Environ Ecol Stat (2011) 18:663–685
DOI 10.1007/s10651-010-0156-6

Design based estimation for ranked set sampling in finite populations

Mohammad Jafari Jozani · Brad C. Johnson

Received: 22 December 2009 / Revised: 10 August 2010 / Published online: 5 September 2010

Abstract In this paper, we consider design-based estimation using ranked set sampling (RSS) in finite populations. We first derive the first and second-order inclusion probabilities for an RSS design and present two Horvitz–Thompson type estimators using these inclusion probabilities. We also develop an alternate Hansen–Hurwitz type estimator and investigate its properties. In particular, we show that this alternate estimator always outperforms the usual Hansen–Hurwitz type estimator in the simple random sampling with replacement design with comparable sample size. We also develop formulae for ratio estimator for all three developed estimators. The theoretical results are augmented by numerical and simulation studies as well as a case study using a well known data set. These show that RSS design can yield a substantial improvement in efficiency over the usual simple random sampling design in finite populations.

Keywords Finite population · Hansen–Hurwitz estimator · Horvitz–Thompson estimator · Inclusion probability · Ratio estimator · Ranked set sampling

1 Introduction

In most sample surveys, a reasonable number of the sampling units (eg. households, businesses, etc.) can be ordered fairly accurately with respect to a variable of interest without actual measurement, and at little cost. However, exact measurement of these units may be very tedious and/or expensive. Ranked set sampling (RSS), as proposed by McIntyre (1952), provides an alternative to simple random sampling (SRS) in these situations. The feature of RSS is that it combines SRS with other sources of information

M. Jafari Jozani · B. C. Johnson (✉)
Department of Statistics, University of Manitoba, Winnipeg, MB R3T 2N2, Canada
e-mail: brad_johnson@umanitoba.ca

such as professional knowledge, judgment, auxiliary information, etc., to yield more representative measurements from the population. In its original form, RSS involves randomly drawing k units (called a set of size k) from the underlying population for which an estimate of the population mean (or total) is required. The units of this set are ranked by means of an auxiliary variable or some other ranking process such as judgmental ranking. For this ranked set, the unit ranked lowest is chosen for actual measurement of the variable of interest. A second set of size k is drawn and ranking is accomplished. The unit in the second lowest position is chosen and the variable of interest for this unit is quantified. Sampling is continued until, from the kth set, the kth ranked unit is measured. This entire process may be repeated m times (or cycles) to obtain a ranked set sample of size $n = mk$ from the underlying population.

As an alternative to the SRS design, RSS has achieved remarkable popularity in applied statistics. Several variations of the RSS design have been proposed and they have been used in many areas of application (see Kaur et al. 1995; Chen et al. 2004). A majority of the research on RSS has been concerned with estimating unknown parameters for infinite populations. However, in practice, we often have finite populations. Surprisingly, there has been little work related to the use of RSS techniques in finite populations, and most of that work has used model-based or model-assisted approaches to obtain a suitable estimator of the population mean. The problem of using RSS for finite populations was first studied by Takahasi and Futatsuya (1988). They showed that the RSS estimator of the population mean is unbiased and they derived an explicit formula for its variance when the set size is $k = 2$ and the underlying population has a discrete uniform distribution. Patil et al. (1995) extended this result to more general finite populations and to larger set sizes. Takahasi and Futatsuya (1998) showed that, when samples are drawn without replacement, the relative precision of the RSS estimator of the population mean relative to the SRS estimator with the same number of units quantified, is bounded above by 1. Bouza (2001) studied model-assisted ranked survey sampling design and proposed the use of a ratio and a simple linear regression super-population model in estimating the population mean. Bouza (2002a,b) studied an application of the RSS technique, with replacement, to estimate the population mean in the presence of non-responses. Further studies related to the theory and application of RSS designs in finite populations are provided in Barabesi and Marcheselli (2004), Ozturk et al. (2005), Deshpande et al. (2006), and Bouza (2009). Al-Saleh and Samawi (2007) studied the inclusion probabilities in RSS and some of its variation when the set size is $k = 2$ or $k = 3$ and, later on, Özdemir and Gökpınar (2007) extended this result to a more general set size and compared the inclusion probability for an RSS technique with that of a SRS with the same sample size. However, both Al-Saleh and Samawi (2007) and Özdemir and Gökpınar (2007) did not examine the second-order inclusion probabilities, which are necessary for determining the variance of Horvitz–Thompson estimators (π-estimators), and for determining if and when variance estimators are available.

In this paper we study the use of a variation of the RSS technique in finite populations described in Deshpande et al. (2006). We focus on a design-based approach to estimate the population mean or total. Our results show that RSS designs can be used in any large or small scale survey sampling to obtain more precise estimates, often at reduced cost.

The outline of the present paper is as follows. Section 2 introduces the notation used throughout. In Sect. 3, we develop the first and second-order inclusion probabilities for the RSS design considered and develop the properties of two Horvitz–Thompson type estimators under this design. Section 4 introduces an alternate Hansen–Hurwitz type estimator and develops its properties. We show that this estimator outperforms its counterpart under SRS design. In Sect. 5, we extend these results to ratio estimators and introduce an alternate estimator which performs extremely well compared to our proposed Horvitz–Thompson estimators. Section 6 provides some numerical and simulation studies as well as a case study using a well known data set. Specifically, we compare the proposed estimators with other estimators under standard designs. Section 7 provides some concluding comments.

2 Notation

Suppose we have a finite population of N elements, labeled $U = \{1, \ldots, N\}$, consisting of bivariate pairs $(y_1, x_1), \ldots, (y_N, x_N)$ where the study variable is y and x is an auxiliary variable. Throughout, we will assume that the $x_1, \ldots, x_N$ are unique. Let $x_{i:U}$ denote the ith ordered x value in the population and, for a simple random sample S, let $x_{i:S}$ denote the ith ordered x value in the sample (with realization $x_{i:s}$).

Suppose we make use of the following ranked set sampling scheme. Fix an m and k in $\mathbb{N}$ such that $mk \leq N$ and, for each $j = 1, \ldots, k$, take a simple random sample (without replacement), say $S_{1,j}$, of size k and select the population element $y_{[j:s_{1,j}]}$ associated with $x_{j:s_{1,j}}$, returning all elements to the population between samples. This results in the *ranked set sample* $rss_1 = \{y_{[1:s_{1,1}]}, y_{[2:s_{1,2}]}, \ldots, y_{[k:s_{1,k}]}\}$. Note, this (ordered) sample may contain repeated elements since it is certainly possible that $x_{j:s_{1,j}} = x_{j':s_{1,j'}}$, $j \neq j'$; $j, j' = 1, \ldots, k$. We repeat this m times (or cycles), obtaining the complete ranked set sample of size mk,

$$rss = \{\underbrace{y_{[1:s_{1,1}]}, \ldots, y_{[k:s_{1,k}]}}_{rss_1}, \quad \underbrace{y_{[1:s_{2,1}]}, \ldots, y_{[k:s_{2,k}]}}_{rss_2}, \quad \ldots, \quad \underbrace{y_{[1:s_{m,1}]}, \ldots, y_{[k:s_{m,k}]}}_{rss_m}\},$$

where $y_{[j:s_{r,j}]}$ denotes the element selected for the jth sample $s_{r,j}$ in the rth cycle.

We denote this RSS sampling design by RSSWR(m, k) since, after each without replacement sample $S_{r,j}$, we return all elements back into the population. This is the level 0 sampling design of Deshpande et al. (2006), where it is used to obtain nonparametric confidence intervals for quantiles in finite populations.

Sums over the ordered sample(s) rss and rss_r will be denoted

$$\sum_{rss_r} g(y_{[j:s_{r,j}]}) = \sum_{j=1}^{k} g(y_{[j:s_{r,j}]}) \quad \text{and} \quad \sum_{rss} g(y_{[j:s_{r,j}]}) = \sum_{r=1}^{m}\sum_{j=1}^{k} g(y_{[j:s_{r,j}]}),$$

where g is any suitable function. It will be convenient to introduce the set based samples rss_r^* and rss^*, which contain only the unique population elements $[i:U]$ contained

in rss_r and rss respectively. Sums over these sets will be denoted $\sum_{rss_r^*} g(y_{[i:U]})$ and $\sum_{rss^*} g(y_{[i:U]})$.

Remark 2.1 A few remarks are in order.

(i) Note that the RSSWR$(m, 1)$ design is equivalent to simple random sampling with replacement with m independent draws, which we denote by SRSWR(m).
(ii) In the RSSWR(m, k) design, the ranking is done according to the auxiliary variable x which, to be effective, should be easily observable and have a high correlation (in absolute value) with the variable of interest y. In our simulation study (see Sect. 6), we consider the effect of the correlation coefficient between x and y on the performance of our proposed estimators of the population total $t_y = \sum_U y_i$.
(iii) In the RSSWR(m, k) design, judgmental ranking can be done without the use of an auxiliary variable x. If ranking is done using x and the correlation coefficient between the auxiliary and response variable is perfect ($|\rho_{xy}| = 1$), then k should be chosen as large as is practical to increase the efficiency of the proposed estimators. However, if $|\rho_{xy}| < 1$, this choice of k may not be optimum.

In what follows, we will be making comparisons to other standard designs and estimators. We briefly describe these here (see for example, Särndal et al. 1992, for the details).

