1}(1-u)^{k-r} du \\ &= \int_{F(y^-)}^{F(y)} k du = k(F(y) - F(y^-)) = kP(y). \end{aligned} \tag{3}$$

The process of ranking may not be error-free. Under such a scenario, the probability mass function of a ranked statistic with rank r is no longer $P_{(r)}(y)$ and hence is denoted as $P_{[r]}(y)$, see Chen *et al.* (2005). Let p_{sr} be the probability that the sth order statistics is judged to have rank r. For the same probability of judging, we have

$$P_{[r]}(x) = \frac{1}{k} \sum_{s=1}^{k} p_{sr} P_{(s)}(x). \tag{4}$$

where $\sum_{r=1}^{k} p_{sr} = 1$, and obviously

$$\frac{1}{k} \sum_{r=1}^{k} P_{[r]}(y) = \frac{1}{k} \sum_{r=1}^{k} \sum_{s=1}^{k} p_{sr} P_{(s)}(y) = \frac{1}{k} \sum_{s=1}^{k} (\sum_{r=1}^{k} p_{sr}) P_{(s)}(y) = P(y).$$

Note that (4) holds under the assumption that the value of the sth order statistic and the event that it receives judgment rank r are independent, see Presnell and Bohn (1999). The review of complete discussion of imperfect ranking and the tests of imperfect ranking can be found in Amiri, Modarres, and Zwanzig (2017), and references therein.

To obtain a total number of $n = km$ units, the whole procedure should be repeated m times. Let $Y_{(r)j}$ denote the measurement on the jth measured unit with rank r. This results in a RSS of size n from the underlying population written as

$$\{Y_{(r)j}; r = 1, \ldots, k, j = 1, \ldots, m\}.$$

It is worth mentioning that, in RSS designs, $\{Y_{(1)j},\ldots,Y_{(k)j}\}$ are independent order statistics (as they are obtained from independent sets) and each $Y_{(r)j}$ provides information about a different stratum of the population.

The data can be represented as

$$\begin{aligned}\mathcal{Y}_1 &= \{Y_{(1)1}, Y_{(1)2},\ldots, Y_{(1)m}\} \overset{i.i.d.}{\sim} F_{(1)}(y),\\ \mathcal{Y}_2 &= \{Y_{(2)1}, Y_{(2)2},\ldots, Y_{(2)m}\} \overset{i.i.d.}{\sim} F_{(2)}(y),\\ &\vdots\\ \mathcal{Y}_k &= \{Y_{(k)1}, Y_{(k)2},\ldots, Y_{(k)m}\} \overset{i.i.d.}{\sim} F_{(k)}(y),\end{aligned}$$

where $F_{(r)}(y)$ is the cdf of the rth order statistic. The RSS data with unequal m is referred to unbalanced RSS. Let us define the mean and variance of order statistics

$$\mu_{(r)} = \int y dF_{(r)}(y), \tag{5}$$

$$\sigma^2_{(r)} = \int (y-\mu_{(r)})^2 dF_{(r)}(y). \tag{6}$$

Their estimates are $\bar{y}_{(r)} = \widehat{\mu}_{(r)} = \frac{1}{m}\sum_{j=1}^{m} y_{(r)j}$ and $\widehat{\sigma}^2_{(r)} = \frac{1}{m-1}\sum_{j=1}^{m}(y_{(r)j}-\widehat{\mu}_{(r)})^2$. The following proposition provides an asymptotic test statistics for the sample mean obtained by the RSS sampling method.

Proposition 1. *Suppose $F_{(r)}(y)$ with $\int y^2 dF_{(r)}(y) < \infty$, $\widehat{F}_{(r)}(y)$ is the edf of the rth row and the parameter of interest be the mean of population, μ. If $\vartheta_i = (\widehat{\mu}_{(r)} - \mu_{(r)})$, then $(\vartheta_1,\ldots,\vartheta_k)$ converges in distribution to a multivariate normal distribution with mean vector zero and covariance matrix $diag(\sigma^2_{(1)}/m,\ldots,\sigma^2_{(k)}/m)$.*

This proposition suggests the following statistic

$$Z = \frac{\frac{1}{k}\sum_{r=1}^{k}\bar{y}_{(r)} - \mu}{S} \xrightarrow[d]{} N(0,1),$$

where

$$S^2 = \frac{1}{k^2}\sum_{r=1}^{k}\frac{\widehat{\sigma}^2_{(r)}}{m}.$$

See Amiri, Jafari Jozani, and Modarres (2014) for the bootstrap methods.

3. Probability proportional to size sampling

Probability proportional to size sampling is a method of sample selection in which the units are selected with probability appropriate to a given measure related to the characteristics under study. It is also known as unequal probability sampling. Here sampling with replacement is considered as explained in Cochran (1977), the PPS with replacement is proposed in Hansen and Hurwitz (1943) to estimate the population total as same as Cochran (1977). This assumption helps developing theory, and it does not violate the Consumer Prices Index (CPI) because the data is generated continuously and the probabilities stay constant. Here the population mean is of interest.

