

Proportion estimation in ranked set sampling in the presence of tie information

Ehsan Zamanzade¹  · Xinlei Wang²

Received: 13 July 2017 / Accepted: 12 March 2018 / Published online: 23 March 2018
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract Ranked set sampling (RSS) is a statistical technique that uses auxiliary ranking information of unmeasured sample units in an attempt to select a more representative sample that provides better estimation of population parameters than simple random sampling. However, the use of RSS can be hampered by the fact that a complete ranking of units in each set must be specified when implementing RSS. Recently, to allow ties declared as needed, Frey (*Environ Ecol Stat* 19(3):309–326, 2012) proposed a modification of RSS, which is to simply break ties at random so that a standard ranked set sample is obtained, and meanwhile record the tie structure for use in estimation. Under this RSS variation, several mean estimators were developed and their performance was compared via simulation, with focus on continuous outcome variables. We extend the work of Frey (2012) to binary outcomes and investigate three nonparametric and three likelihood-based proportion estimators (with/without utilizing tie information), among which four are directly extended from existing estimators and the other two are novel. Under different tie-generating mechanisms, we compare the performance of these estimators and draw conclusions based on both simulation and a data example about breast cancer prevalence. Suggestions are made about the choice of the proportion estimator in general.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s00180-018-0807-x>) contains supplementary material, which is available to authorized users.

✉ Ehsan Zamanzade
e.zamanzade@sci.ui.ac.ir; ehsanzamanzadeh@yahoo.com

Xinlei Wang
swang@smu.edu

¹ Department of Statistics, University of Isfahan, Isfahan 81746-73441, Iran

² Department of Statistical Science, Southern Methodist University, Dallas, TX 75275, USA

Keywords Imperfect ranking · Isotonic estimation · Maximum likelihood · Nonparametric estimation · Ranking tie · Relative efficiency

1 Introduction

Ranked set sampling (RSS) is a cost-efficient sampling strategy that can be used to provide a more representative sample than simple random sampling (SRS). RSS has been found useful in situations where measuring sample units is expensive, time-consuming or difficult but a small set of units can be ranked using methods such as eye inspection, personal judgment or use of a concomitant variable, which do not require formal quantification of the units. First introduced by McIntyre (1952) for estimating the mean pasture yield in Australia, RSS has been widely used in many other fields including forestry (Halls and Dell 1966), medicine (Chen et al. 2005, 2007), biometrics (Mahdizadeh and Zamanzade 2017), environmental monitoring (Nussbaum and Sinha 1997; Kvam 2003; Ozturk et al. 2005), entomology (Howard et al. 1982) and educational studies (Wang et al. 2016).

To draw a (balanced) ranked set sample using a set size m , one first draws a simple random sample of size m^2 from the population of interest. He then divides the sample into m sets of size m randomly. Each set of m units is ranked from smallest to largest without measuring values of these units. From the first set, the unit with rank 1 is selected for actual measurement; from the second set, the unit with rank 2 is selected for measurement and so on. The whole process is repeated n times (cycles) to form a ranked set sample of size $N = m \times n$. The set size m is usually selected to be a small number (e.g. 2–10), to avoid poor-quality ranking.

RSS requires a ranker to provide a unique rank to each unit when ranking a set, and so there is no tie allowed. However, situations when the ranker is not sure about how to rank two or more observations arise frequently in practice. To mitigate this difficulty, Frey (2012) proposed a slightly different version of the RSS scheme, which is denoted by RSS-t in this paper. RSS-t allows the ranker to declare ties as much as he wishes. Further, the ranker breaks the ties at random, but records the tie structure to be used in the estimation process. Frey (2012) then developed several nonparametric estimators of the population mean and discussed two models for rankings with ties: Discrete Perceived Size (DPS) and Tied-If-Close (TIC). He showed via simulation based on the DPS model that using the tie information would improve the estimation efficiency. All these have been done with focus on continuous variables. Although the estimators in Frey (2012) can be applied to binary outcomes, the use of RSS-t for estimating the population proportion p has not been systematically investigated yet.

In the past, the problem of proportion estimation based on a ranked set sample has received considerable attention. Terpstra (2004) and Terpstra and Liudahl (2004) first showed that RSS proportion estimators are more efficient than their SRS counterparts in cases of judgment ranking and ranking via a concomitant variable, respectively. Terpstra and Wang (2008) developed several methods to construct confidence bounds for the population proportion based on RSS. Chen et al. (2005) further proposed to aid the ranking of a binary variable of interest by fitting a logistic regression model, and Chen et al. (2007) investigated the application of RSS, combined with logistic

regression for ranking, to estimation of disease prevalence. Hatefi and Jafari Jozani (2017) considered the problem of estimating malignant breast cancer prevalence using partially rank ordered set samples based on a different strategy for handling ties when more than one concomitant variable is available.

We consider the use of RSS-t with binary outcomes and investigate six estimators of the population proportion p . Throughout this paper, we restrict our attention to balanced RSS when using RSS-t. These estimators can be classified in two ways: (i) two estimators ignoring the tie information and the other four exploiting such information; and (ii) three nonparametric estimators and three likelihood-based estimators. Here, we address main research questions including whether using RSS-t over SRS is beneficial for binary responses, whether utilizing tie information can improve proportion estimation and if yes, which estimator(s) to use in different scenarios. To do so, we compare the performance of the different estimators via simulation under both TIC and DPS models for generating ties. We also present an empirical study using a breast cancer data set, where ranking is done through ordinal concomitant variables and so ties naturally occur.

2 An illustrative example of RSS-t

Let Y be a binary random variable that follows a Bernoulli distribution with probability of success p . To estimate p , suppose we obtain a RSS-t sample with the total sample size $N = mn$, where m is the set size and n is the number of cycles. The sampling scheme of RSS-t is essentially the same as that of RSS, as described in the introduction, except for breaking ties at random when deciding which units to select for actual quantification and meanwhile recording the tie structure. That is, in the i th set of the j th cycle, if there are more than one observation with (judgement) rank i , the ranker randomly select one to measure (say $Y_{[i]j}$) and use an $m \times m$ matrix, say \mathbf{T}^j , to record how the units are tied:

$$\mathbf{T}^j = \begin{bmatrix} I_{1,1}^j & \cdots & I_{1,m}^j \\ \vdots & \ddots & \vdots \\ I_{m,1}^j & \cdots & I_{m,m}^j \end{bmatrix},$$

where $I_{i,k}^j$ is an indicator variable that is one if the unit with rank i is tied for rank k in the j th cycle and zero otherwise ($i = 1, \dots, m, k = 1, \dots, m$). Note that $\sum_{k=1}^m I_{i,k}^j \geq 1$ and “=” occurs if rank i is assigned only to one unit in the set.

We illustrate the sampling scheme by a hypothetical example. Suppose that we are interested in estimating the prevalence of breast cancer in a given population of adults. To determine if one suffers from breast cancer, a comprehensive biopsy procedure is required, which is expensive and inconvenient, especially in some developing countries. However, a medical researcher can simply rank a small set of patients according to their probability of having breast cancer. Let Y be one (*success*) if the test subject suffers from breast cancer and zero (*failure*) otherwise.