SRSWOR(n): In simple random sampling without replacement, the standard Horvitz–Thompson estimator (π-estimator) for $t_y = \sum_U y_i$, its variance and an unbiased variance estimator for this design will be denoted by $\hat{t}_{y,\pi}$, $\mathbb{V}(\hat{t}_{y,\pi})$ and $\hat{\mathbb{V}}(\hat{t}_{y,\pi})$ respectively.
SRSWR(n): In simple random sampling with replacement, the standard Hansen–Hurwitz type estimator (pwr-estimator) for t_y, its variance and an unbiased variance estimator for this design are denoted by $\hat{t}_{y,\mathrm{pwr}}$, $\mathbb{V}(\hat{t}_{y,\mathrm{pwr}})$ and $\hat{\mathbb{V}}(\hat{t}_{y,\mathrm{pwr}})$ respectively.
PPSWR(n): In probability proportional to size sampling with a (strictly positive) auxiliary variable x, probability $p_i = x_i / \sum_U x_i$ of selecting the ith element, and ordered sample $os = \{i_1, \ldots, i_n\}$, the standard Hansen–Hurwitz type estimator for t_y, its variance and an unbiased variance estimator are denoted by $\hat{t}_{y,\mathrm{pps}}$, $\mathbb{V}(\hat{t}_{y,\mathrm{pps}})$ and $\hat{\mathbb{V}}(\hat{t}_{y,\mathrm{pps}})$ respectively.

Throughout, we will use the following notation for common population and sample quantities

$$\bar{y}_U = \tfrac{1}{N}\textstyle\sum_U y_i, \quad S_{yU}^2 = \frac{\sum_U (y_i - \bar{y}_U)^2}{N-1}, \quad S_{yzU} = \frac{\sum_U (y_i - \bar{y}_U)(z_i - \bar{z}_U)}{N-1},$$
$$\bar{y}_s = \tfrac{1}{n}\textstyle\sum_s y_i, \quad S_{ys}^2 = \frac{\sum_s (y_i - \bar{y}_s)^2}{n-1}, \quad S_{yzs} = \frac{\sum_s (y_i - \bar{y}_s)(z_i - \bar{z}_s)}{N-1}.$$

3 Inclusion probabilities and the Horvitz–Thompson estimators

Define the first-order inclusion probability $\pi_{i:U}^{(m)}$ as the probability that element $[i:U]$ is included in rss^* and the second-order inclusion probability $\pi_{ii':U}^{(m)}$ as the probability that both $[i:U]$ and $[i':U]$ are included in rss^* under an RSSWR(m, k) design.

Lemma 3.1 *Let* $\alpha_{i,j} = \binom{i-1}{j-1}\binom{N-i}{k-j}\Big/\binom{N}{k}$. *Then, under the RSSWR*(m, k) *design, the first and second-order inclusion probabilities are given by*

$$\pi_{i:U}^{(m)} = 1 - \prod_{j=1}^{k} \left(1 - \alpha_{i,j}\right)^m,$$

and, for $i \neq i'$,

$$\pi_{ii':U}^{(m)} = 1 - \prod_{j=1}^{k} \left(1 - \alpha_{i,j}\right)^m - \prod_{j=1}^{k} \left(1 - \alpha_{i',j}\right)^m + \prod_{j=1}^{k} \left(1 - \alpha_{i,j} - \alpha_{i',j}\right)^m.$$

Proof Let $S_{r,j}$ be the jth simple random sample (of size k) in the rth cycle and define

$$A_{i,r,j} = \{x_{j:S_{r,j}} = x_{i:U}\}$$

as the event that the jth ordered x in $S_{r,j}$ is equal to the ith ordered x value in the population. It is easy to see that

$$\begin{aligned}\Pr\{A_{i,r,j}\} &= \Pr\{x_{j:S} = x_{i:U}\} \\ &= \binom{i-1}{j-1}\binom{N-i}{k-j}\Big/\binom{N}{k} := \alpha_{i,j}, \quad (\alpha_{i,j} := 0 \text{ for } i < j).\end{aligned}$$

Then, it is clear that $A_i = \bigcup_{r=1}^{m}\bigcup_{j=1}^{k} A_{i,r,j}$ is the event that $x_{i:U}$ (and hence $y_{[i:U]}$) is included in the ranked set sample rss at least once. Since the $S_{r,j}$ are independent, we have that the inclusion probability for $x_{i:U}$ (and hence $y_{[i:U]}$) under the RSSWR(m, k) design is

$$\begin{aligned}\pi_{i:U}^{(m)} := \Pr\{A_i\} &= \Pr\left\{\bigcup_{r=1}^{m}\bigcup_{j=1}^{k} A_{i,r,j}\right\} = 1 - \Pr\left\{\bigcap_{r=1}^{m}\bigcap_{j=1}^{k} A_{i,r,j}^c\right\} \\ &= 1 - \prod_{j=1}^{k}\left(1 - \alpha_{i,j}\right)^m.\end{aligned}$$

To determine the second-order inclusion probabilities $\pi_{ii':U}^{(m)}$ for $i \neq i'$, we have

$$\pi_{ii':U}^{(m)} = \Pr\{A_i \cap A_{i'}\} = 1 - \Pr\{A_i^c\} - \Pr\{A_{i'}^c\} + \Pr\{A_i^c \cap A_{i'}^c\}.$$

But, using the same arguments as above, we have $\Pr\{A_i^c \cap A_{i'}^c\} = \prod_{j=1}^{k}(1-\alpha_{i,j}-\alpha_{i',j})^m$ and hence the result follows. □

Lemma 3.1 leads to the following interesting properties of the inclusion probabilities and allows for some comparisons of the RSSWR design with the SRSWR design.

(i) For $k = 1, \alpha_{i,1} = 1/N$ and we obtain the usual inclusion probabilities for the SRSWR(m) design, namely $\pi_i = 1-(1-1/N)^m$ and $\pi_{ii'} = 1-2(1-1/N)^m + (1-2/N)^m$.

(ii) By definition, $\alpha_{i,j} = \alpha_{N-i+1,j}$ for any $j \in \{1,\ldots,k\}$, and hence $\pi_{i:U}^{(m)} = \pi_{N-i+1:U}^{(m)}$ for all $i = 1,\ldots,N$. That is, the inclusion probabilities under the RSSWR(m,k) design are symmetric with respect to the ordering induced by the $x_{i:U}$.

(iii) Since, $\alpha_{1,j} = k/N$ if $j = 1$ and 0 otherwise, we have

$$\pi_{1:U}^{(m)} = 1-\prod_{j=1}^{k}(1-\alpha_{1,j})^m = 1-\left(1-\frac{k}{N}\right)^m \geq 1-\left(1-\frac{1}{N}\right)^{mk} = \pi_1,$$

with equality only if $k = 1$. Hence, by the symmetry in (ii), the inclusion probability of observing the minimum (or the maximum) value of the population in an RSSWR(m,k) sample is greater than its counterpart in a SRSWR(mk) design whenever $k > 1$, which is given by $\pi_1 = \pi_N = 1-(1-\frac{1}{N})^{mk}$.

(iv) The probability of observing a very unrepresentative sample in the RSSWR (m,k) design is smaller than in the SRSWR(mk) design. For example, probability of observing a sample consisting entirely of $y_{1:U}$ or $y_{N:U}$ in the SRSWR(mk) is $\frac{2}{N^{mk}}$. However, in the RSSWR(m,k) design with $k > 1$, this probability is zero.

(v) The inclusion probabilities developed above are much simpler than those derived in, for example, Al-Saleh and Samawi (2007) and Özdemir and Gökpınar (2007), which make use of some without replacement variations of RSS.

The first and second-order inclusion probabilities lead to two different Horvitz–Thompson type estimators of the population total (or mean).

3.1 A Horvitz–Thompson estimator

Recall that *rss* is the complete ordered sample and that it may contain duplicates. Also recall that *rss** denotes the unique population elements that are contained in *rss*. Then, by the usual theory (cf. Särndal et al. 1992, Chapter 3), the following Theorem holds.

Theorem 3.2 *Consider the RSSWR(m,k) design. Then the estimator*

$$\hat{t}_{y,\mathrm{RSS},\pi} = \sum_{rss^*} \frac{y_{[i:U]}}{\pi_{i:U}^{(m)}} \tag{3.1}$$

is unbiased for $t_y = \sum_U y_i$ *with variance*

$$\mathbb{V}(\hat{t}_{y,\mathrm{RSS},\pi}) = \sum_{i\in U}\sum_{i'\in U} \Delta^{(m)}_{ii':U} \frac{y_{[i:U]}}{\pi^{(m)}_{i:U}} \frac{y_{[i':U]}}{\pi^{(m)}_{i':U}}, \tag{3.2}$$

where $\Delta^{(m)}_{ii':U} := \pi^{(m)}_{ii':U} - \pi^{(m)}_{i:U}\pi^{(m)}_{i':U}$. *An unbiased estimator of* $\mathbb{V}(\hat{t}_{y,\mathrm{RSS},\pi})$ *is given by*

$$\hat{\mathbb{V}}(\hat{t}_{y,\mathrm{RSS},\pi}) = \sum_{i\in rss^*}\sum_{i'\in rss^*} \frac{\Delta^{(m)}_{ii':U}}{\pi^{(m)}_{ii':U}} \frac{y_{[i:U]}}{\pi^{(m)}_{i:U}} \frac{y_{[i':U]}}{\pi^{(m)}_{i':U}}. \tag{3.3}$$

If z *is an additional study variable with corresponding estimate* $\hat{t}_{z,RSS,\pi}$, *then*

$$\mathrm{Cov}(\hat{t}_{y,\mathrm{RSS},\pi}, \hat{t}_{z,\mathrm{RSS},\pi}) = \sum_{i\in U}\sum_{i'\in U} \Delta^{(m)}_{ii':U} \frac{y_{[i:U]}}{\pi^{(m)}_{i:U}} \frac{z_{[i':U]}}{\pi^{(m)}_{i'^3:U}}, \tag{3.4}$$

with unbiased estimator

$$\widehat{\mathrm{Cov}}(\hat{t}_{y,\mathrm{RSS},\pi}, \hat{t}_{z,\mathrm{RSS},\pi}) = \sum_{i\in rss^*}\sum_{i'\in rss^*} \frac{\Delta^{(m)}_{ii':U}}{\pi^{(m)}_{ii':U}} \frac{y_{[i:U]}}{\pi^{(m)}_{i:U}} \frac{z_{[i':U]}}{\pi^{(m)}_{i':U}}. \tag{3.5}$$

The Horvitz–Tompson estimator $\hat{t}_{y,\mathrm{RSS},\pi}$ in (3.1) suffers from a few problems.

