Communications for Statistical Applications and Methods
2018, Vol. 25, No. 1, 1–13

https://doi.org/10.29220/CSAM.2018.25.1.001
Print ISSN 2287-7843 / Online ISSN 2383-4757

On inference of multivariate means under ranked set sampling

Haresh Rochani[1,a], Daniel F. Linder[b], Hani Samawi[a], Viral Panchal[b]

[a]Department of Biostatistics, Georgia Southern University, USA;
[b]Department of Biostatistics, Augusta University, USA

Abstract

In many studies, a researcher attempts to describe a population where units are measured for multiple outcomes, or responses. In this paper, we present an efficient procedure based on ranked set sampling to estimate and perform hypothesis testing on a multivariate mean. The method is based on ranking on an auxiliary covariate, which is assumed to be correlated with the multivariate response, in order to improve the efficiency of the estimation. We showed that the proposed estimators developed under this sampling scheme are unbiased, have smaller variance in the multivariate sense, and are asymptotically Gaussian. We also demonstrated that the efficiency of multivariate regression estimator can be improved by using Ranked set sampling. A bootstrap routine is developed in the statistical software R to perform inference when the sample size is small. We use a simulation study to investigate the performance of the method under known conditions and apply the method to the biomarker data collected in China Health and Nutrition Survey (CHNS 2009) data.

Keywords: multivariate mean, ranked set sampling, hypothesis testing, regression estimator

1. Introduction

As the complexity and cost of biological experiments has grown considerably in recent years, partly due to technological advances (high throughput technologies and more), there is an increasing need to design experiments that maximize the information content of the collected sample. For most standard statistical analyses, where the aim is to estimate some population parameter, maximizing information translates into minimizing the variance associated with a parameter's estimate. In many situations, researchers observe multiple outcomes for each unit in the sample and wish to make inferences on a parameter of the underlying population's joint distribution, routinely this is done via estimating the population mean vector. It is often the case that some or all of the individual components of this response vector are costly, risky (complications due to biopsy), or even destructive (requiring animal sacrifice). In such cases it may be desirable, for monetary or ethical reasons, to extract information from each unit that is sampled, without taking the exact measurement of the response of interest for each unit.

The most common approach for data collection method for making inference about population parameter is simple random sample (SRS) from a population. Even though each subject selected by SRS has an equal chance of being selected from a population to ensure the representativeness of a population, there is no guarantee that the selected sample will truly represent the population. However,

[1] Corresponding author: Jiann-Ping Hsu College of Public Health, Department of Biostatistics, Georgia Southern University, Statesboro, GA 30460, USA. Email: hrochani@georgiasouthern.edu

Published 31 January 2018 / journal homepage: http://csam.or.kr

the only guarantee one can have is that if the sampling process is being repeated over and over again, then the average of the attribute of interest for multiple SRS would provide the good estimator of the population value of the attribute. Ranked set sampling (RSS) (McIntyre, 1952) is a type of sampling scheme which allows researchers to use information from each unit in the sample, without taking every unit's exact measurement. The overall goal of the RSS is to obtain the sample from a population that is more likely to span the full range of the values in the population to have a more representative sample than the SRS of similar sample size. Traditionally, RSS can be used provided there is a reliable ranking mechanism available, which should be cheaper or safer than exact measurement, for the response of interest. The ranked but unmeasured units provide increased information over SRS of the same size improving parameter inference. The additional information provided by ranking is due to the fact that aspects of population structure are encoded through the order statistics. Knowledge of observations' order statistic and exact measurement improve inference since ranked units target different population attributes, unlike the identically distributed unit from a SRS. This has been shown in many works to translate into improvement in parameter inference compared to simple random samples of the same size.

In many situations, the outcome of interest is correlated with some auxiliary variable which may be easier to measure than the outcome of interest. For instance weight may be correlated with fasting blood glucose and may be easily obtained whereas some lab measurement would be necessary for blood glucose measurement. The application of RSS has appeared in series of papers. See for example, Chen (1999), Demir and Çıngı (2000), Huang *et al.* (2016), Jabrah *et al.* (2017), Kaur *et al.* (1996), Samawi and Al-Sagheer (2001).