To draw inference, let us consider a finite population $\{y_1,\ldots,y_N\}$ where the probability corresponding to selecting unit j from this population is $\pi_j = P(\mathcal{Y} = y_j)$, i.e., the probability of unit j being sampled is π_j . Let us consider the following probability mass function for PPS:

$$P(\mathcal{Y}=y) = \sum_{j=1}^{N}\pi_j I(y_j = y) = \sum_{j=1}^{N} P(\mathcal{Y}=y_j)I(y_j=y),$$

where N is the size of population. Consider a sample, $\{\mathcal{Y}_1,\ldots,\mathcal{Y}_n\}$, of size n with replacement from a finite population $\{y_1,\ldots,y_N\}$. Then the estimate of mean is

$$\bar{\mathcal{Y}}_{pps} = \frac{1}{nN}\sum_{i=1}^{n}\frac{\mathcal{Y}_i}{\pi_i^*}, \tag{7}$$

where π_i^* denotes the selection probability of the ith sampled unit, which this estimate is an unbiased estimate with variance

$$V(\bar{\mathcal{Y}}_{pps}) = \frac{1}{nN}\sum_{i=1}^{N} N\pi_i\left(\frac{y_i}{N\pi_i}-\eta\right)^2, \tag{8}$$

where $\eta = \frac{1}{N}\sum_{i=1}^{N} y_i$. The equations (7) and (8) can easily be obtained using the technique given in Cochran (1977), pp. 253. Here a direct approach is used

$$E(\bar{\mathcal{Y}}_{pps}) = \frac{1}{nN}\sum_{i=1}^{n}E\left(\frac{\mathcal{Y}_i}{\pi_i^*}\right) = \frac{1}{N}E\left(\frac{\mathcal{Y}_i}{\pi_i^*}\right) = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\pi_i}\pi_i = \frac{1}{N}\sum_{i=1}^{N} y_i = \eta.$$

The variance is

$$\begin{aligned} V(\bar{\mathcal{Y}}_{pps}) &= \frac{1}{(nN)^2}\sum_{i=1}^{n}V\left(\frac{\mathcal{Y}_i}{\pi_i^*}\right) = \frac{1}{nN^2}V\left(\frac{\mathcal{Y}_i}{\pi_i^*}\right) = \frac{1}{nN^2}\left(\sum_{i=1}^{N}\left(\frac{y_i}{\pi_i}\right)^2\pi_i - \left(\sum_{i=1}^{N}\frac{y_i}{\pi_i}\pi_i\right)^2\right) \\ &= \frac{1}{nN^2}\left(\sum_{i=1}^{N}\frac{y_i^2}{\pi_i} - (N\eta)^2\right) = \frac{1}{nN}\sum_{i=1}^{N}N\pi_i\left(\frac{y_i}{N\pi_i}-\eta\right)^2. \end{aligned}$$

The estimate of the variance from a sample of size n is

$$\widehat{V}(\bar{\mathcal{Y}}_{pps}) = \frac{1}{n(n-1)N^2}\sum_{i=1}^{n}\left(\frac{y_i}{\pi_i} - n\bar{\mathcal{Y}}_{pps}\right)^2. \tag{9}$$

Using the Central Limit Theorem:

$$Z = \frac{\bar{Y}_{pps}-\mu}{\sqrt{\widehat{V}(\bar{Y}_{pps})}} \xrightarrow[d]{} N(0,1).$$

4. Using RSS to achieve PPS

To collect n observations by the combination of RSS and PPS, denoted as RPPS hereafter, let us first obtain k sampling units selected with PPS

$$\{\mathcal{Y}_1,\ldots,\mathcal{Y}_n\},$$

the unit with rank 1 is identified and taken for the measurement, $Y_{(1)1}$, and the remaining are disregarded. The procedure can be repeated for m times to have m iid units with rank 1. Next, another k units are drawn with PPS and the unit with rank 2 is measured, $Y_{(2)1}$. The procedure is continued until m units with rank k are collected. Using this procedure, $n = km$ observations are collected. Then, the sample is

$$\{Y_{(r)1}, Y_{(r)2},\ldots, Y_{(r)m}\},\ r = 1,\ldots,k. \tag{10}$$

The sample mean is estimated by

$$\bar{Y}_{rpps} = \frac{1}{kN}\sum_{r=1}^{k}\frac{1}{m}\sum_{j=1}^{m}\frac{Y_{(r)j}}{\pi_{rj}^*}. \tag{11}$$

where $\pi^*_{rj} = \sum_{k=1}^{N} \pi_k I(Y_{(r)j} = y_k)$, i.e., it is the probability corresponding to selecting unit j from this population that is appeared in cycle r, $\pi_{rj} = P(Y_{(r)j} = y_j)$. Using the definition, the expected value of the rth order statistic in a finite population is

$$\begin{aligned} \mu_{(r)} &= E(Y_{(r)}) = \sum_{j=1}^{N} y_j P_{(r)}(y_j), \\ E\left(\frac{Y_{(r)}}{\pi^*_r}\right) &= \sum_{j=1}^{N} \frac{y_j}{\pi_j} P_{(r)}(y_j). \end{aligned} \tag{12}$$

Obviously under $\pi_r = 1/N$, $N\mu_{(r)} = E\left(\frac{Y_{(r)}}{\pi^*_r}\right)$. Let us define

$$\eta_{(r)} = \frac{1}{N} E\left(\frac{Y_{(r)}}{\pi^*_r}\right). \tag{13}$$

Using the following proposition, we prove that the RPPS for $m = 1$ provides an unbiased estimate with a lower variance than that for standard balanced PPS.