(i) If the coefficient of variation of the ys ($cv(y) := S_{yU}/\bar{y}_U$) is small, then variations in the number of elements in rss^* can cause instability in the estimator $\hat{t}_{y,\mathrm{RSS},\pi}$. Indeed, the sampling distribution of $\hat{t}_{y,\mathrm{RSS},\pi}$ can have multiple modes. Section 3.2 introduces a slightly more stable version of this estimator.

(ii) We need to know all of the x_i in order to calculate the inclusion probabilities. While this is the case in many situations, in others, we only know the x_i in the observed samples $s_{r,j}$, and hence a procedure that does not require knowledge of all of the x_i may be beneficial. This is addressed in Sect. 4.

(iii) While calculation of the variance and variance estimator is easily automated, the form does not simplify and hence theoretical comparisons of efficiency with other designs is difficult at best.

3.2 An alternate Horvitz–Thompson type estimator

To address concern (i), and to reduce the effect of the duplication of population elements in *rss*, we may consider independent estimation within each cycle in the RSSWR(m, k) design. To this end, for each cycle $r = 1, \ldots, m$, we consider the estimator

$$\hat{t}^{(1)}_{y,\pi,r} = \sum_{i\in rss^*_r} \frac{y_{[i:U]}}{\pi^{(1)}_{i:U}},$$

where the $\pi^{(1)}_{i:U}$ are now determined assuming $m = 1$, and the sum is over the distinct population elements in rss_r. We now define the alternate Horvitz–Thompson type estimator of t_y as the arithmetic mean of the estimators from the m cycles,

$$\hat{t}_{y,\mathrm{RSS},m,\pi} = \frac{1}{m}\sum_{r=1}^{m}\sum_{i\in rss_r^*}\frac{y_{[i:U]}}{\pi^{(1)}_{i:U}} = \frac{1}{m}\sum_{r=1}^{m}\hat{t}^{(1)}_{y,\pi,r}, \tag{3.6}$$

Again, by standard results, we obtain the following.

Theorem 3.3 *For the RSSWR(m,k) design, the estimator in (3.6) is unbiased for t_y with variance*

$$\mathbb{V}(\hat{t}_{y,\mathrm{RSS},m,\pi}) = \frac{1}{m}\sum_{i\in U}\sum_{i'\in U}\Delta^{(1)}_{ii':U}\frac{y_{[i:U]}}{\pi^{(1)}_{i:U}}\frac{y_{[i':U]}}{\pi^{(1)}_{i':U}} = \frac{\mathbb{V}(\hat{t}^{(1)}_{y,\pi,r})}{m}, \tag{3.7}$$

where $\Delta^{(1)}_{ii':U} := \pi^{(1)}_{ii':U} - \pi^{(1)}_{i:U}\pi^{(1)}_{i':U}$. An unbiased estimator of $\mathbb{V}(\hat{t}_{y,\mathrm{RSS},m,\pi})$ is given by

$$\hat{\mathbb{V}}(\hat{t}_{y,\mathrm{RSS},m,\pi}) = \frac{1}{m^2}\sum_{r=1}^{m}\sum_{i\in rss_r^*}\sum_{i'\in rss_r^*}\frac{\Delta^{(1)}_{ii':U}}{\pi^{(1)}_{ii':U}}\frac{y_{[i:U]}}{\pi^{(1)}_{i:U}}\frac{y_{[i':U]}}{\pi^{(1)}_{i':U}} = \frac{1}{m^2}\sum_{r=1}^{m}\hat{\mathbb{V}}(\hat{t}^{(1)}_{y,\pi,r}). \tag{3.8}$$

If $m \geq 2$, an alternate unbiased estimator of $\mathbb{V}(\hat{t}_{y,\mathrm{RSS},m,\pi})$ is given by

$$\hat{\mathbb{V}}(\hat{t}_{y,\mathrm{RSS},m,\pi}) = \frac{1}{m(m-1)}\sum_{r=1}^{m}(\hat{t}^{(1)}_{y,\pi,r} - \hat{t}_{y,\mathrm{RSS},m,\pi})^2. \tag{3.9}$$

If z is an additional study (or auxiliary) variable, with corresponding estimate $\hat{t}_{z,\mathrm{RSS},m,\pi}$, then

$$\mathrm{Cov}(\hat{t}_{y,\mathrm{RSS},m,\pi}, \hat{t}_{z,\mathrm{RSS},m,\pi}) = \frac{1}{m}\sum_{i\in U}\sum_{i'\in U}\Delta^{(1)}_{ii':U}\frac{y_{[i:U]}}{\pi^{(1)}_{i:U}}\frac{z_{[i':U]}}{\pi^{(1)}_{i':U}}. \tag{3.10}$$

An unbiased estimator of $\mathrm{Cov}(\hat{t}_{y,\mathrm{RSS},m,\pi}, \hat{t}_{z,\mathrm{RSS},m,\pi})$ is given by

$$\begin{aligned}\widehat{\mathrm{Cov}}(\hat{t}_{y,\mathrm{RSS},m,\pi}, \hat{t}_{z,\mathrm{RSS},m,\pi}) &= \frac{1}{m^2}\sum_{r=1}^{m}\sum_{i\in rss_r^*}\sum_{i'\in rss_r^*}\frac{\Delta^{(1)}_{ii':U}}{\pi^{(1)}_{ii':U}}\frac{y_{[i:U]}}{\pi^{(1)}_{i:U}}\frac{z_{[i':U]}}{\pi^{(1)}_{i':U}} \\ &= \frac{1}{m^2}\sum_{r=1}^{m}\widehat{\mathrm{Cov}}(\hat{t}^{(1)}_{y,\pi,r}, \hat{t}^{(1)}_{z,\pi,r}).\end{aligned} \tag{3.11}$$

Also, if $m \geq 2$, then an alternate unbiased covariance estimator is given by

$$\widehat{\mathrm{Cov}}(\hat{t}_{y,\mathrm{RSS},m,\pi}, \hat{t}_{z,\mathrm{RSS},m,\pi}) = \frac{1}{m(m-1)} \sum_{r=1}^{m} (\hat{t}^{(1)}_{y,\pi,r} - \hat{t}_{y,\mathrm{RSS},m,\pi})(\hat{t}^{(1)}_{z,\pi,r} - \hat{t}_{z,\mathrm{RSS},m,\pi}). \tag{3.12}$$

Proof Equations (3.6)–(3.8), (3.10) and (3.11) follow from standard theory (cf. Särndal et al. 1992). To show (3.9), let $a_r = \hat{t}^{(1)}_{y,\pi,r}$ so that $\hat{t}_{y,\mathrm{RSS},m,\pi} = \bar{a}$. In this notation, we have

$$\mathbb{E}[\hat{\mathbb{V}}(\hat{t}_{y,\mathrm{RSS},m,\pi})] = \mathbb{E}\left[\frac{1}{m(m-1)} \sum_{r=1}^{m} (a_r - \bar{a})^2\right] = \frac{\mathbb{V}(a_r)}{m} = \frac{\mathbb{V}(\hat{t}^{(1)}_{y,\pi,r})}{m},$$

which is equivalent to (3.7). Equation (3.12) follows similarly.

The advantage of the alternate forms of the variance and covariance estimators in (3.9) and (3.12) respectively is that they do not require the calculation of the second-order inclusion probabilities. We have not, however, undertaken a theoretical analysis to determine which variance estimator is more precise.

4 A Hansen–Hurwitz type unbiased estimator

As mentioned, the Horvitz–Thompson type estimators suffer from a certain amount of instability and this effect can be very large if $cv(y)$ is small. To overcome this instability, we propose an alternate estimator of t_y. Consider the RSSWR(m, k) design and let $\delta_{[i:U]}$ denote the number of times $y_{[i:U]}$ appears in the sample *rss*, then

$$\mathbb{E}(\delta_{[i:U]}) = \sum_{r=1}^{m} \sum_{j=1}^{k} \Pr\{A_{i,r,j}\} = \sum_{r=1}^{m} \sum_{j=1}^{k} \alpha_{i,j} = m \sum_{j=1}^{k} \alpha_{i,j} = \frac{mk}{N}.$$

Now, we construct the alternate Hansen–Hurwitz type estimator of t_y as follows

$$\hat{t}_{y,\mathrm{RSS,pwr}} = \sum_{r=1}^{m} \sum_{j=1}^{k} \frac{y_{[j:s_{r,j}]}}{\mathbb{E}(\delta_{[j:s_{r,j}]})} = \frac{N}{mk} \sum_{r=1}^{m} \sum_{j=1}^{k} y_{[j:s_{r,j}]}. \tag{4.1}$$

Before developing the properties of this estimator, we present the following Lemma, which describes a key relationship between the mean of certain functionals of an RSSWR(m, k) sample and the corresponding mean in the underlying finite population. This result mirrors that of the usual RSS from infinite populations.