An outline of the paper is as follows. In Section 2 we introduce the necessary notation and prove that mean estimation is unbiased with a smaller variance for RSS as compared to SRS. In addition, in Section 2, we also derived the limiting distribution of Hotelling's statistics ($\boldsymbol{Q}$) as well as the multivariate regression estimator using RSS. In Section 3, we perform a simulation study to compare the performance of RSS to SRS in terms of estimation as well as hypothesis testing. In Section 4, we apply the method on a real data set in the context of public health. We give concluding remarks and future directions for the method in Section 5.

2. Multivariate mean estimation using ranked set sampling

2.1. Ranked set sampling procedure

In this section, we will briefly describe how a ranked set sample may be collected in this section for a univariate random variable. To select the RSS of size n based on the auxiliary variable (X), the following steps should be performed.

1. Select the SRS of size r, from a population based on the auxiliary variable (X). r is referred as the set size which is typically between 2 and 5 although any size is possible. However, sizes larger than 5 may become impractical (Takahasi and Wakimoto, 1968).

2. Order the auxiliary variable and choose the minimum of ($X_{(1)}$). Measure the multivariate outcome of interest $\boldsymbol{Y}_{(1)}$.

3. Select the SRS of size r and order it based on the auxiliary variable again. Choose the second minimum ($X_{(2)}$) and measure the multivariate outcome of interest $\boldsymbol{Y}_{[2]}$.

4. Repeat this process until the $X_{(r)}$ and $\boldsymbol{Y}_{(r)}$ of r^{th} independent SRS are obtained.

Table 1: Structure of ranked set sampling

Cycle 1	$\left(X_{(1)1}, \boldsymbol{Y}_{(1)1}\right)$	$\left(X_{(2)1}, \boldsymbol{Y}_{(2)1}\right)$	$\cdots$	$\left(X_{(r)1}\boldsymbol{Y}_{(r)1}\right)$
Cycle 2	$\left(X_{(1)2}, \boldsymbol{Y}_{(1)2}\right)$	$\left(X_{(2)2}, \boldsymbol{Y}_{(2)2}\right)$	$\cdots$	$\left(X_{(r)2}, \boldsymbol{Y}_{(r)2}\right)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Cycle m	$\left(X_{(1)m}, \boldsymbol{Y}_{(1)m}\right)$	$\left(X_{(2)m}, \boldsymbol{Y}_{(2)m}\right)$	$\cdots$	$\left(X_{(r)m}, \boldsymbol{Y}_{(r)m}\right)$

5. The entire process of obtaining $(X_{(1)}, X_{(2)}, \ldots, X_{(r)})$ and $(\boldsymbol{Y}_{(1)}, \boldsymbol{Y}_{(2)}, \ldots, \boldsymbol{Y}_{(r)})$ is called a cycle.

6. Repeat m independent cycles to obtain a RSS of size $n = rm$.

Table 1 represents the structure of RSS. For more details about RSS (Jozani and Johnson, 2011; Kowalczyk, 2004; Patil *et al.*, 1995; Takahasi and Futatsuya, 1998).

2.2. Multivariate naive estimator

Our population of interest is an univariate auxilliary variable X and a d dimensional multivariate outcome $\boldsymbol{Y}$, with a covariance structure on the joint distribution of $(\boldsymbol{Y}, X)$ given by $\Sigma = \left[\begin{smallmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{smallmatrix}\right]$. Assuming that we have collected m RSS cycles of set size r, where the ranking has been done on X, we denote the data $(X_{(i)k}, \boldsymbol{Y}_{[i]k})$, $i = 1, 2, \ldots, r$, $K = 1, 2, \ldots, m$. Note that the subscript on X indicates that ranking has been done on X and the subscript on $\boldsymbol{Y}$ indicates that ranking on X may result in imperfect ranking on elements of $\boldsymbol{Y}$. The naive estimator is defined as $\hat{\boldsymbol{\mu}}_{y\text{RSS}} = (1/rm)\sum_{k=1}^{m}\sum_{i=1}^{r}\boldsymbol{Y}_{[i]k}$. It is straightforward to show this is an unbiased estimator of the mean of $\boldsymbol{Y}$. Since $\sum_{i=1}^{r} f_{X_{(i)}}(x) = r f_{X_{(i)}}(x)$ (Dell and Clutter, 1972)

