Submitted to the Brazilian Journal of Probability and Statistics
arXiv: arXiv:0000.0000

On Multiple Imputation for Unbalanced Ranked Set Samples With Applications in Quantile estimation

Saeid Amiri[a], Mohammad Jafari Jozani[b] and Reza Modarres[c]

[a]*Department of Civil, Geologic, and Mining Engineering, Polytechnique Montreal, QC, Canada*
[b]*Department of Statistics, University of Manitoba, Winnipeg, MB, Canada, R3T 2N2*
[c]*Department of Statistics, The George Washington University, Washington DC, USA*

Abstract. We consider multiple imputation (MI) for unbalanced ranked set samples (URSS) by considering them as data sets with missing values. We replace each missing value with a set of plausible values drawn from a predictive distribution that represents the uncertainty about the appropriate value to impute. Using the structure of the MI dataset, we develop algorithms that imitate the structure of URSS to carry out the desired statistical inference. We provide results for the convergence of the empirical distribution functions of imputed samples to the population distribution function, under both URSS and simple random sampling (SRS). We obtain the variances of the imputed URSS, and the expected values of the variance estimators. We also study the problem of quantile estimation using an imputed URSS and propose a hybrid method based on the bootstrap and imputation of URSS data. We apply our results to estimate the mean and quantiles of the mercury in contaminated fish under perfect and imperfect URSS.

1 Introduction

Ranked set sampling is a data collection technique for situations where measuring the variable of interest is difficult and/or costly but (imperfect) ranking of sampling units can be done cheaply. Introduced by McIntyre (1952), the study of ranked set sampling has resulted in a substantial literature. Ranked set sampling has found applications in agriculture, reliability (Mahdizadeh and Zamanzade , 2016), biometrics ((Samawi and Al-Sagheer, 2001)), and medical studies (Hatefi and Jafari Jozani, 2015), among others. Given the rank information, it is well-known that the usual estimator of the population mean using a ranked set sample (RSS) is more efficient than its counterpart under simple random sampling (SRS). Theoretical work has been undertaken on RSS designs including information theory (Jafari Jozani and Ahmadi, 2014), finite population inference (Ozturk and Jafari Jozani, 2014) and tests of perfect judgment ranking Amiri

MSC 2010 subject classifications: Primary 62-XX; secondary 62F40
Keywords and phrases. Nonresponse, Quantile, Ranked Set Sampling

et al. (2016), to name a few. We refer the reader to a monograph in RSS by Chen et al. (2004) that provides invaluable information on ranked set sampling, its many variants and applications. For a recent review of the current developments on RSS, see Wolfe (2012).

To collect a RSS of size k from a population with corresponding population cumulative distribution function (CDF) F, one randomly identifies new k^2 units in the population and randomly divides them into k groups (sets) of size k. Units in each set are ordered using any means other than the actual measurement of the variable of interest. Now, one selects the jth smallest unit from set j, $j = 1,\ldots,k$ for further inspection and taking the final measurement. This constructs one cycle of the ranked set sampling procedure. To increase the sample size to nk, one can repeat this process n times with new sample. Ranked set sampling procedure for one cycle is presented in Table 1, where sampling units are ordered in each set. According to this procedure one needs to select the diagonal elements $\{X^*_{(11)}, X^*_{(22)}, \ldots, X^*_{(kk)}\}$. It should be noted that we use the round parentheses instead the square parentheses because the ranking is often imperfect and $X^*_{(jj)}$ might not necessarily be the jth order statistic in its corresponding set. For the sake of simplicity, we represent the sample as $\{X_{(1)}, X_{(2)}, \ldots, X_{(k)}\}$. The procedure is then repeated until n cycles of k observations are obtained. It is noteworthy that the resulting measurements are independently, but not identically distributed from F.

Table 1 *Display of k^2 observations in k ranked sets of k units with RSS sample asterisked.*

$X^*_{(11)}$	$X_{(12)}$	$\cdots$	$X_{(1(k-1))}$	$X_{(1k)}$
$X_{(21)}$	$X^*_{(22)}$	$\cdots$	$X_{(2(k-1))}$	$X_{(2k)}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
$X_{(k1)}$	$X_{(k2)}$	$\cdots$	$X_{(k(k-1))}$	$X^*_{(kk)}$

For an unbalanced ranked set sample (URSS) the number of observations with rank r, $r = 1,\ldots,k$, are not necessarily equal, and the resulting sample is denoted by $X_{\text{URSS}} = \{\mathcal{X}_r, r =,\ldots,k\}$, where $\mathcal{X}_r = \{X_{(r)j}; j = 1,\ldots,n_r\}$ are the actual observations in a random sample from $F_{(r)}$, the CDF of the rth order statistic in a set of size k from F. The empirical distribution function (EDF) of X_{URSS} is defined by

$$\widehat{F}_{q_n}(t) = \frac{1}{n}\sum_{r=1}^{k}\sum_{j=1}^{n_r} I(X_{(r)j} \leq t) = \sum_{r=1}^{k} q_{n_r}\widehat{F}_{(r)}(t), \tag{1.1}$$

where $\widehat{F}_{(r)}(t) = \frac{1}{n_r}\sum_{j=1}^{n_r} I(X_{(r)j} \leq t)$, when $n_r \geq 1$, $n = \sum_{r=1}^{k} n_r$ and $q_{n_r} = n_r/n$. An estimate of the population mean using (1.1) is the pooled sample mean defined by

$$\bar{X}_{q_n}(t) = \frac{1}{n}\sum_{r=1}^{k}\sum_{j=1}^{n_r} X_{(r)j}. \tag{1.2}$$

Most statistical procedures that are designed for simple random samples (SRS) can be extended to the balanced ranked set sampling design provided that ranked mechanism is consistent, i.e., $F(t) = \frac{1}{k}\sum_{r=1}^{k} F_{(r)}(t)$. In contrast, unbalanced ranked set sampling is a more complex sampling process which does not satisfy the fundamental consistency property. For this design, statistical inference is mostly based on large-sample theory. However, the asymptotic distribution of many URSS estimators can not be readily used as the sample size is usually small. In addition, as discussed in Amiri et al. (2014), the EDF of a URSS does not converge to the CDF and the algorithms developed for bootstrapping balanced RSS can not be applied for URSS situation. In other words, as n_r tends to infinity, $\widehat{F}_{(r)}(t) \xrightarrow{a.s.} F_{(r)}(t)$ and provided that $q_{n_r} \longrightarrow q_r$, for $r = 1, \ldots, n$, we have

$$\widehat{F}_{q_n}(t) - F_q(t) \xrightarrow{\mathcal{L}} 0 \text{ as } n_r \longrightarrow \infty, \tag{1.3}$$

where $\mathcal{L}$ stands for convergence in law and $\widehat{F}_q(t) = \sum_{r=1}^{k} q_r F_{(r)}(x) \neq F(x)$ unless $q_r = \frac{1}{k}$, for $r = 1, \ldots, k$.

To side-step this difficulty, one can transform the URSS data to a balanced RSS. This transformation allows one to apply standard techniques of bootstrap, estimation and testing that are available for balanced RSS data to the completed dataset, which contains both observed and imputed values. In this direction, Amiri et al. (2014) explored different bootstrapping methods.

In this paper, we take another approach for transforming an URSS to a balanced RSS using multiple imputation (MI) techniques and by treating URSS as a dataset with missing values. Instead of filling in a single value for each missing data item, we replace each missing value with a set of plausible values that represent the uncertainty about the right value to impute. This results in multiple imputed data sets where standard statistical analysis can be carried out on each imputed data set to produce multiple analysis results. These results are then combined to obtain one overall analysis. MI accounts for missing data by restoring not only the natural variability in the missing data, but also by incorporating the uncertainty caused by estimating missing data. Uncertainty is accounted for by creating different versions of the missing data and observing the variability between imputed data sets.

We also study a hybrid approach based on MI and bootstrap. We argue that the data obtained from an URSS design with $n = \sum_{r=1}^{k} n_r$ can be modelled as a missing data problem in the context of a balanced RSS design where the r-th smallest unit in a set of size k, $r = 1,\ldots,k$, was to be observed m times instead of n_r times, and $m = \max\{n_1,\ldots,n_k\}$. While MI was developed for SRS with iid structure, we develop MI algorithms for URSS and overcome this problem. Furthermore, we consider estimating the mean and the quantiles under URSS. Statistical inference for quantiles under RSS is difficult and we develop an algorithmic approaches via MI to circumvent this difficulty.

Rubin (1987) proposed MI as a method of handling missing values (nonresponse) in survey sampling to retain the main advantages of single imputation, and avoid its drawbacks by replacing each missing datum with two or more values representing a distribution of likely values. Since Rubin (1987), the theory and application of MI have been advanced in medical studies, high dimensional data analysis, cross-sectional and longitudinal data analysis, survival analysis, among others. In a recent book, Carpenter and Kenward (2012) show MI produces unbiased parameter estimates, is robust to departures from the normality assumption and provides reasonable results when the sample size is small or there is a high rate of missing data. In addition, MI is easily be applied in increasingly complex data structures and is computationally simpler than other methods for imputing missing data such as the maximum likelihood estimation. For more information, the reader is refered to Rubin (1996), Rubin (1987), Rubin and Schenker (1986), Schafer and Olsen (1998), Li et al. (2015).