Lemma 4.1 *Let $g : U \to \mathbb{R}$. Then, for fixed r,*

$$\mathbb{E}\left(\frac{1}{k} \sum_{j=1}^{k} g(y_{[j:S_{r,j}]})\right) = \frac{1}{N} \sum_{i=1}^{N} g(y_i).$$

Proof Recall that $\Pr\{x_{j:S_{r,j}} = x_{i:U}\} = \alpha_{i,j} = \binom{i-1}{j-1}\binom{N-i}{k-j}\Big/\binom{N}{k}$. We have

$$\begin{aligned}\mathbb{E}[g(y_{[j:S_{r,j}]})] &= \sum_{i=1}^{N} \mathbb{E}[\, g(y_{[j:S_{r,j}]}) \mid x_{j:S_{r,j}} = x_{i:U}\,] \Pr\{x_{j:S_{r,j}} = x_{i:U}\} \\ &= \sum_{i=1}^{N} g(y_{[i:U]})\alpha_{i,j}.\end{aligned}$$

Therefore, since $\sum_{j=1}^{k} \alpha_{i,j} = k/N$, we obtain

$$\sum_{j=1}^{k} \mathbb{E}(g(y_{[j:S_{r,j}]})) = \sum_{j=1}^{k}\sum_{i=1}^{N} g(y_{[i:U]})\alpha_{i,j} = \frac{k}{N}\sum_{i=1}^{N} g(y_{[i:U]}) = \frac{k}{N}\sum_{i=1}^{N} g(y_i).$$

□

The same arguments used in the proof of Lemma 4.1 also lead to the following corollary.

Corollary 4.2 *Let $g, h : U \to \mathbb{R}$ and let z be an additional variable. Then, for fixed r,*

$$\mathbb{E}\left(\frac{1}{k}\sum_{j=1}^{k} g(y_{[j:S_{r,j}]})h(z_{[j:S_{r,j}]})\right) = \frac{1}{N}\sum_{i=1}^{N} g(y_i)h(z_i).$$

We are now in a position to prove the following Theorem.

Theorem 4.3 *For the RSSWR(m, k) design, the estimator in (4.1) is unbiased for t_y with variance*

$$\mathbb{V}(\hat{t}_{y,\mathrm{RSS,pwr}}) = \frac{N(N-1)}{mk} S_{yU}^2 - \frac{N^2}{mk^2}\sum_{j=1}^{k} (\mu_{y[j]} - \bar{y}_U)^2, \tag{4.2}$$

where $\mu_{y[j]} = \mathbb{E}(y_{[j:S_{r,j}]})$. If $m \geq 2$, an unbiased estimator of $\mathbb{V}(\hat{t}_{y,\mathrm{RSS,pwr}})$ is given by

$$\hat{\mathbb{V}}(\hat{t}_{y,\mathrm{RSS,pwr}}) = \frac{1}{m(m-1)}\sum_{r=1}^{m} (\hat{t}_{yr} - \hat{t}_{y,\mathrm{RSS,pwr}})^2, \tag{4.3}$$

where $\hat{t}_{yr} = N\sum_{j=1}^{k} y_{[j:s_{r,j}]}/k, r = 1,\ldots,m$. *If* z *is an additional study variable, with corresponding estimate* $\hat{t}_{z,\mathrm{RSS,pwr}}$, *then*

$$\mathrm{Cov}(\hat{t}_{y,\mathrm{RSS,pwr}}, \hat{t}_{z,\mathrm{RSS,pwr}}) = \frac{N(N-1)}{mk} S_{yzU} - \frac{N^2}{mk^2}\sum_{j=1}^{k}(\mu_{y[j]} - \bar{y}_U)(\mu_{z[j]} - \bar{z}_U), \tag{4.4}$$

where $\mu_{z[j]} = \mathbb{E}(z_{[j:S_{r,j}]})$. *If* $m \geq 2$, *an unbiased estimator of* $\mathrm{Cov}(\hat{t}_{y,\mathrm{RSS,pwr}}, \hat{t}_{z,\mathrm{RSS,pwr}})$ *is*

$$\widehat{\mathrm{Cov}}(\hat{t}_{y,\mathrm{RSS,pwr}}, \hat{t}_{z,\mathrm{RSS,pwr}}) = \frac{1}{m(m-1)}\sum_{r=1}^{m}(\hat{t}_{yr} - \hat{t}_{y,\mathrm{RSS,pwr}})(\hat{t}_{zr} - \hat{t}_{z,\mathrm{RSS,pwr}}). \tag{4.5}$$

Proof For each $r = 1,\ldots,m$, let $\hat{t}_{yr} = N\sum_{j=1}^{k} y_{[j:s_{r,j}]}/k$. Then

$$\hat{t}_{y,\mathrm{RSS,pwr}} = \frac{N}{mk}\sum_{r=1}^{m}\sum_{j=1}^{k} y_{[j:s_{r,j}]} = \frac{1}{m}\sum_{r=1}^{m}\left(\frac{N}{k}\sum_{j=1}^{k} y_{[j:s_{r,j}]}\right) = \frac{1}{m}\sum_{r=1}^{m}\hat{t}_{yr}.$$

Taking $g(x) = x$ in Lemma 4.1 immediately shows that $\hat{t}_{yr}$ is unbiased for t_y and hence $\hat{t}_{y,\mathrm{RSS,pwr}}$ is also unbiased for t_y, being the arithmetic average of the $\hat{t}_{yr}$. To show (4.2), we first note that, since the $S_{r,j}$ are independent, so are the $y_{[j:S_{r,j}]}$ and hence

$$\mathbb{V}(\hat{t}_{y,\mathrm{RSS,pwr}}) = \mathbb{V}\left(\frac{N}{mk}\sum_{r=1}^{m}\sum_{j=1}^{k} y_{[j:S_{r,j}]}\right) = \frac{N^2}{m^2k^2}\sum_{r=1}^{m}\sum_{j=1}^{k}\mathbb{V}(y_{[j:S_{r,j}]}).$$

Therefore, we need to determine $\mathbb{V}(y_{[j:S_{r,j}]})$. Letting $\mu_{y[j]} = \mathbb{E}(y_{[j:S_{r,j}]})$, we have

$$\begin{aligned}\mathbb{V}(y_{[j:S_{r,j}]}) = \mathbb{E}[(y_{[j:S_{r,j}]} - \mu_{y[j]})^2] &= \mathbb{E}[((y_{[j:S_{r,j}]} - \bar{y}_U) - (\mu_{y[j]} - \bar{y}_U))^2] \\ &= \mathbb{E}[(y_{[j:S_{r,j}]} - \bar{y}_U)^2] - (\mu_{y[j]} - \bar{y}_U)^2.\end{aligned}$$

Taking $g(x) = (x - \bar{y}_U)^2$ in Lemma 4.1 yields

$$\sum_{j=1}^{k}\mathbb{E}[(y_{[j:S_{r,j}]} - \bar{y}_U)^2] = \frac{k}{N}\sum_{i=1}^{N}(y_i - \bar{y}_U)^2 = \frac{k(N-1)}{N}S_{yU}^2.$$

Hence

$$\mathbb{V}(\hat{t}_{y,\mathrm{RSS,pwr}}) = \frac{N(N-1)}{mk}S_{yU}^2 - \frac{N^2}{mk^2}\sum_{j=1}^{k}(\mu_{y[j]} - \bar{y}_U)^2,$$

which shows (4.2). It remains to show that (4.3) is unbiased. To this end, we write

$$\mathbb{V}(\hat{t}_{y,\mathrm{RSS,pwr}}) = \frac{1}{m^2}\sum_{r=1}^{m}\mathbb{V}(\hat{t}_{yr}) = \frac{\mathbb{V}(\hat{t}_{yr})}{m}.$$

Hence, the same argument in the proof of Theorem 3.3 shows that $\hat{\mathbb{V}}(\hat{t}_{y,\mathrm{RSS,pwr}})$ is unbiased for $\mathbb{V}(\hat{t}_{y,\mathrm{RSS,pwr}})$. For the covariance in (4.4), we have

$$\mathrm{Cov}(\hat{t}_{y,\mathrm{RSS,pwr}}, \hat{t}_{z,\mathrm{RSS,pwr}}) = \frac{1}{m}\mathrm{Cov}(\hat{t}_{yr}, \hat{t}_{zr}) = \frac{N^2}{mk^2}\sum_{j=1}^{k}\mathrm{Cov}(y_{[j:S_{r,j}]}, z_{[j:S_{r,j}]}).$$

But,

$$\begin{aligned}\sum_{j=1}^{k}\mathrm{Cov}(y_{[j:S_{r,j}]}, z_{[j:S_{r,j}]}) &= \sum_{j=1}^{k}\mathbb{E}\left[(y_{[j:S_{r,j}]} - \mu_{y[j]})(z_{[j:S_{r,j}]} - \mu_{z[j]})\right]\\ &= \sum_{j=1}^{k}\mathbb{E}\left[(y_{[j:S_{r,j}]} - \bar{y}_U)(z_{[j:S_{r,j}]} - \bar{z}_U)\right]\\ &\quad - \sum_{j=1}^{k}(\mu_{y[j]} - \bar{y}_U)(\mu_{z[j]} - \bar{z}_U).\end{aligned}$$

