Ranked Set Sampling with Probability Proportional to Size with Application to Economic Data

Saeid Amiri[a1] Hossein Hassani[b] Saeed Heravi[c] Peter Morgan[c]

[a,] *University of Wisconsin-Green Bay, Department of Natural and Applied Sciences, Green Bay, WI, USA*

[b] *Institute for International Energy Studies, Tehran, Iran*

[c] *Cardiff Business School, Cardiff, Wales, UK*

Abstract

The design of rank-based sampling provides a powerful inference alternatives to simple random sampling (SRS) and frequently leads to significant improvements in the precision of estimators. In this paper, we apply the Ranked Set Sampling (RSS) technique to economic data in the form of home scan market research data set for the meat food group. The RSS method is then extended to select a sampling unit 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 home scan 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 smaller variance than the PPS route.

Keywords: Bootstrap method; Monte Carlo simulation; Probability proportional to size; Ranked set sample.

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, where 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

[1]Corresponding author: saeid.amiri1@gmail.com

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 et al. (2005)), and reliability (Mahdizadeh and Zamanzade (2016)), among others. Such applications have attracted widespread attentions; 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 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 above mentioned authors suggested the Probability Proportional to Size sampling (PPS) as an accurate method with which to estimate CPI. It is well known that both RSS and PPS are both superior to the SRS. In this paper, we extend the RSS method, using PPS and evaluate the performance of this new proposed sampling technique. In fact, RSS is not so much as a sampling technique as a data management method; accordingly, the combination of RSS and sampling method like PPS would be of great interest. In this study, the measures of performance considered are the bias and standard deviation of estimate of mean. We considered PPS with replacement to keep the probability constant. PPS deals with finite population, hence it is utmost important to discuss the finite population for proposed RSS, the inference under the RSS finite are considered for different designs, see Deshpande et al. (2006) and Ozturk (2016). A Monte Carlo simulation study and repeated bootstrap sampling were used to estimate the bias and the variances and to evaluate the performance of the two sampling schemes, PPS and RPPS. In addition, the proposed method is applied to the home scan market research data set used by Heravi and Morgan (2014) to evaluate the proposed methods using real data.

Next, we will provide an overview of the data structure of an RSS and presents 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 home scan data set. Section 6 provides some concluding remarks.

2 Ranked Set Sampling

RSS has two stages, in the first stage, units are identified and ranked. In the second stage, measurements are taken from a fraction of the ranked elements. To obtain a RSS of size k, one should choose a 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 denote the RSS sample as $\mathcal{Y} = \{Y_{(1)} \leq \ldots \leq Y_{(k)}\}$. Under $Y \sim F(.)$ 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 $\mu_{(i)}$ 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 case and following Arnold et al. (2008), the pmf can be expressed .

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

this presentation helps us to work with the discrete distribution. Using this expression

$$P_{(r)}(y) = \sum_{j=r}^{k} \binom{k}{j} \Big((1-F(y))^{k-j} (F(y))^{j} - (F(y-1))^{j} (1-F(y-1))^{k-j} \Big). \tag{2}$$

It can be shown that

$$P(y) = \frac{1}{k} \sum_{r=1}^{k} P_{(r)}(y). \tag{3}$$

The process of ranking can be with error. Under such a scenario, the pmf 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. (2004). Let p_{sr} be the probability that the sth order statistics judged to have rank r. For the same probability of judging, we have

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

Obviously

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

The review of complete discussion of imperfect ranking and the tests of imperfect ranking can be found in Amiri et al. (2016), 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.

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 the unequal probability sampling. Here sampling with replacement is of concern as explained in Cochran (1977). The idea of PPS with replacement proposed in Hansen and Hurwitz (1943) to estimate the population total as same as Cochran (1977). Here the population mean is of concern.

To draw inference, let $\pi_j = P(Y = y_j)$ be the probability corresponding to selecting unit i in the population. Let us consider the following probability mass function for PPS:

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

where N is the size of population. Consider a sample of size n with replacement. Then the estimate of mean is

$$\bar{Y}_{pps} = \frac{1}{nN} \sum_{i=1}^{n} \frac{Y_i}{\pi_i}, \tag{5}$$

which is an unbiased estimate with variance

$$V(\bar{y}_{pps}) = \frac{1}{nN} \sum_{i=1}^{N} N\pi_i (\frac{y_i}{N\pi_i} - \eta)^2, \tag{6}$$

where $\eta = \frac{1}{N} \sum_{i=1}^{N} y_i$. The equations (5) and (6) can easily be obtained using the technique given in Cochran (1977), pp. 253. Here a direct approach is used

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

The variance is

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

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

$$\widehat{V}(\bar{Y}_{pps}) = \frac{1}{n(n-1)N^2} \sum_{i=1}^{n} \left(\frac{y_i}{\pi_i} - n\bar{Y}_{pps} \right)^2. \tag{7}$$

Using Central Limit Theorem can show this proposition suggests the following statistic