Proposition 2. *Suppose the samples are selected with unequal probability where $\sum_{i=1}^{N} y_i^2 \pi_i < \infty$ from a finite population, and $\{Y_{(1)}, \ldots, Y_{(k)}\}$ are collected according to the proposed algorithm under a perfect ranking from this population, then*

$$\begin{aligned} E(\bar{Y}_{rpps}) &= E(\bar{Y}_{pps}) = \eta, \\ V(\bar{Y}_{rpps}) &\leq V(\bar{Y}_{pps}). \end{aligned}$$

Proof. Recall that the mean of these observations is

$$\bar{Y}_{rpps} = \frac{1}{kN} \sum_{r=1}^{k} \frac{Y_{(r)}}{\pi^*_r}.$$

Its expected value is then obtained as follows:

$$\begin{aligned} E(\bar{Y}_{rpps}) &= \frac{1}{kN} \sum_{r=1}^{k} E\left(\frac{Y_{(r)}}{\pi^*_r}\right) = \frac{1}{kN} \sum_{r=1}^{k} \left(\sum_{j=1}^{N} \frac{y_j}{\pi_j} P_{(r)}(y_j) \right) = \frac{1}{N} \sum_{j=1}^{N} \left(\frac{y_j}{\pi_j} \frac{1}{k} \sum_{r=1}^{k} P_{(r)}(y_j) \right) \\ &= \frac{1}{N} \sum_{j=1}^{N} \frac{y_j}{\pi_j} P(y_j) = \frac{1}{N} \sum_{j=1}^{N} y_j = \eta = E(\bar{Y}_{pps}). \end{aligned}$$

This statement shows that RPPS provides an unbiased estimate of the population mean, and it is held regardless the ranking procedure. The variance can be obtained using

$$\begin{aligned} V(\bar{Y}_{rpps}) &= \frac{1}{(kN)^2} \sum_{r=1}^{k} V\left(\frac{Y_{(r)}}{\pi^*_r}\right) = \frac{1}{(kN)^2} \sum_{r=1}^{k} \left(E\left(\frac{Y_{(r)}}{\pi^*_r} - N\eta_{(r)}\right)^2 \right) \\ &= \frac{1}{(kN)^2} \sum_{r=1}^{k} \left(E\left(\frac{Y_{(r)}}{\pi^*_r} - N\eta + N\eta - N\eta_{(r)}\right)^2 \right) \\ &= \frac{1}{(kN)^2} \left(\sum_{r=1}^{k} E\left(\frac{Y_{(r)}}{\pi^*_r} - N\eta\right)^2 - \sum_{r=1}^{k} (N\eta_{(r)} - N\eta)^2 \right), \end{aligned}$$

where

$$\begin{aligned} \frac{1}{(kN)^2} \sum_{r=1}^{k} E\left(\frac{Y_{(r)}}{\pi^*_r} - N\eta\right)^2 &= \frac{1}{(kN)^2} \sum_{r=1}^{k} \left(\sum_{j=1}^{N} N^2 \left(\frac{y_j}{N\pi_j} - \eta\right)^2 P_{(r)}(y_j) \right) \\ &= \frac{1}{kN} \sum_{j=1}^{N} \left(\frac{y_j}{N\pi_j} - \eta\right)^2 N P(y_j) = \frac{1}{kN} \sum_{j=1}^{N} N\pi_j \left(\frac{y_j}{N\pi_j} - \eta\right)^2, \end{aligned}$$

hence

$$V(\bar{Y}_{rpps}) = V(\bar{Y}_{pps}) - \frac{1}{(kN)^2}\left(\sum_{r=1}^{k}(N\eta_{(r)} - N\eta)^2\right).$$

It stablishes $V(\bar{Y}_{rpps}) \leq V(\bar{Y}_{pps})$, while $E(\bar{Y}_{rpps}) = E(\bar{Y}_{pps})$, regrdless the randking procedure. □

Proposition 2 can be used to generalize the result for RSS with $n = mk$. According to Propositions 2

$$\begin{aligned} V(\bar{Y}_{rpps}) &= \frac{1}{mk^2N^2}\left(\sum_{j=1}^{N} N^2 k\pi_j\left(\frac{y_j}{N\pi_j} - \eta\right)^2 - N^2\sum_{r=1}^{k}(\eta_{(r)} - \eta)^2\right) \\ &= \frac{1}{mk^2N}\left(\sum_{j=1}^{N} k\left(\frac{y_j^2}{N\pi_j} - \eta^2\right) - N\sum_{r=1}^{k}(\eta_{(r)} - \eta)^2\right). \end{aligned} \tag{14}$$

The estimate of the first part is given in (9), but the unbiased estimate of the second part for $m = 1$ is not trivial, see Zamanzade and Vock (2015) for the discussion of variance. A practical way to accomplish this is using the bootstrap method, the bootstrap of RSS is discussed in Amiri *et al.* (2014). The bootstrap method is a standard tool in statistical analysis that can be used to perform the statistical inference. In this study, the non-parametric bootstrap is used to approximate the population distribution function. The bootstrap can be used to obtain the sampling distribution of a statistic of interest. The bootstrap allows for estimation of the standard error of any well-defined statistic and enables us to draw inferences when the exact or the asymptotic distribution of the statistic of interest is unavailable. To estimate the variance of $\bar{y}_{rpps}$, the bootstrap method can be used, see Algorithm 1.