To draw a ranked set sample of size $N = 10$, using set size $m = 5$, the researcher draws $m^2 \times n = 50$ subjects from the given population and divide them into 10 sets

Table 1 An example of RSS-t with $N = 10$ and $m = 5$

Cycle	Set	Ranked units	Breast cancer status	Y
1	1	$u_{1,1}^1 \& u_{2,1}^1 \& u_{3,1}^1, u_{4,1}^1, u_{5,1}^1$	No	0
	2	$u_{1,2}^1, \mathbf{u}_{2,2}^1, u_{3,2}^1 \& u_{4,2}^1 \& u_{5,2}^1$	No	0
	3	$u_{1,3}^1, u_{2,3}^1, \mathbf{u}_{3,3}^1, u_{4,3}^1 \& u_{5,3}^1$	Yes	1
	4	$u_{1,4}^1, u_{2,4}^1, u_{3,4}^1 \& \mathbf{u}_{4,4}^1, u_{5,4}^1$	No	0
	5	$u_{1,5}^1, u_{2,5}^1, u_{3,5}^1, u_{4,5}^1, \mathbf{u}_{5,5}^1$	Yes	1
2	1	$\mathbf{u}_{1,1}^2, u_{2,1}^2, u_{3,1}^2, u_{4,1}^2, u_{5,1}^2$	No	0
	2	$u_{1,2}^2, \mathbf{u}_{2,2}^2, u_{3,2}^2, u_{4,2}^2, u_{5,2}^2$	Yes	1
	3	$u_{1,3}^2, u_{2,3}^2 \& \mathbf{u}_{3,3}^2 \& u_{4,3}^2, u_{5,3}^2$	Yes	1
	4	$u_{1,4}^2 \& u_{2,4}^2 \& u_{3,4}^2 \& \mathbf{u}_{4,4}^2, u_{5,4}^2$	No	0
	5	$u_{1,5}^2, u_{2,5}^2 \& u_{3,5}^2, u_{4,5}^2, \mathbf{u}_{5,5}^2$	Yes	1

The sample units in each set are listed in an increasing order of perceived probability of having breast cancer and the ties are connected with ampersands

of size 5. In each set, he ranks the subjects according to their perceived probabilities of having breast cancer. This can be done by examining superficial lumps or masses using the fine needle aspiration biopsy technique, which is much faster and cheaper than the comprehensive biopsy procedure. Whenever the researcher cannot determine the exact ranks of two or more subjects in the set, he is allowed to declare ties, and then selects one at random for formal testing. An example of the RSS-t scheme is detailed in Table 1, where each row corresponds to one set, and sample units in each set are listed based on their judgement probability of having breast cancer after the possible ties are broken at random. Let $u_{i,k}^j$ be the i th judgement unit of the k th set in the j th cycle ($i, k \in \{1, \dots, 5\}$ and $j = 1, 2$). In this table, we connect the units that are declared tied with ampersands, and the unit selected for actual quantification with bold face.

Table 1 shows that in the first set of the first cycle, the ranker is able to distinguish the top two subjects but cannot distinguish among the bottom three. Therefore, he selects one of the three at random for measurement and also records this tie structure for potential use in the estimation process. In the second set of the first cycle, the top three subjects are tied. However, since in the second set, the researcher’s interest is to identify the subject with rank 2, this tie structure is irrelevant and so not recorded. Thus, the matrices that contain tie information are specified as follows.

$$\mathbf{T}^1 = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{T}^2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

3 Proportion estimation based on RSS-t

For RSS-t, the full data include not only $\{Y_{[i]j}, i = 1, \dots, m, j = 1, \dots, n\}$, where $Y_{[i]j}$ is the precisely measured value for the unit with (judgement) rank i in the j th cycle after randomly breaking ties, but also $\mathbf{T}^1, \dots, \mathbf{T}^n$, matrices of tie information.

3.1 Nonparametric estimators

Frey (2012) considered six nonparametric mean estimators, $\hat{\mu}_1 - \hat{\mu}_6$, for use with RSS-t: $\hat{\mu}_1$ is the standard RSS mean estimator that ignores the tie information; $\hat{\mu}_2$ is the estimator that uses a strategy proposed in MacEachern et al. (2004) to split each tied unit among the strata corresponding to the ranks for which the unit was tied; $\hat{\mu}_3$ and $\hat{\mu}_4$ are the isotonized versions of $\hat{\mu}_1$ and $\hat{\mu}_2$, respectively; and $\hat{\mu}_5$ and $\hat{\mu}_6$ are the Rao-Blackwellized (RB) versions of $\hat{\mu}_1$ and $\hat{\mu}_3$ (the RB versions of $\hat{\mu}_2$ and $\hat{\mu}_4$ do not lead to new estimators.) Since we focus on balanced RSS-t where $\hat{\mu}_1 = \hat{\mu}_3 = \hat{\mu}_5 = \hat{\mu}_6$ holds, we only need to consider $\hat{\mu}_1, \hat{\mu}_2$, and $\hat{\mu}_4$ for estimating the population proportion when Y is binary. The resulting proportion estimators, denoted by $\hat{p}_{st}, \hat{p}_{sp}$ and \hat{p}_{iso} , are essentially the same as $\hat{\mu}_1, \hat{\mu}_2$, and $\hat{\mu}_4$, respectively, which are discussed below for completeness.

The standard RSS proportion estimator that simply ignores the matrices of tie structures is given by

$$\hat{p}_{st} = \frac{1}{m} \sum_{i=1}^m \hat{p}_{[i]},$$

where $\hat{p}_{[i]} = \sum_{j=1}^n Y_{[i]j} / n$ is the sample proportion within the i th rank stratum.

Our second proportion estimator \hat{p}_{sp} for RSS-t, which corresponds to $\hat{\mu}_2$, is given by

$$\hat{p}_{sp} = \frac{1}{m} \sum_{i=1}^m \hat{p}_{[i],sp},$$

with

$$\hat{p}_{[i],sp} = \frac{\sum_{l=1}^m \sum_{j=1}^n Y_{[l]j} \times w_{l,i}^j}{\sum_{l=1}^m \sum_{j=1}^n w_{l,i}^j}, \tag{1}$$

and

$$w_{l,i}^j = \frac{I_{l,i}^j}{\sum_{k=1}^m I_{l,k}^j},$$

where $I_{l,k}^j$ is the element in the l th row and k th column of \mathbf{T}^j . This estimator incorporates tie information through a splitting strategy; that is, if $Y_{[i]j}$ is tied for r distinct

ranks, we then assign it to the corresponding r rank strata with equally split weights $1/r$. For example, if the measured unit with rank 2 is tied for ranks 2, 3 and 4, we split the Y value among the three strata with ranks 2, 3 and 4 with equal weights $1/3$. Our third proportion estimator \hat{p}_{iso} for RSS-t is the isotonized version of \hat{p}_{sp} , which corresponds to $\hat{\mu}_4$. Let $p_{[i]} = E(Y_{[i]j})$ be the probability of success for sample units with rank i . It is often reasonable to assume that the probability of the i th (judgement) order statistic satisfies the following order constraint, i.e.