In Section 2, we discuss MI for SRS, prove convergence of the EDF of the imputed sample to the population CDF. We obtain the variance of the imputed URSS, and the expected value of the variance estimator. Section 3 examines MI for URSS, shows the convergence of the EDF under MI, obtains the variance of the imputed mean, and the expected value of the variance estimator. In Section 4, we consider the problem of quantile estimation using URSS data and show how MI can be used to make inference about population quantiles using the transformed data. Section 5.1 describes a real data application to examine the performance of our proposed methods for estimating of mean and quantiles. Concluding remarks are given in Section 6. Finally, the Appendix is devoted to the proofs and some of the necessary technical results.

2 Multiple Imputation

Suppose we obtain an i.i.d sample $\{X_1, \ldots, X_N\}$ of size N from a distribution F with mean μ and variance $\sigma^2 < \infty$. Suppose $N - n$ of the observations are missing and we are interested in estimating the population mean μ. Missing data are assumed to be ignorable, hence the observed sample $\mathcal{X} = \{X_1, \ldots, X_n\}$ is a completely random sample. To estimate μ, using the MI technique, we proceed via the MI Algorithm, as discussed below:

MI Algorithm

1. Sample $\mathfrak{r} = (N - n)$ values independently with replacement from the observed sample $\mathcal{X} = \{X_1, \cdots, X_n\}$. These provide the imputed values $\mathcal{X}_m^\diamond = \{X_{1,m}^\diamond, \ldots, X_{r,m}^\diamond\}$.
2. Estimate μ by $\widehat{\mu}_m = \frac{1}{N}\left(\sum_{i=1}^{n} X_{i,m} + \sum_{i=1}^{\mathfrak{r}} X_{i,m}^\diamond\right)$ and the within-imputation variance by
$$U_m = \widehat{Var}(\widehat{\mu}_m) = \frac{1}{N(N-1)}\left\{\sum_{i=1}^{n}(X_{i,m} - \widehat{\mu}_m)^2 + \sum_{i=1}^{\mathfrak{r}}(X_{i,m}^\diamond - \widehat{\mu}_m)^2\right\},$$
where $N = n + \mathfrak{r}$.
3. Repeat Steps 1-2, for $m = 1, \ldots, M$ to obtain M imputed datasets, and M estimates $\widehat{\mu}_m$ and $U_m = \widehat{Var}(\widehat{\mu}_m)$ for $m = 1, \ldots, M$.
4. Let $\widehat{W} = \frac{1}{M}\sum_{m=1}^{M} U_m$ be the average within-imputation variance and $\widehat{B} = \frac{1}{M-1}\sum_{m=1}^{M}(\widehat{\mu}_m - \widehat{\mu})^2$, be the between-imputation variance. The MI estimator of μ is then defined as
$$\widehat{\mu} = \frac{1}{M}\sum_{m=1}^{M}\widehat{\mu}_m, \tag{2.1}$$
with its corresponding variance estimator
$$\widehat{V} = \widehat{W} + (\frac{M+1}{M})\widehat{B}. \tag{2.2}$$

Rubin and Schenker (1986) present a nonparametric MI method, called the approximate Bayesian bootstrap imputation (ABBI), which involves a two-stage sampling procedure to generate proper imputations with minimal distributional assumptions. To obtain ABBI, one should proceed as in the MI Algorithm except that its step 1 is modified as follows: Sample n observations with replacement from the observed $\mathcal{X} = \{X_1, \cdots, X_n\}$. Denote this sample by $\mathcal{X}^*_m = \{X^*_{1,m}, \ldots, X^*_{n,m}\}$. Select the $\mathfrak{r} = N - n$ missing X_i's with replacement from $\mathcal{X}^*_m$. These provide the imputed values $\mathcal{X}^\diamond_m = \{X^\diamond_{1,m}, \ldots, X^\diamond_{r,m}\}$. Rubin (1987) showed that ABBI provides an asymptotically unbiased estimate of μ as both N and M tend to infinity.

The following Proposition shows that the EDF of the completed SRS sample, using MI, converges to the population cumulative distribution function F.

Proposition 1. *Consider the MI of a SRS and let $\mathcal{X}^\diamond_m = \{X^\diamond_{1,m}, ..., X^\diamond_{\mathfrak{r},m}\}$ be an i.i.d sample randomly taken from $\widehat{F}_n(x)$ in Step 1 of the MI Algorithm, where $\widehat{F}_n(x)$ is the EDF of $\mathcal{X} = \{X_1, ..., X_n\}$. Let $N = n + \mathfrak{r}$. If $\widehat{F}^\diamond_N(t)$ is the EDF of the completed sample $\{\mathcal{X}, \mathcal{X}^\diamond_m\}$, then*

$$\|\widehat{F}^\diamond_N(t) - F(t)\|_\infty = \sup_{t\in R} |\widehat{F}^\diamond_N(t) - F(t)| \longrightarrow 0.$$

The following Proposition shows that the EDF of the completed SRS sample, using ABBI, converges to the population cumulative distribution function (CDF).

Proposition 2. *Consider the approximate Bayesian bootstrap imputation and let $\mathcal{X}^\diamond_m = \{X^\diamond_{1,m}, ..., X^\diamond_{\mathfrak{r},m}\}$ be an iid sample randomly taken from $\widehat{F}_n(x)$ in Step 1 of the MI Algorithm using the ABBI method, where $\widehat{F}_n(x)$ is the EDF of $\mathcal{X} = \{X_1, ..., X_n\}$. If $\widehat{F}^\diamond_N(t)$, with $N = n + \mathfrak{r}$, is the EDF of the complete sample $\{\mathcal{X}, \mathcal{X}^\diamond_m\}$, then*

$$\|\widehat{F}^\diamond_N(t) - F(t)\|_\infty = \sup_{t\in R} |\widehat{F}^\diamond_N(t) - F(t)| \longrightarrow 0,$$

The following proposition obtains expressions for the variance of the imputed mean, $Var(\widehat{\mu})$, and the expected value of the variance estimator $E(\widehat{V})$ under MI.

Proposition 3. *The variance of the imputed mean (2.1) and the expected value of the variance estimator (2.2) are, respectively,*

$$Var(\widehat{\mu}) = \frac{\sigma^2}{n} + \frac{1}{M}\left(\frac{N-n-1}{N^2}\right)\sigma^2, \tag{2.3}$$

and

$$E(\widehat{V}) = \frac{1}{(N-1)}(1 - \frac{1}{n} - \frac{N-n-1}{N^2}) + \frac{M+1}{M}\frac{N-n-1}{N^2}\sigma^2. \tag{2.4}$$

From (2.4) one can easily observe that the estimator of the variance of the imputed mean is biased. Under the ABBI method, Kim (2002) derived an expression for the $Var(\widehat{\mu})$ as follows,

$$Var(\widehat{\mu}) = \frac{1}{N^2}\left(\frac{N^2}{n} + \frac{N-n}{M}\left(\frac{N-1}{n} - \frac{N}{n^2}\right)\right)\sigma^2, \tag{2.5}$$

and showed that the ABBI variance estimator is also biased as

$$E(\widehat{V}) = \frac{1}{N^2}\left(\frac{N^2}{n} - \frac{(N-n)N}{n}\left(\frac{3}{N} + \frac{1}{n}\right) + \frac{N-n}{M}\left(\frac{N-1}{n} - \frac{N}{n^2}\right)\right)\sigma^2. \tag{2.6}$$

In order to reduce the bias, Kim (2002) suggests to select d observations from $\mathcal{X}$ in Step 1 of the ABBI algorithm, where $d = \frac{(n-1)(N-n-1)(N-2)}{(N-1)(N-n+1)+N+n+1}$. Parzen et al. (2005) suggests a bias corrected version of ABBI method using a new estimator $\widehat{V}^{\diamond} = f\widehat{V}$, where $f = \frac{Var(\widehat{\mu})}{E(\widehat{V})}$ provided that f does not depend on any unknown parameters. One can use similar methods to reduce the bias of the variance estimator (2.2). One can also derive an unbiased estimator of the variance by directly working with (2.2) and modifying the bias.

3 Imputing URSS

In this section, we use the MI technique to transform URSS data to a balanced RSS. This transformation allows one to apply standard techniques of bootstrap, estimation and testing that are available for balanced ranked set sampling to the completed dataset. To accomplish this, we let $N = \max\{n_i, i = 1,\ldots,k\}$ and using the MI Algorithm, for each $m = 1,\ldots,M$, we obtain $N - n_r$ observations from $\mathcal{X}_r = \{X_{(r),1},\ldots,X_{(r)n_r}\}$ to add to the rth stratum in order to fill in the values that are needed to construct the balanced RSS. This results in a balanced imputed RSS data $\mathcal{X}^{\diamond} = \{\mathcal{X}_1^{\diamond},\ldots,\mathcal{X}_M^{\diamond}\}$, where $\mathcal{X}_m^{\diamond} = \{\mathcal{X}_{1,m}^{\diamond},\ldots,\mathcal{X}_{k,m}^{\diamond}\}$, $m = 1,\ldots,M$, and

$$\mathcal{X}_{r,m}^{\diamond} = \{X_{(r)1,m}^{\diamond}, X_{(r)2,m}^{\diamond},\ldots,X_{(r)n_r,m}^{\diamond}, X_{(r)n_r+1,m}^{\diamond},\ldots,X_{(r)N,m}^{\diamond}\}. \tag{3.1}$$

Here, $X^{\diamond}_{(r)j,m} = X_{(r)j},\ r = 1,\dots,k, j = 1,\dots,n_r$ is the observation from the $r$the stratum. Now, estimates of the mean and its variance can be obtained according to the MI and ABBI methods. Let

$$\widehat{\mu}_m = \frac{1}{Nk}\sum_{r=1}^{k}\sum_{j=1}^{N} X^{\diamond}_{(r)j,m}, \tag{3.2}$$

$$U_m = \frac{1}{k^2}\sum_{r=1}^{k}\frac{1}{N(N-1)}\sum_{j=1}^{N}(X^{\diamond}_{(r)j,m} - \widehat{\mu}_{(r),m})^2, \tag{3.3}$$

where $\widehat{\mu}_{(r),m} = \frac{1}{N}\sum_{j=1}^{N} X^{\diamond}_{(r)j,m}$.