1. Define

$$\begin{aligned}
\mathcal{Z}_1 &= \{z_{(1)1} = y_{(1)1}/\pi^*_{(1)1}, z_{(1)2} = y_{(1)2}/\pi^*_{(1)2}, ..., z_{(1)m} = y_{(1)m}/\pi^*_{(1)m}\}.\\
\mathcal{Z}_2 &= \{z_{(2)1} = y_{(2)1}/\pi^*_{(2)1}, z_{(2)2} = y_{(2)2}/\pi^*_{(2)2}, ..., z_{(2)m} = y_{(2)m}/\pi^*_{(2)m}\}.\\
&\dots\\
\mathcal{Z}_k &= \{z_{(k)1} = y_{(k)1}/\pi^*_{(k)1}, z_{(k)2} = y_{(k)2}/\pi^*_{(k)1}, ..., z_{(k)m} = y_{(k)m}/\pi^*_{(k)m}\}.
\end{aligned}$$

2. Combine all the observations to form $\mathcal{Z}^\diamond = \{\mathcal{Z}_1, \dots, \mathcal{Z}_k\}$ and assign the probability of $1/km$ to each element of $\mathcal{Z}^\diamond$.

3. Randomly draw $\{Z_1, \dots, Z_k\}$ from $\mathcal{Z}^\diamond$, order them as $Z_{(1)} \le \dots \le Z_{(k)}$ and retain $Z^*_{(r)1} = Z_{(r)}$.

4. Perform Step 3 for $r = 1, \dots, k$.

5. Repeat Steps 3–4, m times to obtain $\left\{Z^{*\diamond}_{(r)j},\ j = 1, \dots, m\right\}$ for r=1, . . . ,k.

6. Calculate

$$\bar{Z}^* = \frac{1}{kmN} \sum_{r=1}^{k} \sum_{j=1}^{m} Z^{*\diamond}_{(r)j}.$$

7. Repeat Steps 3–6 B times to obtain the bootstrap samples

$$\bar{Z}^*_b,\ b = 1, \dots, B,$$

and estimate the variance using

$$S^{2*}_Z = \frac{1}{B} \sum_{b=1}^{B} (\bar{Z}^*_b - \bar{\bar{Z}})^2,$$

where $\bar{\bar{Z}}$ is the average of $\bar{Z}^*_b$.

Algorithm 1: Estimate the variance via the Bootstrap method

5. Numerical evaluation

In this section, we first study the performance of the proposed method for estimating the population mean of proposed designs. We then apply our method to a real data set where we also study the performance of our proposed ranked- based technique.

5.1. Simulation

Monte Carlo simulations were used in order to investigate the finite sample properties of the proposed RSS algorithm. We examine certain desirable features such as unbiasedness and smaller variance. Here different balanced RSS with $k = 5$ and different sizes are used to study the performance of discussed methods

$$\begin{aligned}
D_1 &= (1,1,1,1,1), \quad n_1 = 5,\\
D_2 &= (2,2,2,2,2), \quad n_2 = 10,\\
D_3 &= (3,3,3,3,3), \quad n_3 = 15,\\
D_4 &= (4,4,4,4,4), \quad n_4 = 20,\\
D_5 &= (5,5,5,5,5), \quad n_5 = 25.
\end{aligned}$$

The design $D_i = (i,i,i,i,i)$ shows RSS data where each order statistic is gathered i times. Let us consider an artificial finite population,

$$\mathcal{P} = \{y_1, \cdots, y_{100}\} = \{1, \ldots, 100\},$$

which has $N = 100$ and $\frac{1}{100}\sum_{i=1}^{100} y_i = 50.5$. Four different probabilities, $\pi_j = (\pi_{1j}, \ldots, \pi_{100j}), j \in \{I, II, III, IV\}$ are considered, see Table 1. Clearly the values of artificial populations receive different weights to study the proposed methods numerically; π_I gives the largest weighs to $\{Y_{76}, \ldots, Y_{100}\}$, π_{IV} gives the largest weighs to $\{Y_1, \ldots, Y_{25}\}$, π_{II} and π_{III} give the smallest weights to $\{Y_{26}, \ldots, Y_{50}\}$ and $\{Y_{51}, \ldots, Y_{75}\}$, respectively.