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

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

$$\{Y_1, \ldots, Y_k\},$$

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

$$\{Y_{(r)1}, Y_{(r)2}, \ldots, Y_{(r)m}\},\ r = 1, \ldots, k. \tag{8}$$

The sample mean is estimated by

$$\bar{Y}_{rpps} = \frac{1}{kN} \sum_{r=1}^{k} \frac{1}{m} \sum_{j=1}^{m} \frac{Y_{(r)j}}{\pi_{rj}}. \tag{9}$$

A kind of probability proportion is considered in Muttlak and McDonald (1990). In order to draw the inference, they considered the following probability density function pdf

$$g(y) = \frac{yf(y)}{\mu},$$

where $g(y)$ is called weighted pdf and μ is the mean of the unweighed density

of $f(y)$. This density is suggested by Cox (1969) where each unit has probability of selection proportional to its size Y. Actually they considered the infinite pdf, whereas we confine our research work on the finite population according to the sampling technique given in the second section. Using the definition

$$\begin{aligned} \mu_{(r)} &= E(Y_{(r)}) = \sum_{j=1}^{N} y_j P_{(r)}(y_j), \\ E\left(\frac{Y_{(r)}}{\pi_r}\right) &= \sum_{j=1}^{N} y_j P_{(r)}(y_j). \end{aligned} \tag{10}$$

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

$$\eta_{(r)} = \frac{1}{N} E\left(\frac{Y_{(r)}}{\pi_r}\right). \tag{11}$$

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 1. *Suppose $P(y)$ with $\sum_{i=1}^{N} y_i^2 P(y_i) < \infty$ and $\{Y_{(1)}, \ldots, Y_{(k)}\}$ are collected according to the proposed algorithm under perfect ranking, 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 the mean of these observations by

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

Its expected value is then obtained as follows:

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

This statement shows that RPPS provides an unbiased estimate of the population mean. The variance can be obtained using

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

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

$$V(\bar{Y}_{rpps}) \leq V(\bar{Y}_{pps}),$$

while

$$E(\bar{Y}_{rpps}) = E(\bar{Y}_{pps}).$$

□

Proposition 1 can be used to generalize the result for RSS with $n = mk$. According to Propositions 1

$$\begin{aligned} V(\bar{Y}_{rpps}) &= \frac{1}{mk^2N}\left(\sum_{j=1}^{N} k\pi_j\left(\frac{y_j}{N\pi_j} - \eta\right)^2 - N\sum_{r=1}^{k}(\eta_{(r)} - \eta)^2\right) \\ &= \frac{1}{mk^2N}\left(\sum_{j=1}^{N}\left(\frac{ky_j^2}{N^2\pi_j} - k\eta^2\right) - N\sum_{r=1}^{k}(\eta_{(r)} - \eta)^2\right). \end{aligned} \tag{12}$$

The estimate of the first part is given in (7), 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). As a result,

we considered the bootstrap method, a standard tool in statistical analysis that can be used to achieve the statistical inference. In this work, the nonparametric bootstrap is considered that is the empirical distribution function (edf) that serves as a good approximation to 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 $\bar{y}_{rpps}$, we use the bootstrap method, see Algorithm 1.

5 Evaluation of the proposed method

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 Numerical study

This section demonstrates the validity of the proposed algorithm. To study the finite sample properties of algorithms, Monte Carlo experiments are used and the proposed RSS algorithms are explored simultaneously based on the same designs so as to provide a meaningful way to compare the various methods. In order to enable a comparative evaluation of testing procedures we seek 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

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

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 $Z_{(1)} \leq \dots \leq Z_{(k)}$ and retain $Z^*_{(r)1} = Z_{(r)}$.

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

5. Repeat Steps 2–4, m times to obtain $\left\{Z^{*\diamond}_{(r)j},\ j = 1, \dots, m\right\}$.

6. Calculate

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

7. Repeat all steps 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

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 $\sum_{i=1}^{100} y_i = 50.5$. 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.

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 sample with the size of n_i $i = 1, \ldots, 5$ is collected from $\mathcal{P}$ and the sample mean corresponding to (5) is estimated. 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 mean is estimated via (9), 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.

Study the behavior of the proposed methods under imperfect ranking is very important because when ranking process are not perfect, there is often a loss of efficiency. Several mechanisms are presented to produce imperfect RSS

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

Probability	Methods	RSS design				
		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

Probability	Methods	RSS design				
		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)

samples, and also the statistical tests of perfectness of rankings have received attention in RSS literature, see Frey et al. (2007), Vock and Balakrishnan (2011), and Amiri et al. (2016) and references therein. We use the fraction of neighbors technique to generate the perfect 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. Note that RPPS may not always result in improvement over SRS, as PPS does not always give rise to an improvement in precision over SRS.

5.2 Experience with real data

In this part, 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), 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, the price of the goods, and the number of each items. For example, for the meat data, there are around eighteen product attributes. To explore the theoretical part using the real data, the meat sold in London in December of 2005 is considered that includes 5,553 observations. To achieve 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 others. The summary statistics of price (per pack) are given in Table 4. The histogram in Figure 1 shows the data to be positively skewed.