Proposition 4. *Suppose $\mathcal{X} = \{X_{(i)j}, i = 1,\dots,k, j = 1,\dots,n_i\}$ is an URSS from a population with CDF F and $\mathcal{X}^{\diamond} = \{\mathcal{X}^{\diamond}_1,\dots,\mathcal{X}^{\diamond}_M\}$ is the imputed RSS (IRSS) using the MI technique, where $\mathcal{X}^{\diamond}_m = \{\mathcal{X}^{\diamond}_{1,m},\dots,\mathcal{X}^{\diamond}_{k,m}\}$ with $\mathcal{X}^{\diamond}_{r,m}$ defined in (3.1). The variance of the mean estimator $\widehat{\mu} = \frac{1}{M}\sum_{m=1}^{M}\widehat{\mu}_m$, where $\widehat{\mu}_m$ as in (3.2), is given by*

$$Var(\widehat{\mu}) = \frac{1}{k^2}\sum_{r=1}^{k}\left(\frac{1}{n_r} + \frac{1}{M}(\frac{\mathfrak{r}_r - 1}{N^2})\right)\sigma^2_{(r)}, \tag{3.4}$$

where $\sigma^2_{(r)} = \int (x - \mu_{(i)})^2 dF_{(i)}(x)$, $\mu_{(r)} = \int x dF_{(i)}(x)$, and the expected value of the MI variance estimator is given by,

$$E(\widehat{V}) = \frac{1}{k^2}\sum_{r=1}^{k}\left(\frac{1}{(N-1)}(1 - \frac{1}{n_r} - \frac{\mathfrak{r}_r - 1}{N^2}) + \frac{M+1}{M}\frac{\mathfrak{r}_r - 1}{N^2}\right)\sigma^2_{(r)}, \tag{3.5}$$

where $\widehat{V}$ is defined in Step 4 of the MI algorithm.

Following the proof of Proposition 4, one can also prove the following Proposition for the ABBI algorithm.

Proposition 5. *Suppose $\mathcal{X} = \{X_{(i)j}, i = 1,\dots,k, j = 1,\dots,n_i\}$ is an URSS of size $n = \sum_{r=1}^{k} n_r$ from a population with CDF F and corresponding population mean μ and variance $\sigma^2 < \infty$. Suppose also that $\mathcal{X}^{\diamond} = \{\mathcal{X}^{\diamond}_1,\dots,\mathcal{X}^{\diamond}_M\}$ is the imputed RSS (IRSS) using the ABBI in the MI technique, where $\mathcal{X}^{\diamond}_m = \{\mathcal{X}^{\diamond}_{1,m},\dots,\mathcal{X}^{\diamond}_{k,m}\}$ with $\mathcal{X}^{\diamond}_{r,m}$ begin defined as in (3.1). Then, the variance of the mean estimator is given by*

$$Var(\widehat{\mu}) = \frac{1}{k^2N^2}\sum_{r=1}^{k}\left(\frac{N^2}{n_r} + \frac{N - n_r}{M}\left(\frac{N-1}{n_r} - \frac{N}{n_r^2}\right)\right)\sigma^2_{(r)}, \tag{3.6}$$

with the expected value of the ABBI MI estimator of the variance as,

$$E(\widehat{V}) = \frac{1}{k^2N^2}\sum_{r=1}^{k}\left(\frac{N^2}{n_r} - \frac{(N-n_r)N}{n}\left(\frac{3}{N}+\frac{1}{n_r}\right) + \frac{N-n_r}{M}\left(\frac{N-1}{n_r}-\frac{N}{n_r^2}\right)\right)\sigma^2_{(r)}. \quad (3.7)$$

Propositions 4 and 5 show that the variance estimators are biased. To reduce the bias one can use the method proposed by Parzen et al. (2005). Under the ranked set sampling setting, $f = \frac{Var(\widehat{\mu})}{E(\widehat{V})}$ is a function of $\sigma^2_{(r)}$ that need to be estimated from the sample. Demirtas et al. (2007) warns that this modification may be inferior to the approximation Bayesian bootstrap and should be used with caution under SRS. Our experience with RSS data shows that this modification is unnecessary.

4 Quantile Estimation using URSS Data

In this section we show how the MI Algorithm can be used to make inference about the population quantile using URSS data. This is an important problem as there are many cases in which one is interested in making inference about the quantiles of a distribution. For a literature review of quantile estimation under ranked set type sampling designs see Chen (2000), Nourmohammadi et al. (2014) and references therein.

Let $\mathcal{X} = \{X_1, \ldots, X_n\}$ represent a SRS of size n from a population with continuous distribution function $F(\cdot)$ and density function $f(\cdot)$. The p-th quantile of the population is defined as

$$\zeta_p = \inf\{x : F(x) \geq p\},$$

where $F(\zeta_p) = p$. Let $X_{(1)} \leq \ldots \leq X_{(n)}$ be the order statistics of $\mathcal{X}$. Note that a simple method of estimating the p-th quantile uses

$$\widehat{\zeta}_{p,SRS} = \begin{cases} X_{(np)} & \text{if } np \text{ is an integer,} \\ X_{(\lfloor np \rfloor + 1)} & \text{if } np \text{ is not an integer.} \end{cases}$$

For $0 < p < 1$, suppose $F(\cdot)$ is absolutely continuous at ζ_p. The asymptotic distribution of a central order statistic is given by (Serfling, 2009)

$$\sqrt{n} f(F^{-1}(p))\left(\frac{X_{(\lfloor np \rfloor+1)} - F^{-1}(p)}{\sqrt{p(1-p)}}\right) \xrightarrow{\mathcal{L}} N(0,1), \quad \text{as } n \longrightarrow \infty,$$

and $|\ \tilde{\zeta}_{p,SRS} - \zeta_p \ | \leq \frac{2\sqrt{\log n}}{f(\zeta_p)\sqrt{n}}$.

To find the quantile estimate under RSS sample, one can order the entire RSS set of observations $\{X_{(r)j}, r = 1,\ldots,k, j = 1,\ldots,m\}$ to obtain $X^*_{(1)} \leq \ldots \leq X^*_{(n)}$ with $n = mk$. Using (1.1), the EDF of the RSS sample, for multi-cycle RSS data with $n = mk$, is given by

$$\widehat{F}(x) = \frac{1}{mk}\sum_{r=1}^{k}\sum_{j=1}^{m} I(X_{(r)j} \leq x) = \frac{1}{n}\sum_{i=1}^{n} I(X^*_{(i)} \leq x),$$

and the RSS estimator of the p-th quantile is now defined by $\widehat{\zeta}_{p,RSS} = \inf\{x : \widehat{F}(x) \geq p\}$. Chen (2000) studied the estimation of the population quantiles using balanced RSS data when $n = mk$. The sample pth quantile based on balanced RSS data can be expressed as

$$\widehat{\zeta}_{p,RSS} = \begin{cases} X^*_{(np)} & \text{if } np \text{ is integer,} \\ X^*_{(\lfloor np \rfloor + 1)} & \text{if } np \text{ is not integer.} \end{cases}$$

Let the density function f be positive in a neighborhood of ζ_p and continuous at ζ_p. Then, the convergence of the estimate $\widehat{\zeta}_{p,RSS}$ to ζ_p, for sufficiently large n, is guaranteed as

$$P\Big(\mid \widehat{\zeta}_{p,RSS} - \zeta_p \mid < \frac{2\sqrt{\log n}}{f(\zeta_p)\sqrt{n}} \Big) = 1.$$

Also, the asymptotic distribution of $\widehat{\zeta}_{p,RSS}$ is

$$\sqrt{n}(\widehat{\zeta}_{p,RSS} - \zeta_p) \xrightarrow{\mathcal{L}} N(0, \frac{\sigma^2_{n,p}}{f^2(\zeta_p)}),$$

where $\sigma^2_{n,p} = \frac{1}{n}\sum_{r=1}^{n} B(p,r,n+r-1)(1-B(p,r,n+r-1))$, and $B(x,r,s)$ denotes the distribution function of a beta random variable with parameters r and s. Since $B(p,r,n+r-1) = F_{(r)}(\zeta_p)$, the variance does not depend on the unknown parameter and $\sqrt{n}(\widehat{\zeta}_{p,RSS} - \zeta_p)/\frac{\widehat{\sigma}_{n,p}}{f(\widehat{\zeta}_{p,RSS})}$ provides an asymptotic pivotal quantity and a test statistic. This statistic can be used to build confidence intervals and perform test of hypothesis on the population quantiles.