Table 1: The proposed probabilities for $N = 100$

	Probabilities			
data	π_I	π_{II}	π_{III}	π_{IV}
y_1 =1	0.004	0.008	0.008	0.016
⋮	⋮	⋮	⋮	⋮
y_{25} =25	0.004	0.008	0.008	0.016
y_{26} =26	0.008	0.004	0.012	0.004
⋮	⋮	⋮	⋮	⋮
y_{50} =50	0.008	0.004	0.012	0.004
y_{51} =51	0.012	0.012	0.004	0.012
⋮	⋮	⋮	⋮	⋮
y_{75} =75	0.012	0.012	0.004	0.012
y_{76} = 76	0.016	0.016	0.016	0.008
⋮	⋮	⋮	⋮	⋮
y_{100} =100	0.016	0.016	0.016	0.008

To study the estimation of mean using the proposed methods, a PPS sample with the size of n_i $i = 1, \ldots, 5$ is collected from $\mathcal{P}$ and the sample mean corresponding to (7) is calculated. To study its competitor, RPPS, a sample via the discussed procedure with size n_i and the ith design is collected from $\mathcal{P}$ and the sample mean is calculated via (11), the whole procedure is repeated 10,000 times and the mean and variance (number given in the parentheses) are given in Table (2). It shows the estimate of mean using the RPPS has lower variance for different designs and probabilities, which is expected from the theory provided in Section 4.

Studying the behavior of the proposed methods under imperfect ranking is important because when the ranking process is not perfect, there is often a loss of efficiency. The statistical tests of perfectness of rankings have received attention in RSS literature, see Vock and Balakrishnan (2011) and Amiri *et al.* (2017) among others. Several mechanisms are presented to produce imperfect RSS samples, see Amiri *et al.* (2017) and references therein. We use the Fraction of Neighbors technique to generate the imperfect ranking in data; let us denote the ranks using imperfect ranking by $[.]$, we assume $F_{[i]}$ is a mixture of $F_{(i)}$'s. That is,

$$F_{[i]}(x) = (1-\lambda)F_{(i)}(x) + \frac{\lambda}{2}F_{(i-1)}(x) + \frac{\lambda}{2}F_{(i+1)}(x),$$

where λ is the fraction of incorrectly chosen statistics. Here, $\lambda = \frac{1}{3}$ is used and for the extreme judgment order statistics $F_{(0)} := F_{(1)}$ and $F_{(k+1)} := F_{(k)}$. Perfect rankings are obtained by setting $\lambda = 0$. Table 3 includes the estimate of the mean and variance under RSS with imperfect ranking, comparing Table 2 and 3 show that RSS procedure when applied to samples collected using PPS give rise to improved precision than using just PPS. The other method of generating imperfect RSS, following Dell and Clutter (1972), let $X_{[i]j}$ and $X_{(i)j}$ denote the judgment and true order statistics, respectively. Suppose $X_{[i]j} = X_{(i)j} + \epsilon ij$, where $X_{[i]j}$ and ϵij are independent and consider imperfect RSS designs with $\sigma_\epsilon = 5$. Our study shows the

Table 2: Simulation of mean and variance for the proposed approach under perfect ranking.

		RSS design				
Probability	Methods	D_1	D_2	D_3	D_4	D_5
π_I	PPS	50.479(22.876)	50.510(10.941)	50.480(7.364)	50.493(5.590)	50.522(4.459)
	RPPS	50.500(14.986)	50.489(7.613)	50.539(4.932)	50.528(3.763)	50.483(2.959)
π_{II}	PPS	50.504(100.165)	50.409(50.730)	50.461(34.363)	50.441(25.077)	50.479(20.104)
	RPPS	50.362(83.758)	50.501(42.437)	50.557(28.654)	50.426(21.189)	50.522(17.175)
π_{III}	PPS	50.438(311.774)	50.512(153.916)	50.432(104.563)	50.421(78.317)	50.408(61.933)
	RPPS	50.610(259.162)	50.516(129.502)	50.351(85.347)	50.389(64.766)	50.640(51.543)
π_{IV}	PPS	50.602(335.232)	50.223(171.193)	50.489(113.076)	50.573(84.578)	50.684(68.473)
	RPPS	50.432(168.981)	50.727(86.134)	50.523(56.526)	50.455(42.871)	50.550(33.375)

Table 3: Simulation of mean and variance for the proposed approach under imperfect ranking

		RSS design				
Probability	Methods	D_1	D_2	D_3	D_4	D_5
π_I	RPPS	50.674(17.011)	50.628(8.434)	50.726(5.196)	50.717(4.040)	50.706(3.215)
π_{II}	RPPS	51.136(92.866)	51.086(46.733)	51.137(30.886)	51.092(23.316)	51.161(17.947)
π_{III}	RPPS	51.523(307.829)	51.424(146.457)	51.279(98.006)	51.270(72.347)	51.328(58.383)
π_{IV}	RPPS	50.037(214.938)	50.273(110.395)	50.323(72.200)	50.230(54.003)	50.197(44.303)

Table 4: Simulation of mean and variance for the proposed approach under imperfect ranking, $\sigma = 25$

		RSS design				
Probability	Methods	D_1	D_2	D_3	D_4	D_5
π_I	RPPS	50.461(21.863)	50.509(10.629)	50.556(7.066)	50.450(5.333)	50.478(4.355)
π_{II}	RPPS	50.484(91.297)	50.538(45.392)	50.509(30.157)	50.569(22.874)	50.470(18.320)
π_{III}	RPPS	50.181(250.057)	50.487(131.924)	50.581(89.205)	50.649(66.522)	50.388(50.410)
π_{IV}	RPPS	50.681(176.665)	50.272(89.858)	50.516(59.466)	50.533(44.217)	50.564(35.483)

proposed method is robust with respect to imperfect ranking and still superior to it counterpart.