$$p_{[1]} \leq \dots \leq p_{[m]}. \tag{2}$$

However, the estimates $\hat{p}_{[i],sp}$ may violate constraint (2) due to sampling variability. Let $n_i = \sum_{l=1}^m \sum_{j=1}^n w_{l,i}^j$ ($i = 1, \dots, m$) be the total weighted sample size of the i th rank stratum. To isotonize $\{\hat{p}_{[i],sp}\}_{i=1}^m$ so that constraint (2) is imposed, we define $n_{rs} = \sum_{g=r}^s n_g$ and

$$\hat{p}_{[i],iso} = \min_{r \leq i} \max_{s \geq i} \sum_{g=r}^s \frac{n_g \hat{p}_{[g],sp}}{n_{rs}},$$

where $\{\hat{p}_{[i],iso}\}_{i=1}^m$ are known as isotonic regression estimators of $\{p_{[i]}\}_{i=1}^m$, which minimize the weighted least square $\sum_{i=1}^m n_i (p_{[i]} - \hat{p}_{[i],sp})^2$ under constraint (2). Then the isotonized proportion estimator is given by

$$\hat{p}_{iso} = \frac{1}{m} \sum_{i=1}^m \hat{p}_{[i],iso}. \tag{3}$$

We would expect that \hat{p}_{iso} improves \hat{p}_{sp} for small N ; but as $N \rightarrow +\infty$, the chance that (2) is violated approaches zero, and so they become identical.

3.2 Likelihood-based estimators

Terpstra (2004) proposed the maximum likelihood (ML) estimator based on (balanced) RSS data for binary outcomes, which has been found to be more efficient than the standard estimator \hat{p}_{st} asymptotically as well as in certain finite-sample simulation settings. Let $\{Y_{(i)j}; i = 1, \dots, m, j = 1, \dots, n\}$ be a ranked set sample in which there is no ranking error. Then, the log likelihood function can be expressed by

$$L(p) = \sum_{i=1}^m \sum_{j=1}^n \{Y_{(i)j} \log(p_{(i)}) + (1 - Y_{(i)j}) \log(1 - p_{(i)})\},$$

where $p_{(i)} \equiv B_{m+1-i,i}(p)$ is the probability of success for sample units in the i th rank stratum. Here, $B_{m+1-i,i}(p)$ is the cumulative distribution function (CDF) of the beta distribution with parameters $m + 1 - i$ and i , evaluated at the point p , namely

$$B_{m+1-i,i}(p) = \int_0^p (m + 1 - i) \binom{m}{m + 1 - i} t^{m-i} (1 - t)^{i-1} dt.$$

Thus, the ML estimator of p based on RSS is given by $\hat{p}_{ml} = \arg \max_{p \in [0,1]} L(p)$.

Since data from RSS-t constitutes a ranked-set sample, \hat{p}_{ml} can still be used if we ignore the tie information. Next, we propose two new proportion estimators, $\hat{p}_{t.ml}$ and $\hat{p}_{m.ml}$, to incorporate tie information using likelihood-based approaches.

Assume that $Y_{(i)j}$ is tied for two or more units in the set of size m . Then $Y_{(i)j}$ follows a Bernoulli distribution with probability of success $p_{(i),t}$, where

$$p_{(i),t} = \frac{\sum_{k=1}^m I_{i,k}^j B_{m+1-k,k}(p)}{\sum_{k=1}^m I_{i,k}^j},$$

is a mixture of beta CDFs, evaluated at the point p . Then, the log-likelihood function of $\{Y_{(i)j}\}$ that incorporates the tie information can be written as

$$L_t(p) = \sum_{i=1}^m \sum_{j=1}^n \{Y_{(i)j} \log(p_{(i),t}) + (1 - Y_{(i)j}) \log(1 - p_{(i),t})\}.$$

Thus, the ML estimator of p based on RSS-t can be obtained by $\hat{p}_{t.ml} = \arg \max_{p \in [0,1]} L_t(p)$, where the subscript “t” stands for tie.

Another approach to incorporate tie information is to define a pseudo log-likelihood function $L^*(p)$ based on the splitting strategy,

$$L^*(p) = \sum_{i=1}^m \{n_i \hat{p}_{(i),sp} \log(p_{(i)}) + n_i (1 - \hat{p}_{(i),sp}) \log(1 - p_{(i)})\},$$

where $\hat{p}_{(i),sp}$ is defined in (1), but with “[]” replaced by “()” in the subscript of p to indicate the requirement of perfect ranking. The corresponding ML estimator is given by $\hat{p}_{m.ml} = \arg \max_{p \in [0,1]} L^*(p)$. Here, the subscript “m” indicates a mixed strategy used.

In the past, MacEachern et al. (2004) and Frey (2012) found the splitting strategy works well in dealing with ties for continuous outcomes. Also, as mentioned before, ML estimation may work well for binary outcomes. Thus, our pseudo likelihood method attempts to combine the strength from the splitting strategy and ML estimation for binary data with ties.

As shown above, the development of the three likelihood-based estimators, \hat{p}_{ml} , $\hat{p}_{t.ml}$, and $\hat{p}_{m.ml}$, requires the assumption of perfect ranking. So it is important to examine the robustness of these estimators in the presence of ranking errors. Finally, we mention that all these ML estimators are well defined and can be solved easily using standard optimization procedures. This can be seen from the following facts; (i) since $B_{m+1-i,i}(p)$ is a log concave function in p , $L(p)$ and $L^*(p)$ are both concave in p ; and (ii) from Theorem 2 in Mu (2015), we can conclude that $p_{(i),t}$ is strictly log-concave in p and so $L_t(p)$ is strictly concave in p as well.

4 Comparison of proportion estimators

We first compare the six estimators using the example given in Sect. 2. Based on the tie structure and data in Table 1, the three nonparametric estimators produce $\hat{p}_{st} = 0.5$, $\hat{p}_{sp} = 0.463$, and $\hat{p}_{iso} = 0.481$; and the three likelihood-based estimators produce $\hat{p}_{ml} = 0.492$, $\hat{p}_{t.ml} = 0.552$ and $\hat{p}_{m.ml} = 0.531$. Recall that \hat{p}_{st} and \hat{p}_{ml} ignore the tie information among the six. It is interesting to observe that, their values are pretty close but the other four give quite different estimates. In this example, \hat{p}_{st} is larger than the other two of its same kind but \hat{p}_{ml} is smaller. Evidently, this example illustrates that incorporating tie information into estimation can make a noticeable difference, no matter which direction it is in, for either estimation technique.

In what follows, we present a simulation study to formally compare the performance of the six different proportion estimators from Sect. 3, where we consider varying ranking quality and different tie-generating models, as well as different design parameters.