$$\begin{aligned} E\hat{\boldsymbol{\mu}}_{y\text{RSS}} &= \frac{1}{rm}\sum_{k=1}^{m}\sum_{i=1}^{r}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\boldsymbol{y} f_{\boldsymbol{Y}|X_{(i)}=x}(\boldsymbol{y}|x) f_{X_{(i)}}(x)d\boldsymbol{y}dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\boldsymbol{y} f_{\boldsymbol{Y}|X_{(i)}=x}(\boldsymbol{y}|x)\frac{1}{r}\sum_{i=1}^{r} f_{X_{(i)}}(x)d\boldsymbol{y}dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\boldsymbol{y} f_{\boldsymbol{Y}|X_{(i)}=x}(\boldsymbol{y}|x) f_X(x)d\boldsymbol{y}dx = \boldsymbol{\mu}_y. \end{aligned}$$

Similarly for the variance (Dell and Clutter, 1972), by defining $\boldsymbol{\mu}_{[i]} = E\boldsymbol{Y}_{[i]}$ we have

$$\begin{aligned} \text{Var}(\hat{\boldsymbol{\mu}}_{y\text{RSS}}) &= \frac{1}{(rm)^2}\sum_{k=1}^{m}\sum_{i=1}^{r}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(\boldsymbol{y}-\boldsymbol{\mu}_{[i]})(\boldsymbol{y}-\boldsymbol{\mu}_{[i]})^{\top} f_{\boldsymbol{Y}|X=x}(\boldsymbol{y}|x) f_{X_{(i)}}(x)d\boldsymbol{y}dx \\ &= \frac{1}{rm^2}\sum_{k=1}^{m}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})^{\top} f_{\boldsymbol{Y}|X=x}(\boldsymbol{y}|x)\frac{1}{r}\sum_{i=1}^{r} f_{X_{(i)}}(x)d\boldsymbol{y}dx \\ &\quad - 2\frac{1}{(rm)^2}\sum_{k=1}^{m}\sum_{i=1}^{r}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(\boldsymbol{y}-\boldsymbol{\mu}) f_{\boldsymbol{Y}|X=x}(\boldsymbol{y}|x) f_{X_{(i)}}(x)d\boldsymbol{y}dx(\boldsymbol{\mu}_{[i]}-\boldsymbol{\mu})^{\top} \\ &\quad + \frac{1}{(rm)^2}\sum_{k=1}^{m}\sum_{i=1}^{r}(\boldsymbol{\mu}_{[i]}-\boldsymbol{\mu})\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\boldsymbol{Y}|X=x}(\boldsymbol{y}|x) f_{X_{(i)}}(x)d\boldsymbol{y}dx(\boldsymbol{\mu}_{[i]}-\boldsymbol{\mu})^{\top} \end{aligned}$$

$$= \frac{1}{(rm)}(\Sigma_{11}) - 2\frac{1}{r^2m}\sum_{i=1}^{r}(\boldsymbol{\mu}_{[i]}-\boldsymbol{\mu})(\boldsymbol{\mu}_{[i]}-\boldsymbol{\mu})^\top + \frac{1}{r^2m}\sum_{i=1}^{r}(\boldsymbol{\mu}_{[i]}-\boldsymbol{\mu})(\boldsymbol{\mu}_{[i]}-\boldsymbol{\mu})^\top$$

$$= \frac{1}{rm}\left(\Sigma_{11} - \frac{1}{r}\sum_{i=1}^{r}(\boldsymbol{\mu}_{[i]}-\boldsymbol{\mu})(\boldsymbol{\mu}_{[i]}-\boldsymbol{\mu})^\top\right). \tag{2.1}$$

It is clear that $\sum_{i=1}^{r}(\boldsymbol{\mu}_{[i]}-\boldsymbol{\mu})(\boldsymbol{\mu}_{[i]}-\boldsymbol{\mu})^\top$ is positive semi-definite since $\forall u \in \mathbb{R}^d$ we have $u^\top(\boldsymbol{\mu}_{[i]}-\boldsymbol{\mu})(\boldsymbol{\mu}_{[i]}-\boldsymbol{\mu})^\top u \geq 0$. Then $\forall u \in \mathbb{R}^d$ $u^\top(\mathrm{Var}(\hat{\boldsymbol{\mu}}_{y\mathrm{SRS}}) - \mathrm{Var}(\hat{\boldsymbol{\mu}}_{y\mathrm{RSS}}))u \geq 0$, or equivalently $\mathrm{Var}(\hat{\boldsymbol{\mu}}_{y\mathrm{SRS}}) \geq \mathrm{Var}(\hat{\boldsymbol{\mu}}_{y\mathrm{RSS}})$. Under the additional assumption that $r \geq d$ and X is correlated with each component of $\boldsymbol{Y}$ we have strict inequality.