5.2. Experience with real data

In this section, we conduct a comparison of the methods in terms of their applications to real data. To this end, we consider the data set supplied by Taylor Nelson Softres (TNS, now part of the Kantar World Panel)[1], which contains 60 million transactions, from a sample panel of 35,000 households, for about 400,000 products. The households were chosen so that they would cover all ages, genders, and social classes and represent every region of the UK. Householders were required to scan their shopping purchases within their own homes. The main data set contains the details of the transactions, including the bar codes, household numbers, product codes, shop codes, product descriptions, market categories, year/month/week/day of transaction, expenditure and the number of packs bought. For example, for the meat data, there are around eighteen product attributes. To explore the theoretical part using the real data, the subset of these data which lists meat sold in London in December of 2005 is considered that includes 5,553 observations. For the purposes of study, the meat's attributes are categorized to the Frozen meat, Cooked Ham, Total Fresh Foods, prepackaged Fresh (meat,veg,pastry), non organic, pork and so on. The summary statistics of price (per pack) are given in Table 5. Table 5 and the histogram in Figure 1 shows the data is positively skewed, in our discussion, we never used any assumption about the population distribution, and the below anaysis shows the PRSS works well under non-normal population.

We used this data as a population, where $\frac{1}{N}\sum_{i=1}^{N} y_i = 2.3909$. To achieve PPS sampling, we first considered equal probability, i.e., $\pi_{V,i} = \frac{1}{\sum_k f_k} = \frac{1}{N}$ where $f_k, k = 1,\ldots,7$ is the frequency for each category and the summation of all $\pi_{V,i}$ equals one. To explore the proposed methods under unequal probabilities, we consider: $\pi_{VI,i} = \frac{f_j}{\sum_k f_k^2}$ and $\pi_{VII,i} = \frac{f_j^2}{\sum_k f_k^3}$ where the item i belongs to category j. The motivation behind such choices is logical, because the probability of elements is in terms of frequency. Table 6 shows the frequencies and probabilities assigned to observations. The elements and the proposed probabilities $\pi_V, \pi_{VI}, \pi_{VII}$ to generate the observations are denoted as V, VI, VII. We also consider situation when the inclusion probabilities are inversely proportional to the the group frequencies (suggested by a referee); $\pi_{VIII,i} \propto \frac{1}{f_j}$ when unit i belongs to category j. The estimate of mean (variance) under the perfect ranking is given in Table 7. Obviously, for the given probabilities, the RPPS leads to an unbiased estimate of mean and a smaller variance. In addition, to attain a better sample with lower variance, RSS also has the advantage of reducing the cost of data collection when, for example, sample collection is time-consuming and expensive, while the ranking variable is cheap. To sample prices for the actual CPI, the price collectors physically call into the shops, which is inherently expensive. However, if we were to use the last period prices, then we would obtain a variable that is highly correlated with the current prices and can be used as a useful ranking variable. Therefore, in this example, we used the total shopping expenditure as such variable. The estimate of mean (variance) under the imperfect ranking is given in Table 7. Here to generate the imperfect ranking, we consider a concomitant variable, the total shopping expenditure. The result shows that RPPS leads to a better estimate than PPS. However, as expected, comparing the variance of RPPS under imperfect and perfect ranking reveals the variance increases under imperfect ranking.

[1] Office for National Statistics (ONS) was provided the real data

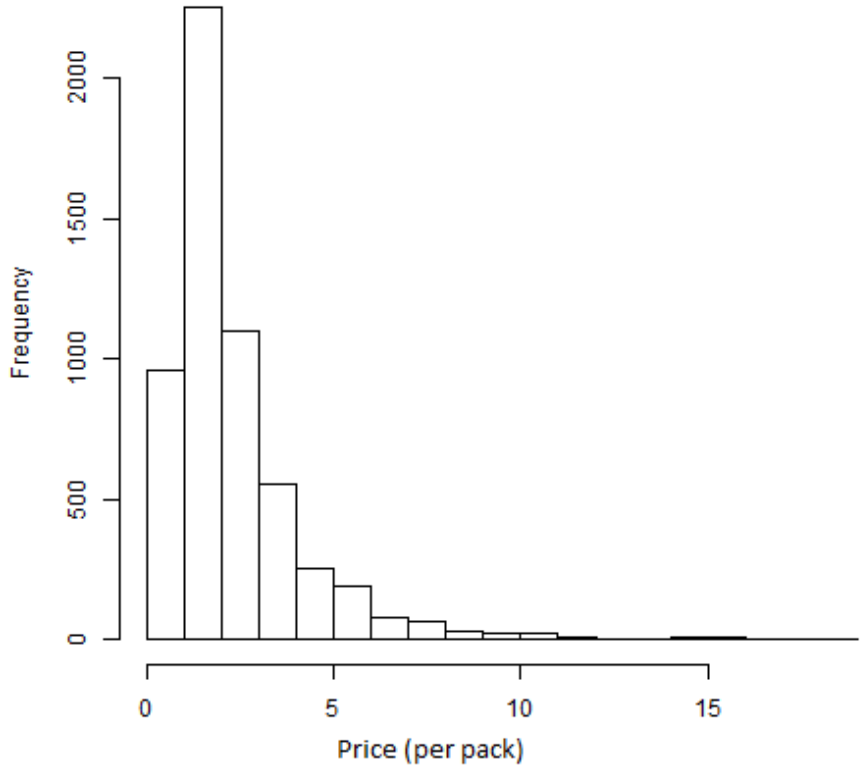

Figure 1: The histogram of the price of meat (per pack)