As we explained in Section 1, under URSS the empirical distribution does not converge to $F(x)$ and so the asymptotic properties mentioned for balanced RSS do not hold for URSS. The problem can be overcame by using MI to construct the imputed balanced RSS sample $\mathcal{X}^\diamond = \{X^\diamond_1,\ldots,X^\diamond_k\}$

where $\mathcal{X}_r^{\diamond} = \{X_{(r)1}^{\diamond}, X_{(r)2}^{\diamond}, \ldots, X_{(r)n_r}^{\diamond}, X_{(r)n_r+1}^{\diamond}, \ldots, X_{(r)N}^{\diamond}\}$, for $r = 1, \ldots, k$, is obtained following the Step 1 of the MI Algorithm for URSS data. To find the quantile estimate under the imputed RSS sample, one can order $\{X_{(r)j}^{\diamond}, r = 1, \ldots, k, j = 1, \ldots, N\}$ by ordering the entire RSS set of observation, denoted by $\{X_{(1)}^{\diamond *} \leq \ldots \leq X_{(n^{\diamond})}^{\diamond *}\}$ with $n^{\diamond} = kN$. The EDF of the imputed RSS sample, for multi-cycle RSS is given by

$$\widehat{F}^{\diamond}(x) = \frac{1}{Nk} \sum_{r=1}^{k} \sum_{j=1}^{N} I(X_{(r)j}^{\diamond} \leq x) = \frac{1}{n^{\diamond}} \sum_{i=1}^{n} I(X_{(i)}^{\diamond *} \leq x).$$

The p-th imputed RSS quantile is defined by $\widehat{\zeta}_{p,RSS}^{\diamond} = \inf\{x : \widehat{F}^{\diamond}(x) \geq p\}$, hence the inference of quantile can be used on the imputed URSS similar to the one based on the balanced RSS design.

4.1 Methods of URSS Estimation

In this section, we propose the following methods for quantile and estimation based on URSS data: Following Amiri et al. (2014), the algorithm Boot resamples from each stratum to provide a balanced RSS and estimate the parameter of interest. The algorithm MI imputes the URSS data to form a balanced RSS and MI-Boots is a hybrid of imputation and bootstrap method. Other algorithms include MI, and Boot-Boot that are explained below.

Algorithm MI: Calculate MI estimate of ζ_p

1. Impute each stratum, $\mathcal{X}_i^{\diamond}, i = 1, \ldots, k$.
2. Combine all strata $\mathcal{X}^{\diamond}$.
3. Calculate the quantile for given p, $\widehat{\zeta}_p(\mathcal{X}^{\diamond})$
4. Repeat the Steps 1-3 for M times. $\widehat{\zeta}_{m,p}(\mathcal{X}^{\diamond}), m = 1, \ldots, M$.
5. Consider the average of repeated estimate as the final estimate, $\widehat{\zeta}_{BM,p} = \frac{1}{M} \sum_{m=1}^{M} \widehat{\zeta}_{m,p}(\mathcal{X}^{\diamond})$.

Algorithm ABBI: Calculate ABBI estimate of ζ_p

1. Double Impute each stratum, $\mathcal{X}_i^{\diamond}, i = 1, \ldots, k$.
2. Combine all strata $\mathcal{X}^{\diamond}$.
3. Calculate the quantile for given p, $\widehat{\zeta}_p(\mathcal{X}^{\diamond})$
4. Repeat the Steps 1-3 for M times. $\widehat{\zeta}_{m,p}(\mathcal{X}^{\diamond}), m = 1, \ldots, M$.
5. Consider the average of repeated estimate as the final estimate, $\widehat{\zeta}_{BM,p} = \frac{1}{M} \sum_{m=1}^{M} \widehat{\zeta}_{m,p}(\mathcal{X}^{\diamond})$.

Algorithm MI-Boot: Calculate estimate of ζ_p using a hybrid of MI and Boot

1. Impute each stratum, $\mathcal{X}_i^{\diamond}, i = 1, \ldots, k$.
2. Combine all strata $\mathcal{X}^{\diamond}$, then resample from $\mathcal{X}^{\diamond}$ and denote as $\mathcal{X}^{\diamond *}$.
3. Calculate the quantile for given p, $\widehat{\zeta}_p(\mathcal{X}^{\diamond *})$
4. Repeat the Steps 1-3 for M times. $\widehat{\zeta}_{m,p}(\mathcal{X}^{\diamond *}), m = 1, \ldots, M$.
5. Consider the average of repeated estimate as the final estimate, $\widehat{\zeta}_{BM,p} = \frac{1}{M}\sum_{m=1}^{M} \widehat{\zeta}_{m,p}(\mathcal{X}^{\diamond *})$.

Algorithm Boot: Calculate bootstrap estimate of ζ_p

1. Resample from each stratum $\mathcal{X}_r = \{X_{(r)j}; j = 1, \ldots, n_r\}$ with maximum size, $\max\{n_r\}_{r=1}^{k}$. Denote them as $\mathcal{X}_i^*, i = 1, \ldots, k$.
2. Combine all strata and calculate the quantile for p, $\widehat{\zeta}_p(\mathcal{X}^*)$
3. Repeat the Steps 1-2 for B times, $\widehat{\zeta}_{p,b}(\mathcal{X}^*), b = 1, \ldots, B$.
4. Consider the average of repeated estimate as the final estimate, $\widehat{\zeta}_{B,p} = \frac{1}{B}\sum_{b=1}^{B} \widehat{\zeta}_{b,p}(\mathcal{X}^*)$.

Algorithm Boot-Boot: Calculate bootstrap estimate of ζ_p

1. Resample from each stratum $\mathcal{X}_r = \{X_{(r)j}; j = 1, \ldots, n_r\}$ with maximum size, $\max\{n_r\}_{r=1}^{k}$. Denote them as $\mathcal{X}_i^*, i = 1, \ldots, k$.
2. Combine all strata and resample from $\mathcal{X}^*$ and denote as $\mathcal{X}^{**}$. Calculate the quantile for p, $\widehat{\zeta}_p(\mathcal{X}^{**})$
3. Repeat the Steps 1-2 for B times, $\widehat{\zeta}_{p,b}(\mathcal{X}^{**}), b = 1, \ldots, B$.
4. Consider the average of repeated estimate as the final estimate, $\widehat{\zeta}_{B,p} = \frac{1}{B}\sum_{b=1}^{B} \widehat{\zeta}_{b,p}(\mathcal{X}^{**})$.

5 Numerical Analyses

We evaluate the performance of MI Algorithm for estimating the population mean and quantiles using both real and simulated URSS data. We generate URSS observations following four URSS designs denoted by $D = (n_1, n_2, \ldots, n_k)$ with different sample sizes $n = n_D = \sum_{r=1}^{k} n_r$ when $k = 5$,

$$\begin{aligned}
D_1 &= (4,7,5,6,7) \qquad \text{with} \quad n_{D_1} = 29,\\
D_2 &= (7,4,5,7,6) \qquad \text{with} \quad n_{D_2} = 29,\\
D_3 &= (5,3,6,7,4) \qquad \text{with} \quad n_{D_3} = 25,\\
D_4 &= (6,7,4,5,4) \qquad \text{with} \quad n_{D_4} = 26.
\end{aligned}$$

In order to obtain D_i, we first generate a balanced RSS, $D_O = (7,7,7,7,7)$, and then delete observations from the strata randomly. Clearly, under D_1

and D_4, the strata with large sample sizes are on the left or the right, but D_2 is somewhat symmetric and the strata with large sizes are settled in the central strata in D_3.

We calculate the mean square of the difference (MSD) between estimators and the true values of the parameters in the population $\sum_{\ell=1}^{3000}(\widehat{\theta}_{Alg,\ell} - \theta)^2$ where ℓ is the Monte Carlo replication number and $\widehat{\theta}_{Alg,\ell}$ is the estimate of parameter with different algorithms discussed in Subsection 4.1. To compare the proposed methods, we defined the relative efficiencies (RE) which is as a measure of performance and define by the ratio of the MSD of proposed methods under URSS to the MSD of the mean estimate under (1.2),,$\sum_{\ell=1}^{3000}(\widehat{\theta}_{Alg,\ell} - \theta)^2 / \sum_{\ell=1}^{3000}(\widehat{\theta}_{SRS,\ell} - \theta)^2$ where $\widehat{\theta}_{SRS,\ell}$ is estimate of parameter under SRS. We note that values smaller than 1 show the better performance of estimation based on RSS relative to its counterpart, the pooled sample mean. We used $M = 400$ imputations and the replicated the experiment 3000 times.

5.1 Mercury Level in Fish

We evaluate the performance of MI Algorithm for estimating the population mean and quantiles based on URSS data generated from a Fishery dataset studied in Nourmohammadi et al. (2015). This dataset contains mercury contamination levels along with the weights and lengths of 3033 of Walleye fish. Sander vitreus (Walleye) fish caught in Minnesota is a freshwater perciform fish native to most of Canada and to the Northern United States.