2.3. Multivariate regression estimator

Regression estimators are used to increase precision in mean estimation by incorporating information in an auxilliary variable. In this case, we assume a linear regression of $\boldsymbol{Y}$ on X

$$\boldsymbol{Y} = \boldsymbol{\mu}_y + \boldsymbol{\beta}(X - \mu_x) + \boldsymbol{\epsilon}, \tag{2.2}$$

where X and $\boldsymbol{\epsilon}$ are independent and $\boldsymbol{\epsilon}$ is a mean zero residual vector with covariance $\Sigma_{\boldsymbol{\epsilon}}$. Then the regression equation with corresponding data from RSS is

$$\boldsymbol{Y}_{[i]k} = \boldsymbol{\mu}_y + \boldsymbol{\beta}(X_{(i)k} - \mu_x) + \boldsymbol{\epsilon}_{(i)k} \quad i = 1, 2, \ldots, r,\ k = 1, 2, \ldots, m. \tag{2.3}$$

It is worth noting that typically the mean of X, μ_x, is unknown. However, since the auxilliary variable X may be much cheaper to measure one may use the r^2m units collected from the first stage of sampling to estimate this quantity as $\bar{\mu}_x = (1/r^2m)\sum_k^m\sum_i^r\sum_j^r X_{ijk}$.

Then the regression estimator for the mean of the response is given by

$$\bar{\boldsymbol{Y}}_{\mathrm{reg}} = \hat{\boldsymbol{\mu}}_{y\mathrm{RSS}} + \hat{\boldsymbol{\beta}}(\bar{\mu}_x - \hat{\mu}_x), \tag{2.4}$$

where

$$\hat{\mu}_x = \frac{1}{rm}\sum_{k=1}^{m}\sum_{i=1}^{r}X_{(i)k}, \quad \hat{\boldsymbol{\beta}} = \frac{\sum_{k=1}^{m}\sum_{i=1}^{r}(X_{(i)k}-\hat{\mu}_x)\left(\boldsymbol{Y}_{[i]k}-\hat{\boldsymbol{\mu}}_{y\mathrm{RSS}}\right)}{\sum_{k=1}^{m}\sum_{i=1}^{r}(X_{(i)k}-\hat{\mu}_x)^2}.$$

It is straightforward to show that $\bar{\mu}_x$ and $\hat{\mu}_x$ are unbiased estimates of μ_x using similar arguments as in the previous section. When (2.3) holds conditional expectation implies that $E\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ and $E\bar{\boldsymbol{Y}}_{\mathrm{reg}} = \boldsymbol{\mu}_y$, so that the regression estimator based on RSS is unbiased. Also

$$\mathrm{Var}\left(\bar{\boldsymbol{Y}}_{\mathrm{reg}}\right) = E_X\mathrm{Var}_Y\left(\bar{\boldsymbol{Y}}_{\mathrm{reg}}|X\right) + \mathrm{Var}_X E_Y\left(\bar{\boldsymbol{Y}}_{\mathrm{reg}}|X\right).$$

Since $E_Y(\bar{\boldsymbol{Y}}_{\mathrm{reg}}|X) = \boldsymbol{\mu}_y + \boldsymbol{\beta}(\bar{\mu}_x - \hat{\mu}_x)$ the second term above is $(1/r^2m)\sum_{i=1}^{r}\sigma^2_{X_{(i)}}\boldsymbol{\beta}\boldsymbol{\beta}^\top$. For the first term $\mathrm{Cov}(\hat{\boldsymbol{\mu}}_{y\mathrm{RSS}}, \boldsymbol{\beta}(\bar{\mu}_x - \hat{\mu}_x)|X) = 0$ so that