Table 5: Summary statistics for the price of meat (per pack)

Min.	1st quantile	Median	Mean	3rd quantile	Max.
0.1299	1.2900	1.8182	2.3909	2.9318	18.1169

Table 6: The frequency of meat's categories and the assigned probabilities

		Probabilities			
Category	f_i	π_V	π_{VI}	π_{VII}	π_{VIII}
Frozen Meat	990	0.0001800828	0.0002046017	0.0002200813	0.0001443001
Cooked Meats Ham	1002	0.0001800828	0.0002070817	0.0002254489	0.000142572
Total Fresh Foods	773	0.0001800828	0.0001597547	0.0001341750	0.0001848087
Prepackaged Fresh*	741	0.0001800828	0.0001531413	0.0001232960	0.0001927897
Non Organic	643	0.0001800828	0.0001328878	0.0000928399	0.0002221729
Pork	309	0.0001800828	0.0000638605	0.0000214402	0.0004623209
Others	1095	0.0001800828	0.0002263019	0.0002692409	0.0001304631

* :Meat,Veg, and Pastry

Table 7: Study of mean and variance for the proposed approach on the real data.

		Perfect Ranking				
		RSS design				
	Methods	D_1	D_2	D_3	D_4	D_5
V	PPS	2.394(0.654)	2.395(0.322)	2.399(0.218)	2.387(0.1572)	2.395(0.128)
	RPPS	2.398(0.350)	2.395(0.122)	2.394(0.089)	2.387(0.0700)	2.391(0.059)
VI	PPS	2.399(0.683)	2.386(0.340)	2.397(0.220)	2.383(0.161)	2.394(0.131)
	RPPS	2.393(0.344)	2.393(0.118)	2.392(0.086)	2.385(0.068)	2.387(0.057)
VII	PPS	2.407(0.984)	2.393(0.482)	2.390(0.324)	2.383(0.237)	2.380(0.189)
	RPPS	2.390(0.679)	2.392 (0.233)	2.391(0.174)	2.388(0.137)	2.389(0.114)
$VIII$	PPS	2.389(1.008)	2.403(0.513)	2.397(0.336)	2.384(0.251)	2.386(0.198)
	RPPS	2.383(0.619)	2.393(0.218)	2.388(0.155)	2.392(0.128)	2.389(0.102)
		Imperfect Ranking				
		RSS design				
	Methods	D_1	D_2	D_3	D_4	D_5
V	RPPS	2.345(0.418)	2.346(0.140)	2.346(0.100)	2.344(0.084)	2.348(0.070)
VI	RPPS	2.347(0.425)	2.356(0.144)	2.349(0.104)	2.353(0.082)	2.356(0.070)
VII	RPPS	2.365(0.770)	2.362(0.254)	2.362(0.188)	2.363(0.152)	2.358(0.125)
$VIII$	RPPS	2.322(0.714)	2.340(0.240)	2.333(0.185)	2.326(0.137)	2.322(0.118)

6. Conclusions

A considerable amount of research has been done to elaborate RSS. Ranked-based sampling techniques are designed to use additional information from inexpensive and easily obtained sources to collect a more representative sample than can be gained from simple random sampling. Due to the unique structure of RSS, researchers are able to have an estimate with lower variabilities, which helps us to draw better inference.

This paper defined a ranked sampling procedure for PPS sampling; we explored the RSS and PPS approaches and considered the possibility of achieving the latter using the former (denoted by RPPS). The properties of these sampling methods were studied theoretically and proved that RPPS outperformed PPS, giving an unbiased estimate with lower variance. The Monte Carlo simulations under perfect/imperfect ranking designs also confirmed the theoretical results obtained. Our findings showed that RPPS is always superior to PPS, with significantly lower variance. In fact, it shows a reduction of up to 50% in the variance for some cases. Taking the TNS database as the population of interest, we also examined the two sampling methods with real data, which were fairly skewed. The results indicated that RPPS provides an unbiased estimate with lower variance and can be considered an efficient sampling technique.

There are many other methods that could be used; for instance, we considered a finite population with sampling with replacement. However, this method can be extended to sample without replacement and consider unbalanced RSS with missing data, but we leave them for future research.

Acknowledgments

The authors would like to thank the Office for National Statistics (ONS) for providing the real data. The opinions expressed here are ours and, of course, not those of the ONS. Data supplied by TNS UK Limited. The use of TNS UK Ltd data in this work does not imply the endorsement of TNS UK Ltd. in relation to the interpretation or analysis of the data. All errors and omissions remain the responsibility of the authorss.
We gratefully acknowledge the constructive comments and suggestions of the anonymous referees, and the associate editor.