By Corollary 4.2, the first term becomes

$$\sum_{j=1}^{k}\mathbb{E}\left[(y_{[j:S_{r,j}]} - \bar{y}_U)(z_{[j:S_{r,j}]} - \bar{z}_U)\right] = \frac{k}{N}\sum_{i=1}^{N}(y_i - \bar{y}_U)(z_i - \bar{z}_U) = \frac{k(N-1)}{N}S_{yzU}.$$

Therefore,

$$\mathrm{Cov}(\hat{t}_{y,\mathrm{RSS,pwr},y}, \hat{t}_{y,\mathrm{RSS,pwr},z}) = \frac{N(N-1)}{mk}S_{yzU} - \frac{N^2}{mk^2}\sum_{j=1}^{k}(\mu_{y[j]} - \bar{y}_U)(\mu_{z[j]} - \bar{z}_U).$$

That (4.5) is unbiased for (4.4) follows from the same argument as in Theorem 3.3. □

Note that, if the ranking of the ys is completely random with respect to the xs, then $\mu_{y[j]} = \bar{y}_U$ for all j and hence $\sum_{j=1}^{k}(\mu_{y[j]} - \bar{y}_U)^2 = 0$, so that

$$\mathbb{V}(\hat{t}_{y,\mathrm{RSS,pwr}}) = \frac{N(N-1)}{mk}S_{yU}^2,$$

which is the variance of the estimator $\hat{t}_{y,\mathrm{pwr}}$ under the SRSWR(mk) design. This says that the $\hat{t}_{y,\mathrm{RSS,pwr}}$ estimator under the RSSWR(m, k) design always performs at least

as well as the $\hat{t}_{y,\text{pwr}}$ estimator under the SRSWR(mk) design. If the ranking is reasonable or perfect (i.e. $y_{i:U} = y_{[i:U]}$), it can perform considerably better. Note also that, $\hat{t}_{y,\text{RSS,pwr}}$ and $\hat{\mathbb{V}}(\hat{t}_{y,\text{RSS,pwr}})$ may be calculated without the knowledge of all of the x_i in the population. That is, we need only know the relative orderings of the $x_{j:s_{r,j}}$ within each $s_{r,j}$. This is a considerable advantage compared to the situation discussed in Sect. 3 and to other designs which require a knowledge of all of the x_i (like the PPSWR design).

5 Ratio estimators under the RSSWR(m, k) design

Let z be an additional auxiliary variable, for which $t_z = \sum_U z_i$ is known (in the sequel, we will take $z = x$, but this is not required). The ratio estimator of $t_y = \sum_U y_i$ is given by $\hat{t}_{yra} = t_z \hat{R}$, where $\hat{R} = \hat{t}_y/\hat{t}_z$ and $\hat{t}_y$ and $\hat{t}_z$ are estimates of the population totals t_y and t_z respectively, which may be based on any sampling design and estimation technique. From general theory (cf. Särndal et al. 1992, Chapter 5), we know that $\hat{t}_{yra}$ is approximately unbiased for t_y and that the approximate variance of $\hat{t}_{yra}$ is given by

$$AV(\hat{t}_{yra}) = \mathbb{V}(\hat{t}_y) + R^2\mathbb{V}(\hat{t}_z) - 2R\text{Cov}(\hat{t}_y, \hat{t}_z),$$

where $R = t_y/t_z$. The variance estimator is given by

$$\hat{\mathbb{V}}(\hat{t}_{yra}) = \frac{t_z^2}{\hat{t}_z^2}\left[\hat{\mathbb{V}}(\hat{t}_y) + \hat{R}^2\hat{\mathbb{V}}(\hat{t}_z) - 2\widehat{R\text{Cov}}(\hat{t}_y, \hat{t}_z)\right],$$

and the form of the variance and covariance estimators depends on the design used to estimate t_y and t_z. For the SRSWOR(n) design, it is well established (cf. Särndal et al. 1992, Chapter 5 again) that the ratio estimator, its approximate variance and a variance estimator, when $\hat{t}_y$ and $\hat{t}_z$ are based on the Horvitz–Thompson estimators, are given by

$$\hat{t}_{y,\pi,ra} = t_z \frac{\hat{t}_{y,\pi}}{\hat{t}_{z,\pi}}, \tag{5.1}$$

$$AV(\hat{t}_{y,\pi,ra}) = \frac{N^2(1-n/N)}{n}\left(S_{yU}^2 + R^2 S_{zU}^2 - 2RS_{yzU}\right), \tag{5.2}$$

$$\hat{\mathbb{V}}(\hat{t}_{y,\pi,ra}) = \frac{N^2(1-n/N)}{n}\left(S_{ys}^2 + R^2 S_{zs}^2 - 2RS_{yzs}\right). \tag{5.3}$$

We now present the following result, which gives the formulae for the ratio estimator under the RSSWR(m, k) design and using the estimators $\hat{t}_{\cdot,\text{RSS,pwr}}$. Results for the ratio estimators based on $\hat{t}_{\cdot,\text{RSS},\pi}$ and $\hat{t}_{\cdot,\text{RSS},m,\pi}$ follow similarly.

Theorem 5.1 *Under the RSSWR(m, k) design with study variable y, auxiliary variable z, and ranking variable x, the ratio estimator*

$$\hat{t}_{y,\text{RSS,pwr},ra} = t_z \frac{\hat{t}_{y,\text{RSS,pwr}}}{\hat{t}_{z,\text{RSS,pwr}}} \tag{5.4}$$

is approximately unbiased for t_y. The approximate variance of $\hat{t}_{y,\mathrm{RSS,pwr},ra}$ is given by

$$AV(\hat{t}_{y,\mathrm{RSS,pwr},ra}) = \mathbb{V}(\hat{t}_{y,\mathrm{RSS,pwr}}) + R^2\mathbb{V}(\hat{t}_{z,\mathrm{RSS,pwr}}) - 2R\mathrm{Cov}(\hat{t}_{y,\mathrm{RSS,pwr}}, \hat{t}_{z,\mathrm{RSS,pwr}}), \tag{5.5}$$

where $\mathbb{V}(\hat{t}_{y,\mathrm{RSS,pwr}})$, $\mathbb{V}(\hat{t}_{z,\mathrm{RSS,pwr}})$ and $\mathrm{Cov}(\hat{t}_{y,\mathrm{RSS,pwr}}, \hat{t}_{z,\mathrm{RSS,pwr}})$ are given in Theorem 4.3. If $m \geq 2$, the variance estimator is given by

$$\hat{\mathbb{V}}(\hat{t}_{y,\mathrm{RSS,pwr},ra}) = \frac{t_z^2}{\hat{t}_{z,\mathrm{RSS,pwr}}^2}\Big[\hat{\mathbb{V}}(\hat{t}_{y,\mathrm{RSS,pwr}}) + \hat{R}^2\hat{\mathbb{V}}(\hat{t}_{z,\mathrm{RSS,pwr}}) - 2\hat{R}\widehat{\mathrm{Cov}}(\hat{t}_{y,\mathrm{RSS,pwr}}, \hat{t}_{z,\mathrm{RSS,pwr}})\Big], \tag{5.6}$$

where $\hat{\mathbb{V}}(\hat{t}_{y,\mathrm{RSS,pwr}})$, $\hat{\mathbb{V}}(\hat{t}_{z,\mathrm{RSS,pwr}})$ and $\widehat{\mathrm{Cov}}(\hat{t}_{y,\mathrm{RSS,pwr}}, \hat{t}_{z,\mathrm{RSS,pwr}})$ are given by Theorem 4.3.

Note that in Theorem 5.1 we may write $AV(\hat{t}_{y,\mathrm{RSS,pwr},ra})$ as

$$AV(\hat{t}_{y,\mathrm{RSS,pwr},ra}) = AV(\hat{t}_{y,\mathrm{pwr},ra}) - \frac{N^2}{mk^2}\sum_{j=1}^{k}\left((\mu_{y[j]} - \bar{y}_U) - R(\mu_{z[j]} - \bar{z}_U)\right)^2, \tag{5.7}$$

which shows that $AV(\hat{t}_{y,\mathrm{RSS,pwr},ra}) \leq AV(\hat{t}_{y,\mathrm{pwr},ra})$ where $\hat{t}_{y,\mathrm{pwr},ra}$ is the ratio estimator under the SRSWR(mk) = RSSWR(mk, 1) design. As mentioned previously, in all above results one can take $z = x$, thus making use of the information contained in x twice—once for ranking and once for the ratio estimate. This is precisely what we do in the numerical results.