$$\begin{aligned} E_X\mathrm{Var}_Y\left(\bar{\boldsymbol{Y}}_{\mathrm{reg}}|X\right) &= E_X\mathrm{Var}_Y\left(\hat{\boldsymbol{\mu}}_{y\mathrm{RSS}}|X\right) + E_X\mathrm{Var}_Y\left(\hat{\boldsymbol{\beta}}(\bar{\mu}_x - \hat{\mu}_x)|X\right) \\ &= \frac{1}{(r^2m)^2}\sum_{k=1}^{m}\sum_{i=1}^{r}\Sigma_{\boldsymbol{\epsilon}} + E_X\left((\bar{\mu}_x - \hat{\mu}_x)^2\,\mathrm{Var}_Y\left(\hat{\boldsymbol{\beta}}|X\right)\right) \\ &= \frac{1}{n}\Sigma_{\boldsymbol{\epsilon}} + \Sigma_{\boldsymbol{\epsilon}}E_X\frac{(\bar{\mu}_x - \hat{\mu}_x)^2}{\sum_{k=1}^{m}\sum_{i=1}^{r}(X_{(i)k}-\hat{\mu}_x)^2}. \end{aligned}$$

2.4. Testing for $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$

Theorem 1. *Let $\{\boldsymbol{Y}_{[i]k}\}$, $i = 1, 2, \ldots r$ and $k = 1, 2, \ldots m$ be a RSS sample from normal with mean vector $\boldsymbol{\mu}$ and variance covariance matrix Σ_{11}. Let*

$$\begin{aligned}
\bar{\boldsymbol{Y}}_{rss} &= \frac{1}{rm} \sum_{k=1}^{m} \sum_{i=1}^{r} \boldsymbol{Y}_{[i]k}, \\
S_{rss} &= \left[\frac{1}{rm - 1} \right] \sum_{k=1}^{m} \sum_{i=1}^{r} \left(\boldsymbol{Y}_{[i]k} - \bar{\boldsymbol{Y}}_{rss} \right) \left(\boldsymbol{Y}_{[i]k} - \bar{\boldsymbol{Y}}_{rss} \right)^{T} \\
\boldsymbol{Q} &= mr \left(\bar{\boldsymbol{Y}}_{rss} - \boldsymbol{\mu}_0 \right)^{\mathsf{T}} S_{rss}^{-1} \left(\bar{\boldsymbol{Y}}_{rss} - \boldsymbol{\mu}_0 \right)
\end{aligned}$$

Then for large sample the limiting distribution of $\boldsymbol{Q}$ is the χ^2-distribution with d degrees of freedom under the Null Hypothesis of $\boldsymbol{\mu} = \boldsymbol{\mu}_0$.

Proof:

$$\begin{aligned}
\bar{\boldsymbol{Y}}_{\text{rss}} &= \frac{1}{rm} \sum_{k=1}^{m} \sum_{i=1}^{r} \boldsymbol{Y}_{[i]k}, \\
\bar{\boldsymbol{Y}}_{\text{rss}} &= \frac{1}{r} \sum_{i=1}^{r} \bar{\boldsymbol{Y}}_{[i]}.
\end{aligned}$$

From Multivariate Central limit theorem $\sqrt{m}(\bar{\boldsymbol{Y}}_{[i]} - \boldsymbol{\mu}_{[i]}) \xrightarrow{d} \mathcal{N}_d(0, \Sigma_{11[i]}/m)$ as $m \to \infty$ where $\Sigma_{11[i]}$ is variance covariance matrix of $\boldsymbol{Y}_{[i]}$.

Since $\bar{\boldsymbol{Y}}_{[i]}$ are independent

$$\sqrt{mr} \left(\bar{\boldsymbol{Y}}_{\text{rss}} - \boldsymbol{\mu} \right) \xrightarrow{d} \mathcal{N}_d \left(0, \frac{\sum_{i=1}^{r} \Sigma_{11[i]}}{mr} \right).$$

$\sqrt{mr}(\bar{\boldsymbol{Y}}_{\text{rss}} - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}_d(0, \Sigma_{11R}/mr)$, where Σ_{11R} is variance covariance matrix of $\boldsymbol{Y}_{\text{rss}}$.