We used this data as population, where $\sum_{i=1}^{N} y_i = 2.3909$. To achieve PPS sampling, we first considered equal probability, i.e., $\pi_{V,i} = \frac{1}{\sum_j f_j} = \frac{1}{N}$ where $f_j, j = 1,\ldots,7$ is the frequency for each category and the summation of overall $\pi_{V,i}$ equals one. To provide unequal probabilities, we consider the frequency for each category: $\pi_{VI,i} = \frac{f_i}{\sum_j f_j^2}$ and $\pi_{VII,i} = \frac{f_i^2}{\sum_j f_j^3}$. The motivation behind such choices is logical, because the probability of elements is in terms of frequency. Table 5 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. The estimate of mean (variance) under the perfect ranking is given in Table 6. 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 certain prices, 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 6. The mechanism of imperfect ranking is the same as explained in the former subsection. Clearly, RPPS leads to a better estimate than PPS although as it is expected, the variance increases in comparison to the RPPS under imperfect ranking.

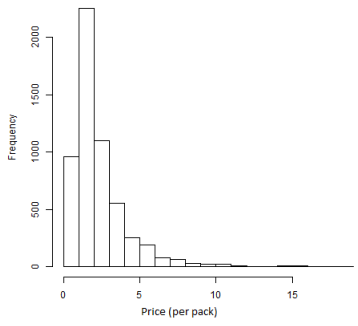

Figure 1: The histogram of the price of meat

Table 4: Summary statistics for the values of meat

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

Table 5: The frequency of meat's categories and the assigned probabilities

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

* :Meat,Veg, and Pastry

Table 6: 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)
		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)

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 consider 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 data base 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 a lower variance, and thus can be considered as an efficient sampling technique.

Throughout we have used a large number of simulated and real data examples to buttress the intuition behind the technique and formal results. There are many other methods that could be done, for instance we considered finite population with sampling with replacement. However, this method can be extended to sampling without replacement and also consider unbalanced RSS with missing data, but we leave these for future research. Another research is to explore the discussed method missing data might lead to unbalanced RSS.

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 authors.

We gratefully acknowledge the constructive comments and suggestions of the anonymous referees, and the associate editor.

References

Al-Saleh, M.F., & Zheng, G. (2002). Estimation of bivariate characteristics using ranked set sampling. *Australian and New Zealand Journal of Statistics*, 44, 221-232.

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.

Amiri, S., Modarres, R., & Zwanzig, S. (2016). Tests of perfect judgment ranking using pseudo-samples. Computational Statistics, 1-14.

Arnold, B.C., Balakrishnan, N., & Nagaraja, H. N. (2008). A first course in order statistics. Society for Industrial and Applied Mathematics.

Vock, M. & Balakrishnan, N. (2011). A Jonckheere-Terpstra-type test for perfect ranking in balanced ranked set sampling. J. Statist. Plann. Infer. 141, 624-630.

Chen, Z., Bai, Z., & Sinha, B.K. (2004). Ranked set sampling: theory and applications. Springer-Verlag, New York.

Chen, H., Stasny, E. A., & Wolfe, D. A. (2005). Ranked set sampling for efficient estimation of a population proportion. Statistics in medicine, 24(21), 3319-3329.

Cochran, W.G. (1977). Sampling Techniques. New York: Wiley.

Cox, D.R. (1969). Some sampling problems in technology. In New Developments in Survey Sampling, N. L. Johnson and J. R. Smith (eds), 506-527. New York: Wiley.

Deshpande, J.V., 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.

Frey, J., Ozturk, O. & Deshpande, J.V. (2007). Nonparametric tests for perfect judgment rankings. J. Amer. Statist. Assoc. 102, 708-717.

Halls, L.K., & Dell, T.R. (1966). Trial of ranked-set sampling for forage yields. Forest Science, 12(1), 22-26.

Hansen, M.H., & Hurwitz, W.N. (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.

Kvam, P. H. (2003). Ranked set sampling based on binary water quality data with covariates. Journal of agricultural, biological, and environmental statistics, 8(3), 271-279.

Muttlak, H. A., & McDonald, L. L. (1990). Ranked set sampling with size-biased probability of selection. Biometrics, 435-445.

Ozturk, O. (2016). Statistical inference based on judgment post-stratified samples in finite population. Survey Methodology, 42(2), 239-262.

Samawi, H.M., & AlSagheer, O.A. (2001). On the estimation of the distribution function using extreme and median ranked set sampling. Biometrical Journal, 43(3), 357-373.

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.

Zamanzade, E., & Vock, M. (2015). Variance estimation in ranked set sampling using a concomitant variable. Statistics & Probability Letters, 105, 1-5.

Mahdizadeh, M., & Zamanzade, E. (2016). A new reliability measure in ranked set sampling. Statistical Papers, 1-31.