4.1 Simulation setups

Suppose $Y \sim \text{Bernoulli}(p)$ is the variable of interest; and X is a continuous variable which can be measured at a negligible cost, satisfying $X|Y = y \sim N(\mu_y, 1)$ for $y = 0, 1$ with $\mu_0 \equiv 0$. Let ρ denote the correlation between X and Y . Then it can be verified that

$$\mu_1 = \frac{\rho}{\sqrt{(1 - \rho^2) p (1 - p)}}. \quad (4)$$

Note that the relationship between X and Y can be modeled via a logistic regression model so that X_i reflects the probability of success of the unit i . In our simulation, X is used for ranking Y and so the correlation ρ is used to measure the ranking quality. We set ρ to 0.9, 0.7, 0.5 and 0.1, representing good, fair, poor and nearly random ranking, respectively. Due to the symmetry considerations of all the six estimators, we set $p = 0.1, 0.2, \dots, 0.5$ and the results for $p = 0.6, 0.7, \dots, 0.9$ would be the same as those corresponding to $1 - p$ except for Monte Carlo errors. For each (ρ, p) combination, we solve for the conditional mean μ_1 using (4).

Frey (2012) described two classes of models for ties in rankings: DPS and TIC. The DPS model involves discretizing X , which can be done by $X^* = \lfloor X/c \rfloor$, where $\lfloor x \rfloor$ is the largest integer less than or equal to x . Rankings are then based on X^* , with units that have the same X^* value being declared tied. The TIC model declares the i th and j th units to be tied whenever $|X_i - X_j| < c$. Due to the transitivity of the TIC model, the i th and j th units may be still declared as ties even if $|X_i - X_j| \geq c$ provided that there are other units to bridge the gap. In either model, $c > 0$ is a user-chosen model parameter. As mentioned in Frey (2012), models in each class can exhibit certain undesirable behavior when the model parameter c and the set size m are modified. For TIC models, adding a unit to a set can increase the number of ties among the units already in the set. For DPS models, we would expect that increasing c leads to more ties, but this is not necessarily true. Frey (2012) also discussed other

differences between the two classes. Thus, it would be interesting to investigate the potential impact of the different tie-generating mechanisms on the relative performance of the estimators. However, Frey (2012) only evaluated the mean estimators under DPS models. Here, we evaluate the performance of our proportion estimators under both classes of models, where we set $c \in \{0.5, 1, 2, 4\}$ for DPS models as in Frey (2012) and $c \in \{0.5, 1, 1.5, 2\}$ for TIC models.

For the RSS design parameters, we set the total sample size $N \in \{15, 30, 90, 180\}$ and the set size $m \in \{3, 5\}$. For each (N, m, p, ρ, c) combination under DPS/TIC, we generate 100,000 RSS-t samples, and estimate the mean square error (MSE) for each proportion estimator. We define the relative efficiency (RE) of a proportion estimator (say \hat{p}) as the ratio of the MSE of the SRS proportion estimator \hat{p}_{srs} vs. the MSE of \hat{p} , given by

$$RE(p) = \frac{p(1-p)/N}{MSE(\hat{p})},$$

where a RE value larger than 1 indicates \hat{p} is more efficient than \hat{p}_{srs} .

4.2 Simulation results

Here, we only report simulated REs for settings with $N = 30$ in Figs. 1 and 2 for DPS models and in Figs. 3 and 4 for TIC models. This is because unless N is very small (e.g., RSS-t is implemented with only one cycle), N seems to have not much impact on the relative performance of the six estimators (see results for settings with $N \in \{15, 90, 180\}$ in Figures S1–S6 for DPS models and in Figures S9–S14 for TIC models in the Supplementary Material; for readers' reference, results for one-cycle settings with $N = m \in \{3, 5\}$ are reported as well in Figures S7–S8 for DPS models and in Figures S15–S16 for TIC models). We further mention that, although the RE of \hat{p}_{iso} is always higher than that of \hat{p}_{sp} in all the cases considered, their performance is so close that the two curves overlap in all these figures and cannot be distinguished from each other. Thus, we simply omit \hat{p}_{sp} in our discussion below.

Results under DPS models

From comparing Fig. 1 with Fig. 2, we can see that as the set size m increases, RE generally increases if it is larger than 1 but decreases if it is smaller than 1 for all the six estimators.

The relative performance of the six estimators varies as p varies, and we distinguish two cases: (i) p is close to 0.5; that is, $p \in (\delta, 1 - \delta)$, where δ is a number in the range $(0, 0.5)$, but is likely to be around 0.25; and (ii) p is close to 0 (or 1), i.e., $p \in (0, \delta]$ or $p \in [1 - \delta, 1)$. When p is close to 0.5, $\hat{p}_{m.ml}$ is usually the best except for cases with nearly random ranking (i.e. $\rho = 0.1$), where \hat{p}_{ml} often outperforms $\hat{p}_{m.ml}$, and both are better than the other estimators. When p is close to 0 (or 1), $\hat{p}_{m.ml}$ seems to the best for good-quality ranking (i.e. $\rho = 0.9$); but in general, the nonparametric estimator \hat{p}_{iso} is preferred. This is because the performance of \hat{p}_{iso} is the best or close

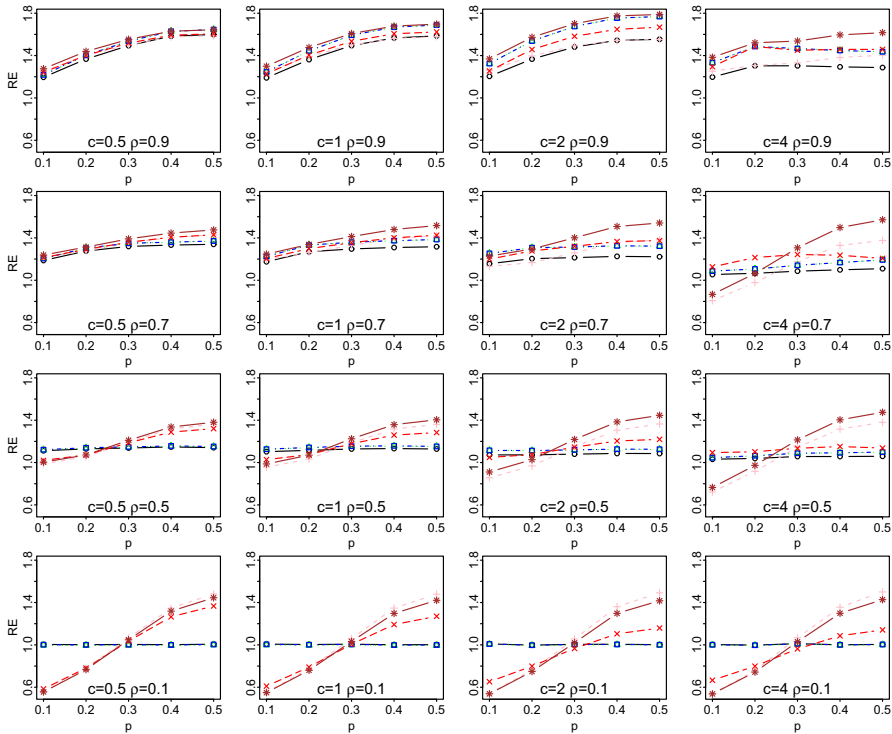


Fig. 1 Simulation: comparing relative efficiency of \hat{p}_{st} (represented by \circ), \hat{p}_{sp} (represented by Δ), \hat{p}_{iso} (represented by \square), \hat{p}_{ml} (represented by $+$), $\hat{p}_{t.ml}$ (represented by \times) and $\hat{p}_{m.ml}$ (represented by $*$) under DPS models. This figure appears in color in the electronic version of this paper

to the best in nearly all the settings. As the ranking quality ρ decreases or the sample size N increases, δ tends to increase, meaning that the central interval where $\hat{p}_{m.ml}$ or \hat{p}_{ml} works best becomes narrower.