References

Abbasi AM, Shad MY (2021). "Sensitive proportion in ranked set sampling." *PloS one*, **16**(8), e0256699. doi:10.1371/journal.pone.0256699.

Abbasi AM, Yousaf Shad M (2022). "Estimation of population proportion using concomitant based ranked set sampling." *Communications in Statistics-Theory and Methods*, **51**(9), 2689–2709. doi:10.1080/03610926.2021.1916529.

Al-Saleh MF, Zheng G (2002). "Theory & Methods: Estimation of bivariate characteristics using ranked set sampling." *Australian & New Zealand Journal of Statistics*, **44**(2), 221–232. doi:10.1111/1467-842X.00224.

Amiri S, Jafari Jozani M, Modarres R (2014). "Resampling unbalanced ranked set samples with applications in testing hypothesis about the population mean." *Journal of agricultural, biological, and environmental statistics*, **19**(1), 1–17. doi:10.1007/s13253-013-0153-y.

Amiri S, Modarres R, Zwanzig S (2017). "Tests of perfect judgment ranking using pseudo-samples." *Computational Statistics*, **32**(4), 1309–1322. doi:10.1007/s00180-016-0698-7.

Arnold BC, Balakrishnan N, Nagaraja HN (2008). *A first course in order statistics*. SIAM.

Chen H, Stasny EA, Wolfe DA (2005). "Ranked set sampling for efficient estimation of a population proportion." *Statistics in medicine*, **24**(21), 3319–3329.

Cochran WG (1977). *Sampling techniques*. John Wiley & Sons.

Dell T, Clutter J (1972). "Ranked set sampling theory with order statistics background." *Biometrics*, pp. 545–555. doi:10.2307/2556166.

Deshpande JV, Frey J, Ozturk O (2006). "Nonparametric ranked-set sampling confidence intervals for quantiles of a finite population." *Environmental and Ecological Statistics*, **13**(1), 25–40. doi:10.1007/s10651-005-5688-9.

Halls LK, Dell TR (1966). "Trial of ranked-set sampling for forage yields." *Forest Science*, **12**(1), 22–26. doi:10.1093/forestscience/12.1.22.

Hansen MH, Hurwitz WN (1943). "On the theory of sampling from finite populations." *The Annals of Mathematical Statistics*, **14**(4), 333–362.

Heravi S, Morgan P (2014). "Sampling schemes for price index construction: a performance comparison across the classification of individual consumption by purpose food groups." *Journal of Applied Statistics*, **41**(7), 1453–1470. doi:10.1080/02664763.2014.881466.

Kvam PH (2003). "Ranked set sampling based on binary water quality data with covariates." *Journal of Agricultural, Biological, and Environmental Statistics*, **8**(3), 271–279. doi:10.1198/1085711032156.

Mahdizadeh M, Zamanzade E (2018). "A new reliability measure in ranked set sampling." *Statistical Papers*, **59**(3), 861–891. doi:10.1007/s00362-016-0794-3.

Ozturk O (2019). "Post-stratified probability-proportional-to-size sampling from stratified populations." *Journal of Agricultural, Biological and Environmental Statistics*, **24**(4), 693–718. doi:10.1007/s13253-019-00370-6.

Ozturk O (2020). "Probability-proportional-to-size ranked-set sampling from stratified populations." *Survey Methodology*, **46**(2), 243–265.

Presnell B, Bohn LL (1999). "U-statistics and imperfect ranking in ranked set sampling." *Journal of Nonparametric Statistics*, **10**(2), 111–126. doi:10.1080/10485259908832756.

Samawi HM, Al-Sagheer OA (2001). "On the estimation of the distribution function using extreme and median ranked set sampling." *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, **43**(3), 357–373. doi:10.1002/1521-4036(200106)43:3<357::AID-BIMJ357>3.0.CO;2-Q.

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.

Vock M, Balakrishnan N (2011). "A Jonckheere–Terpstra-type test for perfect ranking in balanced ranked set sampling." *Journal of Statistical Planning and Inference*, **141**(2), 624–630. doi:10.1016/j.jspi.2010.07.005.

Zamanzade E, Mahdizadeh M (2020). "Using ranked set sampling with extreme ranks in estimating the population proportion." *Statistical methods in medical research*, **29**(1), 165–177. doi:10.1177/0962280218823793.

Zamanzade E, Vock M (2015). "Variance estimation in ranked set sampling using a concomitant variable." *Statistics & Probability Letters*, **105**, 1–5. doi:10.1016/j.spl.2015.04.034.

Affiliation:

Saeid Amiri
Department of Neurology and Neurosurgery, Montreal Neurological Institute, McGill, CANADA
E-mail: saeid.amiri@gmail.com
Hossein Hassani
Institute for International Energy Studies, Tehran, Iran
Saeed Heravi
Cardiff Business School, Cardiff, Wales, UK

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume VV, Issue II
doi:10.17713/ajs.v000.i00

http://www.ajs.or.at/
http://www.osg.or.at/
Submitted: yyyy-mm-dd
Accepted: yyyy-mm-dd