Information on fish intake rates is included in the Exposure Factors Handbook USEPA (2011) and is often utilized in multiplicative risk models to provide estimates of risk to human health as a result of exposure to chemicals. Christophi et al. (2005) considered the distribution of a hazard index for a specified chemical in consumed fish. This index requires estimates of the mean and quantiles of the concentration of a chemical contaminant such as mercury in fish along with estimates of ingestion rate of fish and the chemical-specific reference dose.

We use the fish dataset to study the performance of the the proposed methods for estimating the quantiles of the mercury levels. As explained in Nourmohammadi et al. (2015), measuring the mercury level in fish body is a costly and time consuming process. To obtain better samples from the fish population one can use a RSS design using length or weight of fish to rank.

To generate an URSS, we treat the 3033 records as our population and compute the true mean and quantiles to use in the MSD. We consider both perfect and imperfect rankings. Under perfect ranking, we use mercury levels to perform the ranking which the consistency, $F(t) = \frac{1}{k}\sum_{r=1}^{k} F_{(r)}(t)$, is held. We note that imperfect ranking is only provided for comparison purpose. For imperfect ranking , rankings is performed using the fish weight. The correlation coefficient between the mercury level and the fish weight is about 0.4. The Table 2 displays the REs for estimating the mean mercury level in fish body. The proposed methods outperform the pooled sample mean under the proposed designs.

Table 2 *The relative efficiencies of proposed methods under URSS for estimating the average mercury in fish body compared with the pooled sample mean with SRS. Note that values less than 1 are desirable.*

		RE			RE	
Design	Method	perfect	imperfect	Design	perfect	imperfect
D_1	MI	0.379	0.896	D_2	0.386	0.803
	ABBI	0.379	0.893		0.386	0.806
	MI-Boot	0.381	0.897		0.390	0.806
	Boot	0.380	0.894		0.386	0.805
	Boot-Boot	0.383	0.902		0.384	0.811
D_3	MI	0.352	0.839	D_4	0.375	0.891
	ABBI	0.352	0.841		0.375	0.891
	MI-Boot	0.354	0.843		0.375	0.895
	Boot	0.352	0.842		0.376	0.891
	Boot-Boot	0.353	0.842		0.381	0.895

Next, we compare the estimated quantiles URSS quantiles with the true quantiles of the population. The quantiles are calculated for different values of p. We report their relative efficiencies (RE), which are the ratios of the MSD of our proposed methods under RSS to the MSD of quantile estimate under SRS. The upper panel of Tables 3 displays the REs under perfect ranking and the last column shows the average of REs. The estimated REs that are less than 1 indicate that the URSS based estimators are more accurate than their SRS counterparts. We observe that the REs of RSS based estimators are less than 1 and stay stable over all the value of p. The average of REs shows that MI-Boot and Boot-Boot have a better performance than other estimators. The second panel of Table 3 presents the performance of the methods under imperfect ranking. The results show that the proposed method under RSS behave better than SRS and Boot-Boot provides more accurate estimates.

5.2 Simulated data

To perform simulation studies, we generate SRS, perfect and imperfect URSS data from $N(0,1)$ (symmetric) and $Exp(1)$ (skewed) distributions us-

Table 3 *The relative efficiencies of proposed methods under perfect URSS for estimating quantiles of the mercury level in fish body compared with their corresponding SRS estimators.*

Under perfect ranking											
		Quantile									
Design	Method	10	20	30	40	50	60	70	80	90	avg.
D_1	MI	0.794	0.593	0.595	0.447	0.460	0.468	0.501	0.499	0.612	0.552
	ABBI	0.782	0.583	0.591	0.444	0.456	0.466	0.500	0.499	0.612	0.548
	MI-Boot	0.755	0.473	0.441	0.361	0.350	0.362	0.397	0.402	0.539	0.453
	Boot	0.766	0.524	0.507	0.408	0.407	0.420	0.446	0.439	0.528	0.493
	Boot-Boot	0.768	0.458	0.419	0.346	0.338	0.352	0.389	0.376	0.521	0.440
D_2	MI	0.617	0.599	0.575	0.507	0.512	0.494	0.481	0.566	0.608	0.551
	ABBI	0.619	0.593	0.561	0.502	0.505	0.490	0.479	0.566	0.609	0.547
	MI-Boot	0.576	0.473	0.436	0.412	0.393	0.378	0.384	0.442	0.523	0.446
	Boot	0.545	0.536	0.525	0.474	0.466	0.440	0.423	0.484	0.534	0.491
	Boot-Boot	0.559	0.451	0.422	0.397	0.385	0.363	0.366	0.422	0.505	0.430
D_3	MI	0.622	0.643	0.594	0.509	0.475	0.452	0.522	0.544	0.710	0.563
	ABBI	0.614	0.632	0.573	0.504	0.468	0.448	0.517	0.538	0.688	0.553
	MI-Boot	0.597	0.512	0.458	0.407	0.366	0.349	0.400	0.436	0.642	0.463
	Boot	0.597	0.605	0.553	0.474	0.420	0.398	0.451	0.485	0.677	0.517
	Boot-Boot	0.600	0.495	0.448	0.394	0.350	0.335	0.390	0.425	0.647	0.453
D_4	MI	0.526	0.503	0.492	0.470	0.506	0.513	0.563	0.605	0.743	0.546
	ABBI	0.528	0.503	0.489	0.465	0.498	0.508	0.551	0.598	0.718	0.539
	MI-Boot	0.514	0.406	0.360	0.382	0.414	0.408	0.441	0.498	0.658	0.453
	Boot	0.509	0.438	0.414	0.436	0.472	0.480	0.513	0.564	0.705	0.503
	Boot-Boot	0.520	0.387	0.342	0.372	0.404	0.400	0.433	0.486	0.653	0.444
Under perfect imranking											
Dist.	Method	10	20	30	40	50	60	70	80	90	avg.
D_1	MI	1.059	0.891	0.646	0.529	0.536	0.567	0.614	0.697	0.936	0.719
	ABBI	1.054	0.872	0.640	0.525	0.530	0.565	0.613	0.697	0.936	0.719
	MI-Boot	1.066	0.791	0.542	0.458	0.436	0.486	0.551	0.628	0.873	0.647
	Boot	1.033	0.816	0.579	0.492	0.481	0.517	0.558	0.630	0.839	0.660
	Boot-Boot	1.076	0.765	0.523	0.449	0.427	0.469	0.529	0.608	0.865	0.634
D_2	MI	1.0061	0.827	0.698	0.650	0.537	0.594	0.643	0.767	0.989	0.745
	ABBI	1.010	0.814	0.687	0.648	0.532	0.590	0.640	0.769	0.987	0.741
	MI-Boot	1.008	0.734	0.579	0.564	0.449	0.501	0.555	0.691	0.925	0.667
	Boot	0.960	0.747	0.632	0.613	0.492	0.539	0.572	0.688	0.896	0.682
	Boot-Boot	1.007	0.701	0.561	0.552	0.437	0.485	0.528	0.660	0.922	0.650
D_3	MI	1.184	0.916	0.692	0.619	0.550	0.511	0.588	0.723	0.887	0741
	ABBI	1.184	0.906	0.674	0.612	0.534	0.507	0.581	0.715	0.872	0.731
	MI-Boot	1.171	0.821	0.582	0.532	0.449	0.428	0.507	0.656	0.850	0.662
	Boot	1.150	0.864	0.641	0.579	0.490	0.463	0.531	0.669	0.841	0.692
	Boot-Boot	1.171	0.801	0.561	0.523	0.430	0.413	0.490	0.635	0.856	0.653
D_4	MI	0.888	0.693	0.593	0.546	0.597	0.633	0.667	0.776	1.045	0.715
	ABBI	0.892	0.693	0.588	0.543	0.590	0.622	0.657	0.762	1.021	0.707
	MI-Boot	0.891	0.606	0.486	0.485	0.503	0.530	0.599	0.699	1.001	0.644
	Boot	0.851	0.606	0.519	0.519	0.551	0.590	0.635	0.733	1.001	0.667
	Boot-Boot	0.906	0.578	0.457	0.468	0.491	0.519	0.588	0.690	1.001	0.633

Note that values less than 1 are desirable.

ing designs $D_1,\dots,D_4$. We then implement our proposed approaches and obtain corresponding estimators under each method. The results appear in Tables 4 and 5. One can easily observe that the proposed methods in this paper outperform SRS and Boot-IMP provides more accurate estimates. We also considered the performance under the imperfect ranking. There are several techniques of generating imperfect RSS, see Frey et al. (2007), Vock and Balakrishnan (2011), among others. Here, we considered a variant of the fraction of neighbors method, Vock and Balakrishnan (2011). Let $F_{[i]}$ be the distribution of imperfect RSS, which is the mixture $F_{(i)}$, $F_{(i-1)}$ and $F_{(i+1)}$, $F_{[i]} = \frac{\lambda}{2}F_{(i-1)} + (1-\lambda)F_{(i)} + \frac{\lambda}{2}F_{(i+1)}$, where λ is the fraction of incorrectly chosen statistics and $F_{(0)} := F_{(1)}$ and $F_{(k+1)} := F_{(k)}$. For a more realistic model, Amiri et al. (2016) suggested λ to depend on i, in which case the probability of incorrect ranking varies with the order. It would be reasonable to assume that it is more likely to be make more error the middle rankings than the extreme ones. Hence, we define $F_{[i]} = \lambda_{1i}F_{(i-1)} + \lambda_{2i}F_{(i)} + \lambda_{3i}F_{(i+1)}$ and consider the following weights, respectively.

$$\begin{aligned}
(\lambda_{11},\lambda_{21},\lambda_{31}) &= (0,1/2,1/2), \qquad (5.1)\\
(\lambda_{12},\lambda_{22},\lambda_{32}) &= (1/4,1/2,1/4),\\
(\lambda_{13},\lambda_{23},\lambda_{33}) &= (1/3,1/3,1/3),\\
(\lambda_{14},\lambda_{24},\lambda_{34}) &= (1/4,1/2,1/4),\\
(\lambda_{15},\lambda_{25},\lambda_{35}) &= (1/2,1/2,0).
\end{aligned}$$

The results are presented in Tables 4 and 5. The results support the discussion in the previous section, i.e., MI-Boot and Boot-Boot provide the best performance.