We note that among 640 simulation scenarios under DPS, there are only 12 scenarios where the maximum RE among the six estimators falls below 1. A closer examination reveals that all the 12 scenarios are for nearly random ranking, and the lowest RE value of the best estimator among these scenarios is 0.993. This seems to suggest that as long as we choose an appropriate estimator to use, using RSS-t over SRS would not incur loss of estimation efficiency even in situations when it is not helpful. As to whether to utilize the tie information when estimating p from RSS-t data, the answer is clearly yes as long as ranking is better than random guessing.

Results under TIC models

From comparing Fig. 3 with Fig. 4, we find that the impact of the set size m on RE is not as clear as in DPS models. Although in many cases, increasing m increases

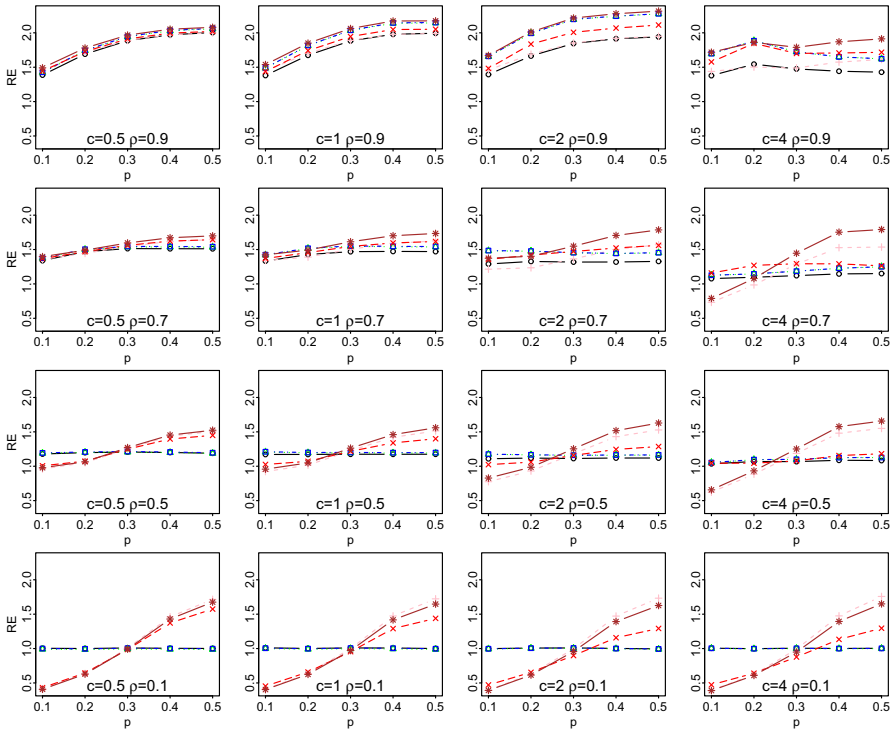


Fig. 2 Simulation: comparing relative efficiency of \hat{p}_{st} (represented by \circ), \hat{p}_{sp} (represented by Δ), \hat{p}_{iso} (represented by \square), \hat{p}_{ml} (represented by $+$), $\hat{p}_{t.ml}$ (represented by \times) and $\hat{p}_{m.ml}$ (represented by $*$) to \hat{p}_{SRS} as a function of population proportion p , for various (c, ρ) settings with $(N, m) = (30, 5)$ under DPS models. This figure appears in color in the electronic version of this paper

RE when $RE > 1$, this is obviously not true for some estimators. For example, when $p = 0.1$ and 0.2 , the RE values of $\hat{p}_{m.ml}$ are above 1.2 for the case of $N = 30, m = 3, c = 1.5$ and $\rho = 0.7$; but when m is increased to 5 while fixing the other parameters, the RE of $\hat{p}_{m.ml}$ decreases to values slightly above 1. Similarly, the RE values of \hat{p}_{ml} jump from above 1 to below 1 when increasing m to 5 from 3.

When p is close to 0.5, $\hat{p}_{m.ml}$ is usually the best except for cases with nearly random ranking, where \hat{p}_{ml} and $\hat{p}_{m.ml}$ are the best two, with one slightly outperforming the other or comparable performance otherwise. When p is close to 0 (or 1), \hat{p}_{iso} is preferred as its overall performance is the best. As in the DPS model, δ tends to increase as ρ decreases or N increases.

Among 640 simulation scenarios under TIC, there are only 23 scenarios where the maximum RE among the six estimators falls below 1. Again, all these scenarios are for nearly random ranking, and the lowest value of the best estimator among these scenarios is 0.990. Thus, under TIC models, as long as an appropriate estimator is chosen, loss of estimation efficiency is not a concern either, even if it is not helpful to use RSS-t over SRS. As to whether to utilize the tie information when estimating p , the answer is again yes in general situations.

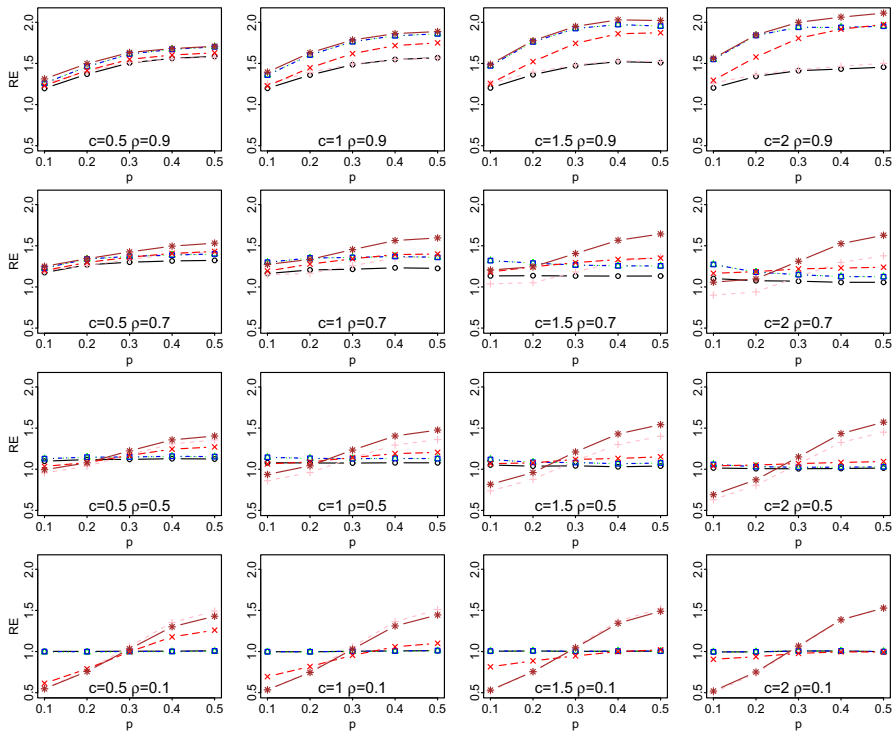


Fig. 3 Simulation: comparing relative efficiency of \hat{p}_{st} (represented by \circ), \hat{p}_{sp} (represented by Δ), \hat{p}_{iso} (represented by \square), \hat{p}_{ml} (represented by $+$), $\hat{p}_{t.ml}$ (represented by \times) and $\hat{p}_{m.ml}$ (represented by $*$) to \hat{p}_{SFS} as a function of population proportion p , for various (c, ρ) settings with $(N, m) = (30, 3)$ under TIC models. This figure appears in color in the electronic version of this paper