5.1 An alternate estimator

The above discussion on ratio estimators suggests the alternate estimator of t_y as follows

$$\tilde{t}_{y,\mathrm{RSS}} = N\frac{\hat{t}_{y,\mathrm{RSS},\pi}}{\hat{N}_{\mathrm{RSS},\pi}} \quad \text{where} \quad \hat{N}_{\mathrm{RSS},\pi} = \sum\nolimits_{rss^*}\frac{1}{\pi_{i:N}^{(m)}}. \tag{5.8}$$

This estimator avoids most of the instability problems associated with $\hat{t}_{y,\mathrm{RSS},\pi}$ mentioned earlier since the number of terms in the denominator is the same as the number of terms in the numerator. The numerical results in the next section show that this estimator can perform substantially better than $\hat{t}_{y,\mathrm{RSS},\pi}$, especially when $cv(y)$ is small. Again, by standard results, we have

Theorem 5.2 *Let* $\tilde{t}_{y,\mathrm{RSS}}$ *and* $\hat{N}_{\mathrm{RSS},\pi}$ *be as in* (5.8)*. Then,* $\tilde{t}_{y,\mathrm{RSS}}$ *is approximately unbiased for* t_y *with approximate variance*

$$AV(\tilde{t}_{y,\mathrm{RSS}}) = \mathbb{V}(\hat{t}_{y,\mathrm{RSS},\pi}) + \bar{y}_U^2 \mathbb{V}(\hat{N}_{\mathrm{RSS},\pi}) - 2\bar{y}_U \mathrm{Cov}(\hat{t}_{y,\mathrm{RSS},\pi}, \hat{N}_{\mathrm{RSS},\pi}),$$

where $\mathbb{V}(\hat{t}_{y,\mathrm{RSS},\pi})$ *is given in Theorem* 3.2*,*

$$\mathbb{V}(\hat{N}_{\mathrm{RSS},\pi}) = \sum_{i\in U}\sum_{i'\in U} \frac{\Delta^{(m)}_{ii':U}}{\pi^{(m)}_{i:U}\pi^{(m)}_{i':U}},$$

and

$$\mathrm{Cov}(\hat{t}_{y,\mathrm{RSS},\pi}, \hat{N}_{\mathrm{RSS},\pi}) = \sum_{i\in U}\sum_{i'\in U} \frac{\Delta^{(m)}_{ii':U}}{\pi^{(m)}_{i:U}\pi^{(m)}_{i':U}} y_i.$$

A variance estimator is given by

$$\hat{\mathbb{V}}(\tilde{t}_{y,\mathrm{RSS}}) = \left(\frac{N}{\hat{N}_{\mathrm{RSS},\pi}}\right)^2 \left[\hat{\mathbb{V}}(\hat{t}_{y,\mathrm{RSS},\pi}) + \tilde{y}_s^2 \hat{\mathbb{V}}(\hat{N}_{\mathrm{RSS},\pi}) - 2\tilde{y}_s \widehat{\mathrm{Cov}}(\hat{t}_{y,\mathrm{RSS},\pi}, \hat{N}_{\mathrm{RSS},\pi})\right],$$

where $\hat{\mathbb{V}}(\hat{t}_{y,\mathrm{RSS},\pi})$ *is given in Theorem* 3.2*,* $\tilde{y}_s = \hat{t}_{y,\mathrm{RSS},\pi}/\hat{N}_{\mathrm{RSS},\pi}$*,*

$$\hat{\mathbb{V}}(\hat{N}_{\mathrm{RSS},\pi}) = \sum_{i\in rss^*}\sum_{i'\in rss^*} \frac{\Delta^{(m)}_{ii':U}}{\pi^{(m)}_{ii':U}\pi^{(m)}_{i:U}\pi^{(m)}_{i':U}},$$

and

$$\widehat{\mathrm{Cov}}(\hat{t}_{y,\mathrm{RSS},\pi}, \hat{N}_{\mathrm{RSS},\pi}) = \sum_{i\in rss^*}\sum_{i'\in rss^*} \frac{\Delta^{(m)}_{ii':U}}{\pi^{(m)}_{ii':U}\pi^{(m)}_{i:U}\pi^{(m)}_{i':U}} y_i.$$

One could also use the results of Theorem 3.3 to develop similar results for the estimator

$$\tilde{t}_{y,\mathrm{RSS},m,\pi} = N\frac{\hat{t}_{y,\mathrm{RSS},m,\pi}}{\hat{N}_{\mathrm{RSS},m,\pi}}, \qquad \text{where} \qquad \hat{N}_{\mathrm{RSS},m,\pi} = \frac{1}{m}\sum_{r=1}^{m}\sum_{i\in rss_r^*}\frac{1}{\pi^{(m)}_{i:U}}.$$

The one disadvantage of these estimators is that they are biased, although it can be shown that (cf. Särndal et al. 1992)

$$\frac{[\mathbb{E}(\tilde{t}_{y,\mathrm{RSS}}) - t_y]^2}{\mathbb{V}(\tilde{t}_{y,\mathrm{RSS}})} \le \frac{\mathbb{V}(\hat{N}_{\mathrm{RSS}})}{N^2}.$$

Table 1 Summary statistics for the y-values in populations $U_1, \ldots, U_4$

	Min	Q_1	Median	Q_3	Max	$\bar{y}_U$	S^2_{yU}	$cv(y)$
U_1	66.50	92.97	99.74	106.77	130.01	99.83	102.64	0.10
U_2	1.85	68.23	121.70	198.35	681.82	147.77	11317.57	0.72
U_3	468.76	496.23	499.89	503.60	528.87	500.15	53.49	0.01
U_4	3,325.20	4,814.24	5,301.08	17,764.83	25,614.76	9,474.86	48,720,400.33	0.74

6 Numerical results and simulations

In this section, we first evaluate the performance of our proposed estimators for four simulated populations, each of size $N = 1,000$. We then apply our results to a real data set. In our simulation study: population U_1 consists of 1,000 realizations from a $N(\mu, \sigma)$ distribution with $\mu = 100$ and $\sigma = 10$; population U_2 consists of 1,000 realizations from a Gamma(α, λ) distribution with shape parameter $\alpha = 2$ and rate $\lambda = 1/75$; population U_3 is roughly symmetric and highly peaked and consists of 1,000 realizations from the Laplace density $f(y) = (1/10)\exp\{-|x - 500|/5\}$; and population U_4 is bimodal and right skewed, consisting of 700 realizations from a $N(5{,}000, 500)$ distribution and 300 realizations from a $N(20{,}000, 2{,}500)$ distribution. Table 1 shows the summary statistics for these populations.

6.1 Perfect ranking

We first present some results under the assumption of perfect ranking, (i.e. $y_{[i:U]} = y_{i:U}$). We report the design effect for each estimator, which we defined as

$$\text{deff}(\hat{t}) = \frac{\mathbb{V}(\hat{t})}{N^2(1 - n/N)S^2_{yU}/n},$$

where $n = \sum_U \pi_{i:U}$ is the expected number of unique elements in the RSSWR(m, k) design and the denominator is the variance of the usual π-estimator under the SRSWOR(n) design. Table 2 shows the design effects for the three proposed estimators (3.1), (3.6) and (4.1). The exact variances were calculated using (3.2), (3.7) and (4.2) respectively. For each of the populations $U_1, \ldots, U_4$, five different RSSWR(m, k) designs were analyzed. The values of m and k were such that $mk = 30$ and $k = 2, 3, 5, 10$ and 15. Note that the effect for the SRSWR(mk) design with $mk = 30$ and $N = 1,000$ is well known to be approximately 1.015 (see Särndal et al. 1992, Section 3.8.2).

Notice that the $\hat{t}_{y,\text{RSS},\pi}$ estimator does not perform well for populations U_1 and U_3. Both of these populations have fairly small coefficients of variation (0.10 and 0.02, respectively). Recall that the $\hat{t}_{y,\text{RSS},\pi} = (N/mk)\sum_{rss^*} y_{[i:U]}$, which is a sum over distinct elements in the ordered sample. When $cv(y)$ is small, all of the y's are of the same order of magnitude and hence any variation in the number of terms in the

Table 2 Design effects for perfect ranking with populations U_1, U_2, U_3 and U_4 with $mk = 30$ and $k = 2, 3, 5, 10$ and 15

Population	m	k	$\mathbb{E}(n_{rss^*})$	$\mathbb{V}(n_{rss^*})$	$\hat{t}_{y,\mathrm{RSS},\pi}$	$\hat{t}_{y,\mathrm{RSS},m,\pi}$	$\hat{t}_{y,\mathrm{RSS,pwr}}$	$\tilde{t}_{y,\mathrm{RSS}}$
U_1	15	2	29.57	0.41	2.072	0.723	0.690	0.685
	10	3	29.58	0.41	1.902	0.596	0.528	0.527
	6	5	29.58	0.40	1.722	0.509	0.363	0.366
	3	10	29.60	0.39	1.532	0.560	0.208	0.215
	2	15	29.61	0.38	1.443	0.714	0.147	0.155
U_2	15	2	29.57	0.41	0.746	0.725	0.725	0.718
	10	3	29.58	0.41	0.596	0.572	0.571	0.569
	6	5	29.58	0.40	0.436	0.410	0.407	0.410
	3	10	29.60	0.39	0.275	0.251	0.243	0.250
	2	15	29.61	0.38	0.208	0.189	0.176	0.183
U_3	15	2	29.57	0.41	67.560	2.304	0.722	0.716
	10	3	29.58	0.41	66.817	3.896	0.576	0.574
	6	5	29.58	0.40	65.692	7.453	0.422	0.424
	3	10	29.60	0.39	63.706	17.186	0.265	0.271
	2	15	29.61	0.38	62.207	27.385	0.197	0.204
U_4	15	2	29.57	0.41	0.792	0.774	0.773	0.765
	10	3	29.58	0.41	0.655	0.634	0.633	0.629
	6	5	29.58	0.40	0.501	0.477	0.475	0.475
	3	10	29.60	0.39	0.338	0.316	0.309	0.314
	2	15	29.61	0.38	0.268	0.250	0.238	0.244

summation causes significant instability in the estimator $\hat{t}_{y,\mathrm{RSS},\pi}$. When the coefficient of variation of the y's is moderate or large then, due to the symmetry of the inclusion probabilities, the duplicated elements in the ordered sample are just as likely to be very large or very small, hence mitigating the effect variations in the number of terms in the summation. The estimator $\hat{t}_{y,\mathrm{RSS},m,\pi}$ also suffers from this instability, but to a lesser extent, since duplication between cycles is no longer a factor. We also comment that π-estimation under the SRSWR(n) design also suffers from this instability. For the designs chosen here, the design effect is approximately $1 + 0.014/cv^2(y)$. The estimator $\hat{t}_{y,\mathrm{RSS},m,\pi}$ performs quite well for U_1, U_2 and U_4 but exhibits problems with U_3, and these problems increase as k increases. On the other hand, the estimator $\tilde{t}_{y,\mathrm{RSS}}$ defined in Theorem 5.2 performs very well for the examples considered here.