Table 4 *The relative efficiencies of proposed methods under perfect URSS for estimating quantiles of $N(0,1)$ compared with their corresponding SRS estimators. Note that values less than 1 are desirable.*

Under perfect ranking											
		Quantile									
Design	Method	10	20	30	40	50	60	70	80	90	avg.
D_1	MI	0.765	0.682	0.528	0.485	0.501	0.464	0.494	0.516	0.556	0.554
	ABBI	0.748	0.674	0.523	0.483	0.498	0.461	0.493	0.516	0.556	0.550
	MI-Boot	0.663	0.539	0.420	0.393	0.394	0.363	0.398	0.407	0.446	0.44
	Boot	0.719	0.603	0.458	0.443	0.456	0.420	0.442	0.453	0.462	0.495
	Boot-Boot	0.664	0.522	0.401	0.381	0.378	0.355	0.388	0.389	0.420	0.433
D_2	MI	0.734	0.624	0.571	0.534	0.503	0.508	0.526	0.584	0.594	0.575
	ABBI	0.735	0.614	0.560	0.526	0.496	0.505	0.526	0.584	0.594	0.571
	MI-Boot	0.596	0.494	0.469	0.416	0.396	0.390	0.414	0.464	0.514	0.461
	Boot	0.610	0.565	0.528	0.490	0.460	0.458	0.460	0.511	0.534	0.512
	Boot-Boot	0.559	0.465	0.459	0.411	0.383	0.378	0.393	0.440	0.502	0.443
D_3	MI	0.711	0.708	0.600	0.524	0.455	0.448	0.502	0.574	0.768	0.587
	ABBI	0.701	0.704	0.581	0.517	0.446	0.444	0.498	0.572	0.747	0.578
	MI-Boot	0.625	0.583	0.483	0.417	0.345	0.360	0.386	0.471	0.668	0.482
	Boot	0.660	0.682	0.565	0.483	0.399	0.398	0.433	0.516	0.729	0.540
	Boot-Boot	0.609	0.565	0.470	0.403	0.334	0.348	0.373	0.459	0.662	0.469
D_4	MI	0.556	0.495	0.467	0.438	0.504	0.477	0.520	0.627	0.730	0.534
	ABBI	0.556	0.494	0.465	0.430	0.496	0.472	0.509	0.620	0.712	0.528
	MI-Boot	0.470	0.386	0.367	0.347	0.397	0.385	0.423	0.521	0.650	0.438
	Boot	0.500	0.422	0.406	0.401	0.462	0.449	0.482	0.583	0.692	0.488
	Boot-Boot	0.461	0.365	0.349	0.335	0.386	0.377	0.419	0.505	0.644	0.426
Under imperfect ranking											
Design	Method	10	20	30	40	50	60	70	80	90	avg.
D_1	MI	1.042	0.848	0.691	0.598	0.522	0.542	0.633	0.739	0.841	0.717
	ABBI	1.016	0.832	0.684	0.593	0.520	0.540	0.628	0.739	0.841	0.710
	MI-Boot	0.969	0.737	0.595	0.521	0.444	0.466	0.532	0.644	0.762	0.630
	Boot	0.977	0.777	0.635	0.554	0.480	0.496	0.562	0.655	0.753	0.654
	Boot-Boot	0.954	0.723	0.575	0.501	0.428	0.448	0.510	0.608	0.752	0.611
D_2	MI	1.038	0.912	0.685	0.627	0.561	0.589	0.630	0.725	0.919	0.742
	ABBI	1.032	0.902	0.677	0.617	0.553	0.585	0.627	0.726	0.918	0.737
	MI-Boot	0.957	0.825	0.608	0.522	0.464	0.494	0.536	0.642	0.824	0.652
	Boot	0.947	0.846	0.640	0.572	0.506	0.531	0.561	0.652	0.812	0.674
	Boot-Boot	0.923	0.806	0.585	0.505	0.448	0.475	0.508	0.616	0.801	0.629
D_3	MI	1.005	0.829	0.699	0.607	0.515	0.555	0.606	0.706	0.924	0.716
	ABBI	0.995	0.811	0.692	0.597	0.507	0.548	0.598	0.698	0.907	0.705
	MI-Boot	0.952	0.744	0.630	0.524	0.426	0.469	0.507	0.629	0.866	0.638
	Boot	0.950	0.781	0.668	0.569	0.471	0.504	0.541	0.649	0.868	0.666
	Boot-Boot	0.942	0.730	0.620	0.509	0.414	0.452	0.487	0.609	0.850	0.666
D_4	MI	0.889	0.624	0.603	0.531	0.520	0.556	0.686	0.724	1.051	0.687
	ABB()	0.892	0.623	0.600	0.525	0.511	0.548	0.672	0.708	1.021	0.677
	MI-Boot	0.810	0.556	0.514	0.446	0.439	0.480	0.600	0.646	0.975	0.607
	Boot	0.799	0.558	0.539	0.485	0.482	0.525	0.647	0.687	0.997	0.635
	Boot-Boot	0.788	0.531	0.485	0.426	0.426	0.466	0.592	0.631	0.974	0.591

Table 5 *The relative efficiencies of proposed methods under perfect URSS for estimating quantiles of $Exp(1)$ compared with their corresponding SRS estimators. Note that values less than 1 are desirable.*

Under perfect ranking											
		Quantile									
Design	Method	10	20	30	40	50	60	70	80	90	avg.
D_1	MI	0.758	0.665	0.515	0.463	0.465	0.511	0.493	0.511	0.591	0.552
	ABB	0.738	0.657	0.511	0.460	0.460	0.509	0.493	0.511	0.591	0.547
	MI-Boot	0.881	0.619	0.421	0.383	0.363	0.418	0.390	0.410	0.492	0.486
	Boot	0.780	0.625	0.450	0.416	0.416	0.473	0.440	0.460	0.514	0.508
	Boot-Boot	0.971	0.650	0.424	0.385	0.358	0.411	0.382	0.396	0.476	0.494
D_2	MI	0.617	0.631	0.576	0.555	0.482	0.497	0.491	0.547	0.612	0.556
	ABBI	0.625	0.634	0.565	0.548	0.475	0.492	0.491	0.547	0.611	0.554
	MI-Boot	0.727	0.621	0.501	0.464	0.386	0.394	0.378	0.448	0.490	0.489
	Boot	0.578	0.607	0.545	0.523	0.442	0.450	0.429	0.487	0.533	0.510
	Boot-Boot	0.796	0.631	0.505	0.470	0.387	0.386	0.370	0.431	0.480	0.495
D_3	MI	0.666	0.633	0.565	0.547	0.439	0.435	0.484	0.552	0.714	.559
	ABBI	0.670	0.633	0.551	0.537	0.436	0.430	0.480	0.551	0.705	0.554
	MI-Boot	0.789	0.592	0.493	0.441	0.343	0.355	0.376	0.484	0.629	0.500
	Boot	0.657	0.621	0.543	0.509	0.392	0.395	0.415	0.497	0.678	0.523
	Boot-Boot	0.836	0.606	0.502	0.440	0.342	0.351	0.365	0.472	0.631	0.505
D_4	MI	0.522	0.410	0.479	0.489	0.536	0.510	0.548	0.587	0.828	0.545
	ABBI	0.521	0.411	0.480	0.486	0.524	0.506	0.535	0.580	0.814	0.539
	MI-Boot	0.644	0.404	0.437	0.421	0.449	0.427	0.448	0.505	0.736	0.496
	Boot	0.542	0.377	0.438	0.451	0.496	0.479	0.500	0.545	0.796	0.513
	Boot-Boot	0.741	0.417	0.435	0.418	0.452	0.421	0.444	0.496	0.731	0.506
Under imperfect ranking											
Design	Method	10	20	30	40	50	60	70	80	90	
D_1	MI	1.282	1.022	0.706	0.563	0.497	0.543	0.583	0.664	0.896	0.750
	ABBI	1.289	1.010	0.698	0.558	0.492	0.540	0.581	0.663	0.898	0.747
	MI-Boot	1.560	1.080	0.680	0.516	0.434	0.463	0.482	0.565	0.762	0.726
	Boot	1.390	1.009	0.668	0.530	0.464	0.498	0.514	0.584	0.772	0.714
	Boot-Boot	1.749	1.151	0.697	0.524	0.428	0.449	0.456	0.527	0.700	0.742
D_2	MI	1.187	0.958	0.684	0.641	0.572	0.581	0.602	0.699	0.919	0.760
	ABBI	1.206	0.956	0.677	0.633	0.562	0.576	0.599	0.698	0.921	0.758
	MI-Boot	1.433	1.040	0.668	0.592	0.495	0.485	0.497	0.600	0.792	0.733
	Boot	1.245	0.947	0.654	0.615	0.533	0.528	0.530	0.619	0.806	0.719
	Boot-Boot	1.600	1.096	0.678	0.599	0.499	0.473	0.468	0.555	0.753	0.746
D_3	MI	1.249	0.930	0.709	0.588	0.595	0.510	0.628	0.651	0.893	0.750
	ABBI	1.270	0.931	0.696	0.581	0.586	0.503	0.622	0.639	0.873	0.744
	MI-Boot	1.510	0.986	0.679	0.513	0.507	0.429	0.534	0.566	0.806	0.725
	Boot	1.308	0.940	0.686	0.555	0.548	0.464	0.567	0.588	0.829	0.720
	Boot-Boot	1.632	1.024	0.694	0.518	0.495	0.419	0.503	0.537	0.783	0.733
D_4	MI	0.913	0.769	0.725	0.571	0.556	0.573	0.647	0.756	0.960	0.718
	ABBI	0.919	0.771	0.725	0.569	0.549	0.566	0.637	0.746	0.935	0.713
	MI-Boot	1.117	0.801	0.728	0.529	0.478	0.498	0.560	0.679	0.874	0.696
	Boot	0.972	0.727	0.678	0.534	0.514	0.542	0.608	0.723	0.906	0.689
	Boot-Boot	1.270	0.834	0.739	0.526	0.470	0.487	0.543	0.654	0.860	0.709