5 An empirical study

The performance of the six proportion estimators has been evaluated under two existing classes of tie-generating models through simulation. In practical situations, ranking can be done through ordinal variables that are associated with the (binary) response variable of interest and so ties may frequently happen. This can be thought of as another tie-generating mechanism, of which DPS might become a special case if the ordinal variable is defined by discretizing a continuous variable. Here, we conduct an empirical study to further examine the performance under this third mechanism using real data, Wisconsin Breast Cancer Data (WBCD), including 699 patients from a doctor’s clinic. In particular, some ordinal variables in this dataset may be defined based on a natural and error-prone process instead of using any underlying continuous variables.

WBCD is available online at UCI machine learning repository (Lichman 2013). It contains a binary variable (say Y) indicating tumour status and related variables of 9 visually cytological characteristics that are measured using an ordinal scale and are coded from 1 to 10, including clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin,

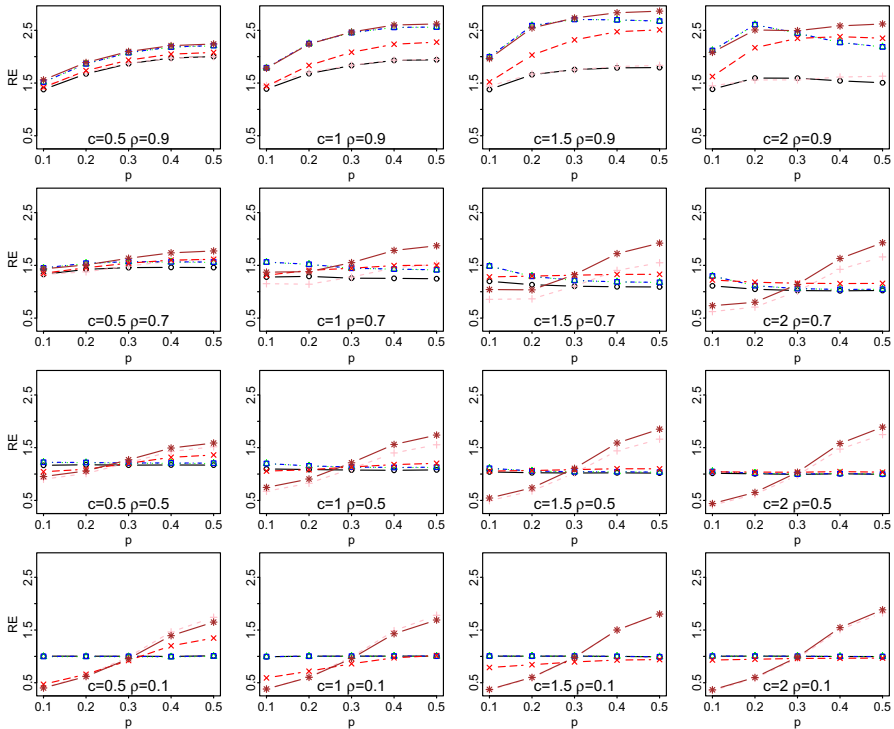


Fig. 4 Simulation: comparing relative efficiency of \hat{p}_{st} (represented by \circ), \hat{p}_{sp} (represented by Δ), \hat{p}_{iso} (represented by \square), \hat{p}_{ml} (represented by $+$), $\hat{p}_{t.ml}$ (represented by \times) and $\hat{p}_{m.ml}$ (represented by $*$) to \hat{p}_{SFS} as a function of population proportion p , for various (c, ρ) settings with $(N, m) = (30, 5)$ under TIC models. This figure appears in color in the electronic version of this paper

normal nucleoli, and mitoses. We note that although diagnosing the tumour status of a patient is expensive and requires a comprehensive biopsy procedure, these cytological variables can be easily measured and therefore can be used for ranking. In our study, WBCD is treated as a hypothetical population, and we are interested in estimating the incidence rate p of breast cancer in this “population”, where 241 out of 699 patients have breast cancer, and so $p = 0.344$.

We consider three combinations for (N, m) , $(15, 3)$, $(15, 5)$ and $(30, 3)$; and for each, we draw 100,000 samples using RSS-t from WBCD, where sampling is all done with replacement. We select uniformity of cell size, epithelial cell size and mitoses as ranking variables, each having correlation 0.818, 0.683, 0.431 with Y and approximately representing relatively good, fair, poor ranking quality, respectively. Figure 5 shows bar plots of these concomitant variables. Clearly, the frequency of level 1 or 2 is much higher than the other levels. Therefore, ranking ties can easily arise in a set of size 3 or 5.

Table 2 reports estimated bias, relative efficiency, and percent of sample size reduction (PSSR) for the scenarios considered. For each of the six estimators, the bias is estimated by the difference between the mean of the proportion estimates based on

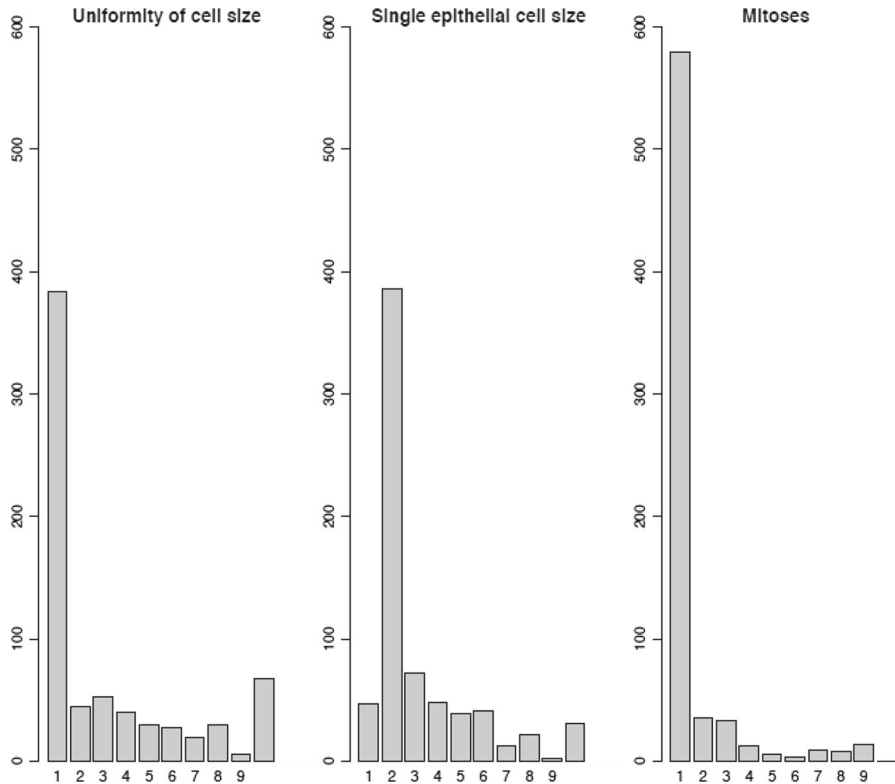


Fig. 5 Empirical Study using WBCD: bar plots of ranking variables in WBCD: uniformity of cell size, epithelial cell size and mitoses