Notice that, for U_1, the design effect of the $\hat{t}_{y,\mathrm{RSS},m,\pi}$ estimator first decreases and then increases as k increases from 2 to 15. As mentioned, it is preferable to choose k as large as practically possible, however, for this design, choosing k too large results in more variation in the within cycle sample sizes, and hence reduces the efficiency. This effect is even more dramatic for population U_3 due to the very small coefficient of variation.

The estimators $\hat{t}_{y,\mathrm{RSS,pwr}}$ and $\tilde{t}_{y,\mathrm{RSS}}$ perform very well for all of the populations considered here and exhibit the largest (and most consistent) improvement in precision

over SRSWOR. Further simulation studies in Jafari Jozani and Johnson (2010) indicate that the sampling distributions of $\hat{t}_{y,\mathrm{RSS},\pi}$ and $\hat{t}_{y,\mathrm{RSS},\pi,m}$ can be far from normal (especially for populations U_2 and U_4). On the other hand, the sampling distributions of $\hat{t}_{y,\mathrm{RSS,pwr}}$ and $\tilde{t}_{y,\mathrm{RSS}}$, for these examples, are approximately normal, perhaps slightly right skewed for populations U_2 and U_4, and this is a definite advantage for approximate confidence intervals based on normal approximations.

6.2 Imperfect ranking

We now consider the case when the ranking of the y values is imperfect. For these results, we generate the x_i's according to the following model

$$x_i \mid y_i \sim N\left(\mu_x + \rho\frac{\sigma_x}{\sigma_y}(y_i - \bar{y}_U),\ \sigma_x\sqrt{1-\rho^2}\right),$$

where $\mu_x = 100, \sigma_x = 10, \rho \in \{0.3, 0.5, 0.7, 0.9\}$ and, for convenience, $\sigma_y = S_{yU}$. Note that, due to the symmetry of the inclusion probabilities, the auxiliary variable x may be negatively correlated with y.

Table 3 gives the finite population correlation coefficient $\rho(x, y)$ between the x and y values, the ratio $cv(x)/cv(y)$ and the design effects (over SRS) for the following estimators:

$\hat{t}_{y,\mathrm{pps}}$:	See Särndal et al. (1992)	$\hat{t}_{y,\pi,ra}$:	See Särndal et al. (1992)
$\hat{t}_{y,\mathrm{RSS},\pi}$:	Theorem 3.2	$\hat{t}_{y,\mathrm{RSS},\pi,ra}$:	see Theorem 5.1
$\hat{t}_{y,\mathrm{RSS},m,\pi}$:	Theorem 3.3	$\hat{t}_{y,\mathrm{RSS},m,\pi,ra}$:	See Theorem 5.1
$\hat{t}_{y,\mathrm{RSS,pwr}}$:	Theorem 4.3	$\hat{t}_{y,\mathrm{RSS,pwr},ra}$:	Theorem 5.1
$\tilde{t}_{y,\mathrm{RSS}}$:	Theorem 5.2		

We make the following observations regarding Table 3.

(1) It is well known (cf. Särndal et al. 1992) that, under the SRSWOR(n) design, the ratio estimator in (5.1) will be more efficient that the usual π-estimator provided that $\rho(x, y) > cv(x)/[2cv(y)]$. A similar dependence on $cv(x)/cv(y)$ also appears to hold for the RSSWR ratio estimators considered here. While a detailed theoretical analysis has not been attempted, the results in Theorem 4.3 and (5.7) suggest that, at least for $\hat{t}_{y,\mathrm{RSS,pwr},ra}$, the dependence is a function of both $\rho(x, y)$ and

$$\rho(\mu_{y[j]}, \mu_{x[j]}) = \frac{\sum_{j=1}^{k}(\mu_{y[j]} - \bar{y}_U)(\mu_{x[j]} - \bar{x}_U)}{\sqrt{\sum_{j=1}^{k}(\mu_{y[j]} - \bar{y}_U)^2 \sum_{j=1}^{k}(\mu_{x[j]} - \bar{x}_U)^2}}.$$

(2) The estimators $\hat{t}_{y,\mathrm{RSS,pwr}}$ and $\tilde{t}_{y,\mathrm{RSS}}$ again perform very well, especially when the correlation between y and x is reasonably high.

(3) Notice that, when ratio estimation is appropriate, the estimator $\hat{t}_{y,\mathrm{RSS},\pi,ra}$ performs very well here. The instability of $\hat{t}_{y,\mathrm{RSS},\pi}$ due to variations in the sample size is now mitigated by the fact that both $\hat{t}_{y,\mathrm{RSS},\pi,ra}$ and $\hat{t}_{x,\mathrm{RSS},\pi,ra}$ are based on the same sample size.

Table 3 Design effects in the RSSWR(10, 5) design with imperfect ranking

			Non-ratio estimators					Ratio estimators			
Pop.	$\rho(x, y)$	$cv(x)/cv(y)$	$\hat{t}_{y,\text{pps}}$	$\hat{t}_{y,\text{RSS},\pi}$	$\hat{t}_{y,\text{RSS},m,\pi}$	$\hat{t}_{y,\text{RSS,pwr}}$	$\tilde{t}_{y,\text{RSS}}$	$\hat{t}_{y,\pi,ra}$	$\hat{t}_{y,\text{RSS},\pi,ra}$	$\hat{t}_{y,\text{RSS},m,\pi,ra}$	$\hat{t}_{y,\text{RSS,pwr},ra}$
U_1	0.90	0.981	0.212	2.808	0.638	0.491	0.491	0.204	0.200	0.204	0.205
	0.69	0.983	0.638	3.018	0.858	0.712	0.702	0.611	0.560	0.571	0.572
	0.48	0.987	1.067	3.175	1.022	0.876	0.858	1.018	0.857	0.873	0.874
	0.28	0.991	1.494	3.270	1.122	0.976	0.954	1.424	1.099	1.116	1.118
U_2	0.90	0.139	0.725	0.574	0.532	0.529	0.528	0.770	0.433	0.436	0.436
	0.69	0.140	0.792	0.753	0.720	0.717	0.707	0.826	0.638	0.648	0.649
	0.49	0.140	0.864	0.896	0.869	0.868	0.850	0.882	0.804	0.821	0.822
	0.29	0.141	0.940	0.993	0.971	0.969	0.947	0.938	0.923	0.944	0.946
U_3	0.90	6.905	38.119	112.122	7.651	0.550	0.548	36.261	14.694	14.526	14.525
	0.70	6.946	41.964	112.298	7.835	0.734	0.723	39.489	15.204	14.984	14.983
	0.51	6.957	45.121	112.426	7.969	0.867	0.850	42.340	16.083	15.837	15.837
	0.31	6.952	47.847	112.517	8.064	0.961	0.940	44.966	17.171	16.914	16.914
U_4	0.90	0.137	0.735	0.575	0.535	0.533	0.532	0.772	0.437	0.439	0.440
	0.70	0.138	0.799	0.712	0.679	0.677	0.669	0.826	0.605	0.613	0.614
	0.51	0.138	0.866	0.875	0.849	0.847	0.831	0.879	0.786	0.802	0.803
	0.31	0.138	0.936	0.983	0.962	0.961	0.939	0.933	0.914	0.935	0.936

The column $\rho(x, y)$ is the finite population correlation coefficient between x and y. The design effect of the ratio estimator are calculated assuming the auxiliary variable z is equivalent to x

(4) We finally comment that, when ratio estimation is appropriate, the estimator $\hat{t}_{y,\mathrm{RSS,pwr}}$ performs quite well, although $\hat{t}_{y,\mathrm{RSS},\pi,ra}$ has a very slight advantage in some cases. However, recall that $\hat{t}_{y,\mathrm{RSS,pwr},ra}$ only requires knowledge of t_x and the relative rankings of the x_i in the samples. The estimators $\hat{t}_{y,\mathrm{RSS},\pi,ra}$ and $\hat{t}_{y,\mathrm{RSS},m,\pi,ra}$ require a-priori knowledge of the relative ordering of all of the x_i in order to calculate the inclusion probabilities.

6.3 Case study

We conclude the numerical results section with a case study. We apply the techniques described in this paper to a data set attributed to Platt et al. (1988). The original data set consists of measurements made on 399 longleaf pine (*Pinus palustris*) trees. The data set for the current example consists of the truncated version of the data set given in Chen et al. (2004), which contains the diameter (in centimeters) at breast height (x) and the height (in feet) (y) of 396 trees. For the purposes of this example, we take these 396 trees as the population U. Variations of this data set have been analyzed in many papers on ranked set sampling (see for example, Bhoj 2001; Deshpande et al. 2006). The idea is that, while the diameter at breast height is easy to measure, the height is usually more difficult to obtain.