6 Conclusion

This article draws on the imputation literature with minimal distributional assumptions in order to transform URSS data to a balanced RSS. This transformation allows one to apply standard techniques of bootstrap, estimation and testing that are available for balanced ranked set samples to the completed dataset. To this end, we first study MI of a SRS, prove that its EDF converges to the population CDF under MI, obtain the variance of the imputed mean, and the expected value of the variance estimator. We extend these results to MI for URSS data and provide different methods for estimating the population quantiles. We use a real data application and study the performance of our proposed methods in estimating the mean and the quantiles of the mercury level in a fish population using both perfect and imperfect unbalanced ranked set sampling designs. We consider a hybrid method based on the bootstrap and imputing URSS. The overall recommendations are the hybrid estimates based on imputation and bootstrap (MI-Boot) and Boot-Boot.

Acknowledgements

Mohammad Jafari Jozani acknowledge the research supports of the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Amiri, S., Jafari Jozani, M., & Modarres, R. (2014). Resampling Unbalanced Ranked Set Samples With Applications in Testing Hypothesis About the Population Mean. Journal of Agricultural, Biological, and Environmental Statistics, 19(1), 1-17.

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

Chen, Z. (2000). On the ranked-set sample quantiles and their application. J. Stat. Plan. Infe. 83, 125-135.

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

Christophi, C. A., & Modarres, R. (2005). Approximating the distribution function of risk. Computational statistics & data analysis, 49(4), 1053-1067.

Carpenter J.R. and Kenward M.G. (2012). Multiple imputation and its application. John Wiley and Sons Ltd. publication. 2nd Edition.

Demirtas H, Arguelles LM, Chung H, Hedeker D. (2007) On the performance of bias-reduction techniques for variance estimation in approximate Bayesian bootstrap imputation. Computational Statistics and Data Analysis. 51(8):4064-4068.

Frey, J., Ozturk, O., & Deshpande, J.V. (2007). Nonparametric tests for perfect judgment rankings. Journal of the American Statistical Association, 102(478), 708-717.

Hatefi, A. and Jafari Jozani, M. (2015). Proportion estimation based on a partially rank ordered set sample with multiple concomitants in a breast cancer study. Statistical Method in Medical Research. To appear.

Jafari Jozani, M., & Ahmadi, J. (2014). On uncertainty and information properties of ranked set samples. Information Sciences, 264, 291-301.

Kim, J. K. (2002). A note on approximate Bayesian bootstrap imputation. Biometrika 89, 470-477.

Li P., Stuart E.A. and Allison D.B. (2015). Multiple Imputation: A Flexible Tool for Handling Missing Data. JAMA Guide to Statistics and Methods, 314(18), 1966–1967.

McIntyre, G.A. (1952). A method for unbiased selective sampling, using ranked sets. Australian Journal of Agriculture Research, 3, 385-390.

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

Nourmohammadi, M., Jafari Jozani, M., & Johnson, B. C. (2014). Confidence intervals for quantiles in finite populations with randomized nomination sampling. Computational Statistics & Data Analysis, 73, 112-128.

Nourmohammadi, M., Jafari Jozani, M. J., & Johnson, B. C. (2015). Distribution-free tolerance intervals with nomination samples: Applications to Mercury contamination in fish. Statistical Methodology, 26, 16-33.

Ozturk, O., & Jafari Jozani, M. (2014). Inclusion probabilities in partially rank ordered set sampling. Computational Statistics & Data Analysis, 69, 122–132.

Parzen, M., Lipsitz, S.R., Fitzmaurice, G.M. (2005). A note on reducing the bias of the approximate Bayesian bootstrap imputation variance estimator. Biometrika 92,971–974.

Rubin, D. B. (1996). Multiple Imputation after 18+ years (with discussion). J. Amer. Statist. Assoc. 91, 473-520.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. Wiley: New York.

Rubin, D. B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. J. Amer. Statist. Assoc. 81 366–374.

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

Serfling, R. J. (2009). Approximation theorems of mathematical statistics. John Wiley & Sons.

Schafer J.L. and Olsen K.M. (1998). Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. Multivariate Behavioral Research, 33 (4), 545–571.

U.S. EPA. (2011) Exposure Factors Handbook 2011 Edition (Final). U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-09/052F.

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

Wolfe, D.A. (2012). Ranked Set Sampling: Its Relevance and Impact on Statistical Inference. ISRN Probability and Statistics.

7 Appendix

7.1 Proof of Proposition 1

For the ease in notation, we drop the index m from $X^{\diamond}_{i,m}$ and simply work with $X^{\diamond}_i$. Also for convenience, we represent $\mathcal{X}^{\diamond}_i$ as $\mathcal{X}^{\diamond}_i = \{X^{\diamond}_{n+1,i}, \ldots, X^{\diamond}_{N,i}\}$

instead $\mathcal{X}_i^\diamond = \{X_{1,i}^\diamond, \ldots, X_{\mathfrak{r},i}^\diamond\}$ which was defined in Proposition 1. Using the Glivenko-Cantelli Theorem, we have

$$\|\widehat{F}_n(t) - F(t)\|_\infty = 0. \tag{7.1}$$

Using the imputed observations, we can show that

$$\begin{aligned}
|\widehat{F}_N^\diamond(t) - \widehat{F}_n(t)| &= \left|\frac{1}{N}(\sum_{i=1}^n I(X_i \le t) + \sum_{i=n+1}^N I(X_i^\diamond \le t)) - \frac{1}{n}\sum_{i=1}^n I(X_i \le t)\right| \\
&= \left|(\frac{n}{N} - 1)\frac{1}{n}\sum_{i=1}^n I(X_i \le t) + (1 - \frac{n}{N})\frac{1}{\mathfrak{r}}\sum_{i=n+1}^N I(X_i^\diamond \le t)\right|.
\end{aligned}$$

As $\lim_{\mathfrak{r}\longrightarrow\infty} \frac{1}{\mathfrak{r}} \sum_{i=n+1}^N I(X^\diamond \le t) = \widehat{F}_n(t)$, it follows that $\lim_{\mathfrak{r}\longrightarrow\infty} |\widehat{F}_N^\diamond(t) - \widehat{F}_n(t)| = |\widehat{F}_n(t) - \widehat{F}_n(t)| = 0$. The result follows uisng the inequality $|\widehat{F}_N^\diamond(t) - F(t)| \le |\widehat{F}_N^\diamond(t) - \widehat{F}_n(t)| + |\widehat{F}_n(t) - F(t)|$.