$B = 100,000$ replicates and the true population proportion $p = 0.344$. The RE is defined as in Sect. 4. The PSSR of an estimator \hat{p} measures what percent of sample units can be reduced for \hat{p} to achieve the same precision as the SRS estimator \hat{p}_{SRS} , defined as

$$PSSR = \left(1 - \frac{\widehat{Var}(\hat{p})}{p(1-p)/N} \right) \times 100,$$

where $\widehat{Var}(\hat{p})$ is the estimated variance of \hat{p} .

From Table 2, we find that the three nonparametric estimators are (nearly) unbiased in all the settings considered. Among them, both \hat{p}_{iso} and \hat{p}_{sp} outperform \hat{p}_{st} that does not utilize the tie information; and \hat{p}_{iso} is (slightly) better than \hat{p}_{sp} for $N = 15$, and their performance becomes nearly identical for $N = 30$. Overall, \hat{p}_{iso} is the best nonparametric estimator.

Among the three likelihood-based estimators, $\hat{p}_{t.ml}$ consistently overestimates p (or more precisely, its estimates tend to be closer to 0.5 than they should be); and $\hat{p}_{m.ml}$

Table 2 Empirical Study using WBCD: comparing performance of different proportion estimators based on 100,000 RSS-t samples (winning values in each setting are bolded)

(N, m)	Ranking variables								
	Uniformity of cell size			Single epithelial cell size			Mitoses		
	$(\rho = 0.818)$			$(\rho = 0.683)$			$(\rho = 0.431)$		
	Bias	RE	PSSR	Bias	RE	PSSR	Bias	RE	PSSR
(15, 3)									
\hat{p}_{St}	0.00	1.45	31	0.00	1.33	25	0.00	1.07	6
\hat{p}_{Sp}	0.00	1.66	40	0.00	1.47	32	-0.01	1.12	11
\hat{p}_{iso}	0.00	1.67	40	0.00	1.47	32	0.00	1.13	11
\hat{p}_{ml}	0.00	1.45	31	0.00	1.36	26	0.02	1.19	18
$\hat{p}_{t.ml}$	0.02	1.40	31	0.02	1.31	26	0.02	1.08	10
$\hat{p}_{m.ml}$	0.00	1.66	40	0.00	1.52	34	0.02	1.34	27
(15, 5)									
\hat{p}_{St}	0.00	1.77	44	0.00	1.53	35	0.00	1.11	10
\hat{p}_{Sp}	-0.01	2.09	53	-0.01	1.71	42	-0.01	1.17	15
\hat{p}_{iso}	-0.01	2.11	53	-0.01	1.74	43	-0.01	1.18	16
\hat{p}_{ml}	0.00	1.76	43	0.01	1.54	35	0.03	1.22	23
$\hat{p}_{t.ml}$	0.02	1.67	42	0.02	1.49	35	0.02	1.16	17
$\hat{p}_{m.ml}$	0.00	2.07	52	0.00	1.76	43	0.03	1.40	33
(30, 3)									
\hat{p}_{St}	0.00	1.46	31	0.00	1.34	25	0.00	1.07	7
\hat{p}_{Sp}	0.00	1.68	41	0.00	1.49	33	0.00	1.14	13
\hat{p}_{iso}	0.00	1.68	41	0.00	1.49	33	0.00	1.15	13
\hat{p}_{ml}	0.00	1.47	32	0.00	1.38	28	0.02	1.19	21
$\hat{p}_{t.ml}$	0.02	1.35	31	0.02	1.27	27	0.02	1.04	12
$\hat{p}_{m.ml}$	0.00	1.69	41	0.00	1.55	36	0.02	1.33	30

and \hat{p}_{ml} appear to be unbiased when the ranking quality is fair or good; but they both become biased (again, estimates tend to be closer to 0.5) when the ranking quality is poor. In terms of RE and PSSR, $\hat{p}_{m.ml}$ clearly outperforms $\hat{p}_{t.ml}$ and \hat{p}_{ml} . Thus, $\hat{p}_{m.ml}$ is the best likelihood-based estimator.

Table 2 shows that as the ranking quality ρ decreases, the RE decreases for every estimator as expected. When the ranking quality is good, the performance of \hat{p}_{iso} and $\hat{p}_{m.ml}$ is quite close, with \hat{p}_{iso} being slightly better than $\hat{p}_{m.ml}$ for $N = 15$ and $\hat{p}_{m.ml}$ slightly better for $N = 30$. When the ranking quality decreases, it seems that the performance of \hat{p}_{iso} deteriorates faster than $\hat{p}_{m.ml}$, and so $\hat{p}_{m.ml}$ becomes a clear winner. Even when the ranking quality is poor, about 30% reduction in sample size from using RSS-t over SRS can be achieved by $\hat{p}_{m.ml}$. In case that information about ranking quality is not available, we would recommend the use of $\hat{p}_{m.ml}$ in this example for its consistently top performance. This agrees with the conclusion from our simulation in Sect. 4 that when p is not extreme, $\hat{p}_{m.ml}$ is generally preferred for $N \in \{15, 30\}$.

Furthermore, by comparing results of the setting (15, 3) with those of (15, 5), we find that the RE of every estimator increases with the set size m while the total sample size N is fixed. By comparing results of (15, 3) with (30, 3), we find that the RE of every estimator increases with N while m is fixed with an exception of $\hat{p}_{t.ml}$ that shows the opposite pattern.

6 Conclusion

We have extended the work of Frey (2012) about RSS-t to binary outcomes. Besides the direct extension of the three nonparametric estimators considered in Frey (2012) and the maximum likelihood estimator proposed in Terpstra (2004), we have added two new likelihood-based estimators into the pool to incorporate the tie information. We have thoroughly examined the performance of the six proportion estimators, through simulation using data generated from both DPS and TIC models and a data example involving a natural tie-generating process via the use of ordinal ranking variables.