For this data set, we have $\bar{y}_U = 52.67677$, $S^2_{yU} = 3261.6826$, $\bar{x}_U = 20.96414$, $S^2_{xU} = 310.3290$ and $\rho(x, y) = 0.9072982$ and, after adjusting the degrees of freedom for the variances to $N-1$ instead of N, these agree with those reported in Chen et al. (2004). Since the data had repeated x values, for the purpose of ranking, we took $x_i + u_i$ where the u_i were 396 realizations from a $N(0, 0.000001)$ distribution (re-centered so that $\bar{u} = 0$).

Table 4 shows the variances and design effects for the various estimators as well as the sample size that would be required to achieve the same variance under the SRSWOR design (n^*). Notice that $\hat{t}_{y,\mathrm{pps}}$ (the PPSWR(mk) estimator) performs extremely well in this example. Unfortunately, since we would seldom know the diameter at breast height for an entire population of trees, neither $\hat{t}_{y,\mathrm{pps}}$, nor the estimators based on the inclusion probabilities, are actually feasible in this example. On the other hand,

Table 4 Variances and design effects for various estimators for the trees data set

Non-ratio estimators				Ratio estimators			
Estimator	Variance	Deff.	n^*	Estimator	Variance	Deff.	n^*
$\hat{t}_{y,\mathrm{pps}}$	11.411	0.109	167	$\hat{t}_{y,\pi,ra}$	20.292	0.194	115
$\hat{t}_{y,\mathrm{RSS},\pi}$	72.079	0.690	41	$\hat{t}_{y,\mathrm{RSS},\pi,ra}$	19.740	0.189	117
$\hat{t}_{y,\mathrm{RSS},m,\pi}$	70.207	0.672	42	$\hat{t}_{y,\mathrm{RSS},m,\pi,ra}$	20.413	0.195	114
$\hat{t}_{y,\mathrm{RSS,pwr}}$	70.156	0.672	42	$\hat{t}_{y,\mathrm{RSS,pwr},ra}$	20.441	0.196	114
$\tilde{t}_{y,\mathrm{RSS}}$	68.908	0.660	43				

The RSS designs used $m = 10$ and $k = 3$. The column n^* shows the sample size required under the SRSWOR design to achieve the same variance of the considered estimators

the estimators $\hat{t}_{y,\mathrm{RSS,pwr}}$ and $\hat{t}_{y,\mathrm{RSS,pwr},ra}$ require only the width at breast height of the mk^2 inspected trees, and this is entirely feasible. Again, in this example, the estimator $\tilde{t}_{y,\mathrm{RSS}}$ performs better that the Horvitz–Thompson and Hansen–Hurwitz type estimators in estimating the population total (or mean).

We can also examine the efficiency from a cost perspective. Suppose the cost of ranking per unit is c_1 and the cost of obtaining a y measurement is c_2. Then the RSSWR(m, k) design will be more cost effective if $c_1 mk^2 + c_2 mk < c_2 n^*$ or $c_1 < c_2(n^* - mk)/mk^2$, where n^* is the sample size required under the SRSWOR design to achieve the same variance. In our example, the RSSWR$(10, 3)$ design, with the estimator $\hat{t}_{y,\mathrm{pwr}}$, would be more cost effective if the cost of obtaining a y measurement is more than about 8 times the cost of ranking per unit.

We also comment that Chen et al. (2004) report a simulated design effect of about 0.27 for this example based on 20 repeated samples. In experiments, we were able to obtain simulated design effects of 0.27 or less (based on 20 replications) for $\hat{t}_{y,\mathrm{RSS,pwr}}$ about 4% of the time, whereas the average was very close to the stated exact design effect of 0.672.

7 Concluding remarks

In this paper, we studied the use of a variation of the RSS design in finite populations. We derived the first and second-order inclusion probabilities of population elements for this RSS design. These inclusion probabilities are very important. They give an insight on how the RSS design has more control on which element enters the sample as compared to SRS with or without replacement. They can also be used to construct different designed-based π-estimators of the population parameters. To address these issues, we proposed a number of designed-based estimators of the population mean (or total) using the obtained inclusion probabilities. These estimators have been shown to be more efficient than the usual estimators under SRSWOR designs of comparable sample sizes. However, the performance of the Horvitz–Thompson type estimators and their associated ratio estimators based on the RSS design compared to their counterparts under SRSWOR design depends highly on the value of $cv(y)$, the coefficient of variation of the variable of interest in the underlying population. The consistent performance of the estimators $\hat{t}_{y,\mathrm{RSS,pwr}}$, $\hat{t}_{y,\mathrm{RSS,pwr},ra}$ and $\tilde{t}_{y,\mathrm{RSS}}$ suggest that these are to be preferred in most situations. But, according to our simulation study, if $cv(y)$ is not too small then the Horvitz–Thompson type estimators and their associated ratio estimators can perform extremely well compared to the usual estimator under the SRSWOR design. In this situation the Horvitz–Thompson type estimators perform as good as the Hansen–Hurwitz type estimators for almost all considered populations. We also examined the efficiency of our proposed estimators from a cost perspective. In addition, we examined the impact of imperfect ranking on the performance of our proposed estimators by considering ranking according to an auxiliary variable with different values of the correlation coefficient with the variable of interest. Our simulation study suggests that, for auxiliary variables that are highly correlated with the variable of interest, the design effect of our proposed estimators compared to that of SRSWOR are less than 1. The results of this paper indicate that ranked set sampling

should be seriously considered as a technique for improving the precision of estimates when sampling from a finite population as it can increase precision considerably for little extra effort. Note that the results of this paper can easily be adapted to difference and regression estimators as well as more complex designs such as stratified or multi-stage sampling designs.

Acknowledgments Both authors were partially supported by the Natural Sciences and Engineering Research Council of Canada. The authors would also like to thank Alexandre Leblanc for providing comments on an earlier version of this manuscript and the anonymous referees for their comments and suggestions.

References

Al-Saleh MF, Samawi HM (2007) A note on inclusion probability in ranked set sampling and some of its variations. Test 16:198–209

Barabesi L, Marcheselli M (2004) Design-based ranked set sampling using auxiliary variables. Environ Ecol Stat 11:415–430

Bhoj DS (2001) Ranked set sampling with unequal samples. Biometrics 57:957–962

Bouza CN (2001) Model-assisted ranked survey sampling. Biom J 43:249–259

Bouza CN (2002a) Estimation of the mean in ranked set sampling with non responses. Metrika 56:171–179

Bouza CN (2002b) Ranked set sub-sampling the non-response strata for estimating the difference of means. Biom J 44:903–915

Bouza CN (2009) Ranked set sampling and randomized response procedures for estimating the mean of a sensitive quantitative character. Metrika 70:267–277

Chen Z, Bai Z, Sinha B (2004) Ranked set sampling: theory and applications. Lecture Notes in Statistics. Springer, New York

Deshpande JV, Frey J, Ozturk O (2006) Nonparametric ranked-set sampling confidence intervals for quantiles of a finite population. Environ Ecol Stat 13:25–40

Jafari Jozani M, Johnson BC (2010) Ranked set sampling estimation in finite populations. Technical Report 2010-001, University of Manitoba, Department of Statistics

Kaur A, Patil G, Sinha A, Taillie C (1995) Ranked set sampling: an annotated bibliography. Environ Ecol Stat 2:25–54

McIntyre GA (1952) A method for unbiased selective sampling using ranked sets. Aust J Agric Res 3:385–390

Özdemir YA, Gökpınar F (2007) A generalized formula for inclusion probabilities in ranked set sampling. Hacet J Math Stat 36:89–99

Ozturk O, Bilgin OC, Wolfe DA (2005) Estimation of population mean and variance in flock management: a ranked set sampling approach in a finite population setting. J Stat Comput Simul 75:905–919

Patil GP, Sinha AK, Taillie C (1995) Finite population corrections for ranked set sampling. Ann Inst Stat Math 47:621–636

Platt WJ, Evans GW, Rathbun SL (1988) The population-dynamics of a long lived conifer (Pinus palustris). Am Nat 131:491–525

Särndal CE, Swensson B, Wretman J (1992) Model assisted survey sampling. Springer Series in Statistics. Springer, New York

Takahasi K, Futatsuya M (1988) Ranked set sampling from a finite population. Proc Inst Stat Math 36:55–68

Takahasi K, Futatsuya M (1998) Dependence between order statistics in samples from finite population and its application to ranked set sampling. Ann Inst Stat Math 50:49–70

Author Biographies

Mohammad Jafari Jozani joined the Department of Statistics at the University of Manitoba as an Assistant Professor in 2009. He completed all his studies in Iran. He obtained a B.Sc. degree in Statistics from Allameh Tabatabaie University, and M.Sc. and Ph.D. degrees in Mathematical Statistics from Shahid

Beheshti University, Tehran, Iran. Mohammad's research has primarily been in Bayesian statistics, statistical decision theory, and estimation in restricted parameter spaces. Recently he has begun to do research in more applied areas such as ranked set sampling, credibility theory and image processing.

Brad C. Johnson joined the Department of Statistics at the University of Manitoba as an Assistant Professor in 2005. Brad received his Ph.D. in Statistics from Purdue University. He also holds Master's of Science degrees from Purdue University and the University of Manitoba, and a Bachelor of Science degree in Statistics from the University of Manitoba. His research interests include applied probability, Markov chains and survey sampling.