Proof of Proposition 2

Similar to the proof of Proposition 1, we drop the index m from $X_{i,m}^\diamond$ and simply work with $X_i^\diamond$. Using the Glivenko-Cantelli Theorem, equation (7.1), and the imputed observations, we can show that

$$\begin{aligned}
&\left|\widehat{F}_{n+\mathfrak{r}}^\diamond(t) - \widehat{F}_n(t)\right| \\
= &\left|\frac{1}{n+\mathfrak{r}}(\sum_{i=1}^n I(X_i \le t) + \sum_{i=n+1}^{n+\mathfrak{r}} I(X_i^\diamond \le t)) - \frac{1}{n}\sum_{i=1}^n I(X_i \le t)\right| \\
= &\left|(\frac{n}{N} - 1)\frac{1}{n}\sum_{i=1}^n I(X_i \le t) + (1 - \frac{n}{N})\frac{1}{\mathfrak{r}}\sum_{i=n+1}^{n+\mathfrak{r}} I(X_i^\diamond \le t) \pm (1 - \frac{n}{N})\frac{1}{n}\sum_{i=1}^n I(X_i^* \le t)\right| \\
= &\left|(\frac{n}{N} - 1)\frac{1}{n}\sum_{i=1}^n I(X_i \le t) + (1 - \frac{n}{N})\frac{1}{n}\sum_{i=1}^n I(X_i^* \le t) + \right. \\
&\left.(1 - \frac{n}{N})\frac{1}{\mathfrak{r}}\sum_{i=n+1}^{n+\mathfrak{r}} I(X_i^\diamond \le t) - (1 - \frac{n}{N})\frac{1}{n}\sum_{i=1}^n I(X_i^* \le t)\right| \\
\le &\left|(\frac{n}{N} - 1)\frac{1}{n}\sum_{i=1}^n I(X_i \le t) + (1 - \frac{n}{N})\frac{1}{n}\sum_{i=1}^n I(X_i^* \le t)\right| \\
&+\left|(1 - \frac{n}{N})\frac{1}{\mathfrak{r}}\sum_{i=n+1}^{n+\mathfrak{r}} I(X_i^\diamond \le t) - (1 - \frac{n}{N})\frac{1}{n}\sum_{i=1}^n I(X_i^* \le t)\right| \\
= &A + B.
\end{aligned}$$

It follows that

$$\begin{aligned}\lim_{n\to\infty}(A) &= \left|(\frac{n}{N}-1)\frac{1}{n}\sum_{i=1}^{n}I(X_i\le t)+(1-\frac{n}{N})\frac{1}{n}\sum_{i=1}^{n}I(X_i^*\le t)\right| \\ &= \lim_{n\to\infty}(1-\frac{n}{N})\lim_{n\to\infty}(-\widehat{F}_n(x)+\widehat{F}_n^*(x))=0, \qquad (7.2)\end{aligned}$$

and

$$\begin{aligned}\lim_{\mathfrak{r}\to\infty}B &= \lim_{\mathfrak{r}\to\infty}\left|(1-\frac{n}{N})\frac{1}{\mathfrak{r}}\sum_{i=n+1}^{n+\mathfrak{r}}I(X_i^\diamond\le t)-(1-\frac{n}{N})\frac{1}{n}\sum_{i=1}^{n}I(X_i^*\le t)\right| \\ &= (1-\frac{n}{N})\lim_{\mathfrak{r}\to\infty}(\widehat{F}_{\mathfrak{r}}^\diamond(x)-\widehat{F}_n^*(x))=0. \qquad (7.3)\end{aligned}$$

Using (7.2) and (7.3), $\lim_{n\to\infty}\lim_{\mathfrak{r}\to\infty}|\widehat{F}_N^\diamond(t)-\widehat{F}_n(t)|=0$ and by considering (7.1), the results follow by a similar argument as in the proof of Proposition 1.

7.2 Proof of Proposition 3

Since the M imputed estimators $\{\widehat{\mu}_m\}_{m=1,\ldots,M}$ are identically distributed, one can easily show that

$$Var(\widehat{\mu})=\frac{1}{M}Var(\widehat{\mu}_m)+(1-\frac{1}{M})Cov(\widehat{\mu}_m,\widehat{\mu}_{m'}), \qquad (7.4)$$

where $m\le m'$. Define $\mathcal{X}=\{X_1,\ldots,X_n\}$, hence

$$\begin{aligned}Var(\widehat{\mu}_m) &= Var(E(\frac{n\bar{X}+\mathfrak{r}\bar{X}^\diamond}{N}|\mathcal{X}))+E(Var(\frac{n\bar{X}+\mathfrak{r}\bar{X}^\diamond}{N}|\mathcal{X})) \\ &= Var(\bar{X})+E(\frac{\mathfrak{r}^2}{N^2}Var(\bar{X}^\diamond)) \\ &= \frac{\sigma^2}{n}+\frac{\mathfrak{r}^2}{N^2}E(\frac{\mathfrak{r}-1}{\mathfrak{r}}\frac{S^2}{\mathfrak{r}}) \\ &= \frac{\sigma^2}{n}+\frac{\mathfrak{r}-1}{N^2}\sigma^2. \qquad (7.5)\end{aligned}$$

$$\begin{aligned}Cov(\widehat{\mu}_m,\widehat{\mu}_{m'}) &= Cov(\frac{n\bar{X}+\mathfrak{r}\bar{X}_m^\diamond}{N},\frac{n\bar{X}+\mathfrak{r}\bar{X}_{m'}^\diamond}{N}) \\ &= \frac{1}{N^2}(Cov(n\bar{X},n\bar{X})+2Cov(n\bar{X},\mathfrak{r}\bar{X}_m^\diamond)+Cov(\mathfrak{r}X_m^\diamond,\mathfrak{r}X_{m'}^*)) \\ &= \frac{1}{N^2}(n^2\frac{\sigma^2}{n}+2n\mathfrak{r}\frac{\sigma^2}{n}+\mathfrak{r}^2\frac{\sigma^2}{n}) \\ &= \frac{\sigma^2}{n}. \qquad (7.6)\end{aligned}$$

By substituting (7.5) and (7.6) in (7.4), one obtains (2.3). In order to prove (2.4), note that

$$\widehat{U}_m = \frac{1}{N(N-1)}(\sum_{i=1}^{n} X_i^2 + \sum_{i=n+1}^{N} X_{i,m}^{2\diamond} - N\widehat{\mu}_m^2).$$

It follows that

$$\begin{aligned}
E(\widehat{U}_m) &= \frac{1}{N(N-1)}(E(\sum_{i=1}^{n} X_i^2) + E(\sum_{i=n+1}^{N} X_{i,m}^{2\diamond}) - NE(\widehat{\mu}_m^2)) \\
&= \frac{1}{N(N-1)}(NE(X_i^2) - N(Var(\widehat{\mu}_m) + E(\widehat{\mu}_m)^2)) \\
&= \frac{1}{(N-1)}(E(X_i^2) - \mu^2 - Var(\widehat{\mu}_m)) \\
&= \frac{1}{(N-1)}(\sigma^2 - \frac{\sigma^2}{n} - \frac{\mathfrak{r}-1}{N^2}\sigma^2) \\
&= \frac{1}{(N-1)}(1 - \frac{1}{n} - \frac{\mathfrak{r}-1}{N^2})\sigma^2. \qquad (7.7)
\end{aligned}$$

Since $\widehat{B} = (M-1)^{-1}(\sum_{m=1}^{M} \widehat{\mu}_m - M\widehat{\mu}^2)$ it follows that

$$\begin{aligned}
E(\widehat{B}) &= \frac{1}{(M-1)}(\sum_{m=1}^{M} E(\widehat{\mu}_m) - ME(\widehat{\mu}^2)) \\
&= \frac{M}{M-1}(Var(\widehat{\mu}_m) + E(\widehat{\mu}_m)^2 - (Var(\widehat{\mu}) + E(\widehat{\mu})^2)) \\
&= \frac{M}{M-1}(Var(\widehat{\mu}_m) - Var(\widehat{\mu})) \\
&= \frac{M}{M-1}(\frac{\sigma^2}{n} + \frac{\mathfrak{r}-1}{N^2}\sigma^2 - \frac{\sigma^2}{n} - \frac{1}{M}(\frac{\mathfrak{r}-1}{N^2}\sigma^2)) \\
&= \frac{M}{M-1}(\frac{\mathfrak{r}-1}{N^2}\sigma^2 - \frac{1}{M}(\frac{\mathfrak{r}-1}{N^2}\sigma^2)) \\
&= \frac{\mathfrak{r}-1}{N^2}\sigma^2, \qquad (7.8)
\end{aligned}$$

by substituting (7.7) and (7.8) in (2.2) can establish (2.4).

7.3 Proof of Proposition 4

Consider the imputed URSS, and note that $\widehat{\mu}_m = \frac{1}{k}\sum_{r=1}^{k} \widehat{\mu}_{(r),m}$ and

$$Var(\widehat{\mu}_m) = \frac{1}{k^2}\sum_{r=1}^{k} Var(\widehat{\mu}_{(r),m}) \qquad (7.9)$$

where $Var(\widehat{\mu}_{(r)m})$ is given in Proposition 3. One can readily obtain (3.4). In order to prove (3.5), using (3.3)

$$E(U_m) = \frac{1}{k^2}\sum_{r=1}^{k}\frac{1}{(N-1)}(1-\frac{1}{n_r}-\frac{\mathfrak{r}_r-1}{N^2}).$$

We also have

$$E(\widehat{B}) = \frac{M}{M-1}(Var(\mu_m)-Var(\widehat{\mu})).$$

where $Var(\widehat{\mu}_m)$ is given in (7.9) and $Var(\widehat{\mu})$ is given in (3.5).

Department of Civil, Geologic, and Mining Engineering, Polytechnique Montreal, QC, Canada,E-mail: saeid.amiri1@gmail.com
Department of Statistics, University of Manitoba, Winnipeg, MB, Canada, R3T 2N2,
E-mail: m_jafari_jozani@umanitoba.ca, and Department of Statistics, The George Washington University, Washington DC, USA ,
E-mail: reza@gwu.edu