Our results suggest that using RSS-t over SRS, combined with an appropriate choice of the proportion estimator, can greatly improve the efficiency of proportion estimation. Unless ranking is close to random, utilizing tie information is helpful in the estimation process, which leads to considerable efficiency gain when the quality of ranking is good. We also find that the relative performance of the different proportion estimators can depend on the value of p , ranking quality and tie-generating mechanism, as detailed in Sects. 4 and 5. However, in a very wide range of settings, we find that \hat{p}_{iso} works well for rare or common events, and $\hat{p}_{m.ml}$ is the best choice otherwise, regardless of the tie-generating mechanism. Thus, in the most common situations where one can get a rough estimate of p but is not sure about the other factors, we recommend that $\hat{p}_{m.ml}$ be used for p close to 0.5 and \hat{p}_{iso} be used otherwise.

This paper focuses on proportion estimation from (balanced) ranked set samples with tie structures recorded. One future task can be to extend our work to judgment post-stratification (JPS) with binary outcomes, where empty strata may arise often with small sample sizes so that different versions of the isotonic estimator no longer lead to identical estimates. It would be interesting to investigate how various proportion estimators perform for JPS in the presence of tie information.

We mention that deriving theoretical properties of the proposed proportion estimators is very difficult because the added tie structure of RSS-t depends on the researcher's ability to rank the sample units. For the same reason, Frey (2012) did not present any theoretical justification for mean estimators proposed for RSS-t. Even how the original RSS mean estimator (without utilizing the tie information) performs when ties occur (and then are randomly broken) has not been theoretically examined in the past. Studying these estimators formally in presence of ties may help researchers deeply understand their behaviors, and this has an ample space for future research. On the other hand, obtaining such theoretical results may require large-sample arguments. It is well known that ranked set sampling is a cost efficient method, and so the small-sample properties (typically studied via simulation) are more relevant in practice.

Finally, we note that a natural way to draw statistical inference for RSS-t samples is via resampling. We can easily adapt a procedure called BRSSR (*Bootstrap RSS by*

row), proposed by Modarres et al. (2006). The BRSSR method first resamples each judgment stratum separately. Let $\mathbf{Y} = \{Y_{[i]j}, i = 1, \dots, m, j = 1, \dots, n\}$ be the original RSS sample. Then a bootstrap sample $\mathbf{Y}^b = \{Y_{[i]j}^b, i = 1, \dots, m, j = 1, \dots, n\}$ is obtained by drawing $Y_{[i]j}^b$ with replacement from the discrete uniform distribution on the set $\{Y_{[i]1}, \dots, Y_{[i]n}\}$, for each $i = 1, \dots, m$, respectively. However, for RSS-t, the sample should also include $\mathbf{T}^1, \dots, \mathbf{T}^n$, matrices of tie information. Therefore, we propose to draw the i th row of \mathbf{T}^j , say $\mathbf{T}_i^{j,b}$, along with $Y_{[i]j}^b$ to construct a RSS-t bootstrap sample $(\mathbf{Y}, \mathbf{T})^b = \left\{ \left(Y_{[i]j}^b, \mathbf{T}_i^{j,b} \right), i = 1, \dots, m, j = 1, \dots, n \right\}$; and statistical inference for RSS-t can be done based on multiple copies of the bootstrap samples $[(\mathbf{Y}, \mathbf{T})^b]_{b=1}^B$. For example, to construct a $(1 - \alpha)\%$ confidence interval for p based on some proportion estimator (say \hat{p}), we can generate B bootstrap samples using the method described above from a RSS-t sample and then compute the corresponding B bootstrap estimates $(\hat{p}^1, \dots, \hat{p}^B)$. Then the $(1 - \alpha)\%$ confidence interval can be given by $(\hat{p}_{\frac{\alpha}{2}}, \hat{p}_{1-\frac{\alpha}{2}})$, where \hat{p}_α is the α th sample quantile of $(\hat{p}^1, \dots, \hat{p}^B)$. Other methods of interval estimation such as random grouping and jackknife can be also considered for potentially better performance. The code for implementing the modified BRSSR procedure and our numerical experiments is publicly available at goo.gl/sf7DJW.

Acknowledgements We thank Professor Johan Lim for his comments on an earlier version of this paper, Professor Jesse Frey for sharing his R-Code and the UCI machine learning repository for online data use. We are also thankful to two anonymous referees and an associate editor for their valuable comments which improved an earlier version of this paper.

Compliance with ethical standards

Conflict of interest No potential conflict of interest was reported by the authors.

References

- Chen H, Stasny EA, Wolfe DA (2005) Ranked set sampling for efficient estimation of a population proportion. *Stat Med* 24:3319–3329
- Chen H, Stasny EA, Wolfe DA (2007) Improved procedures for estimation of disease prevalence using ranked set sampling. *Biom J* 49(4):530–538
- Frey J (2012) Nonparametric mean estimation using partially ordered sets. *Environ Ecol Stat* 19(3):309–326
- Halls LK, Dell TR (1966) Trial of ranked-set sampling for forage yields. *For Sci* 12:22–26
- Hatefi A, Jafari Jozani M (2017) An improved procedure for estimation of malignant breast cancer prevalence using partially rank ordered set samples with multiple concomitants. *Stat Methods Med Res* 26(6):2552–2566
- Howard RW, Jones SC, Mauldin JK, Beal RH (1982) Abundance, distribution, and colony size estimates for *Reticulitermes* spp. (Isopter: Rhinotermitidae) in Southern Mississippi. *Environ Entomol* 11:1290–1293
- Kvam PH (2003) Ranked set sampling based on binary water quality data with covariates. *J Agri Biol Environ Stat* 8:271–279
- Lichman M (2013) UCI machine learning repository. School of Information and Computer Science, University of California, Irvine, CA. <http://archive.ics.uci.edu/ml>. Accessed 14 Feb 2018
- MacEachern SN, Stasny EA, Wolfe DA (2004) Judgement post-stratification with imprecise rankings. *Biometrics* 60:207–215

- Mahdizadeh M, Zamanzade E (2017) To appear in efficient body fat estimation using multistage pair ranked set sampling. *Stat Methods Med Res.* <https://doi.org/10.1177/0962280217720473>
- McIntyre GA (1952) A method for unbiased selective sampling using ranked set sampling. *Aust J Agric Res* 3:385–390
- Modarres R, Hui TP, Zheng G (2006) Resampling methods for ranked set samples. *Comput Stat Data Anal* 51(2):1039–1050
- Mu X (2015) Log-concavity of a mixture of beta distributions. *Stat Probab Lett* 99:125–130
- Nussbaum BD, Sinha BK (1997) Cost effective gasoline sampling using ranked set sampling. In: Proceedings of the section on statistics and the environment, pp 83–87. American Statistical Association
- Ozturk O, Bilgin O, 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
- Terpstra JF (2004) On estimating a population proportion via ranked set sampling. *Biom J* 46(2):264–272
- Terpstra JF, Liudahl LA (2004) Concomitant-based rank set sampling proportion estimates. *Stat Med* 23:2061–2070
- Terpstra JF, Wang P (2008) Confidence intervals for a population proportion based on a ranked set sample. *J Stat Comput Simul* 78:351–366
- Wang X, Lim J, Stokes L (2016) Using ranked set sampling with cluster randomized designs for improved inference on treatment effects. *J Am Stat Assoc* 111(516):1576–1590