Therefore,

$$\sqrt{mr} \left(\bar{\boldsymbol{Y}}_{\text{rss}} - \boldsymbol{\mu} \right) \Sigma_{11R}^{-\frac{1}{2}} \xrightarrow{d} \mathcal{N}_d(0, \mathbb{1})$$

and hence

$$mr \left(\bar{\boldsymbol{Y}}_{\text{rss}} - \boldsymbol{\mu}_0 \right)^{\mathsf{T}} \Sigma_{11R}^{-1} \left(\bar{\boldsymbol{Y}}_{\text{rss}} - \boldsymbol{\mu}_0 \right) \sim \chi^2_{(d)}.$$

Since $S_{\text{rss}}^{-1}/\Sigma_{11R}^{-1} \xrightarrow{d} 1$ (See Appendix for more detail)

$$Q = mr \left(\bar{\boldsymbol{Y}}_{\text{rss}} - \boldsymbol{\mu}_0 \right)^{\mathsf{T}} \Sigma_{11R}^{-1} \frac{S_{\text{rss}}^{-1}}{\Sigma_{11R}^{-1}} \left(\bar{\boldsymbol{Y}}_{\text{rss}} - \boldsymbol{\mu}_0 \right) \sim \chi^2_{(d)}.$$

□

2.5. Testing for $H_0 : \boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)}$

Theorem 2. *Let* $\{\boldsymbol{Y}^{(t)}_{[i]k}\}$, $i = 1, 2, \ldots r$, $k = 1, 2, \ldots m$ *and* $t = 1, 2$ *are two RSS samples from* $\mathcal{N}_d(\boldsymbol{\mu}^{(1)}, \Sigma_{11})$ *and* $\mathcal{N}_d(\boldsymbol{\mu}^{(2)}, \Sigma_{11})$. *Let*

$$\bar{\boldsymbol{Y}}^{(1)}_{rss} = \frac{1}{r_1 m_1} \sum_{k=1}^{m_1} \sum_{i=1}^{r_1} \boldsymbol{Y}^{(1)}_{[i]k},$$

$$\bar{\boldsymbol{Y}}^{(2)}_{rss} = \frac{1}{r_2 m_2} \sum_{k=1}^{m_2} \sum_{i=1}^{r_2} \boldsymbol{Y}^{(2)}_{[i]k},$$

$$S_{rss} = \left[\frac{1}{r_1 m_1 + r_2 m_2 - 2}\right] \left[\sum_{k=1}^{m_1} \sum_{i=1}^{r_1} \left(\boldsymbol{Y}^{(1)}_{[i]k} - \bar{\boldsymbol{Y}}^{(1)}_{rss}\right)\left(\boldsymbol{Y}^{(1)}_{[i]k} - \bar{\boldsymbol{Y}}^{(1)}_{rss}\right)^T + \sum_{k=1}^{m_2} \sum_{i=1}^{r_2} \left(\boldsymbol{Y}^{(2)}_{[i]k} - \bar{\boldsymbol{Y}}^{(2)}_{rss}\right)\left(\boldsymbol{Y}^{(2)}_{[i]k} - \bar{\boldsymbol{Y}}^{(2)}_{rss}\right)^T\right].$$

Then, $\boldsymbol{Q} = \{(r_1 m_1 \cdot r_2 m_2)/(r_1 m_1 + r_2 m_2)\}(\bar{\boldsymbol{Y}}^{(1)}_{rss} - \bar{\boldsymbol{Y}}^{(2)}_{rss})^{\intercal} S^{-1}_{rss} (\bar{\boldsymbol{Y}}^{(1)}_{rss} - \bar{\boldsymbol{Y}}^{(2)}_{rss})$, *for large samples, has the limiting distribution as* χ^2 *with d degrees of freedom under* $H_0 : \boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)}$.

Proof: The proof is similar to that as in Theorem 1. □

2.6. Small samples

For small to moderate samples, for SRS, under H_0 the $\boldsymbol{Q}$ statistics is distributed as $\{(N-1)\,d\}/(N-p)$ $F_{d,N-d}$ (Seber, 2009). As explicit distribution of $\boldsymbol{Q}$ statistics is not known, for small or moderate size of RSS samples, we recommend performing hypothesis testing by Bootstrap method. Resampling method for RSS was proposed by (Chen *et al.*, 2004; Modarres *et al.*, 2006). They suggest a natural method to obtain bootstrap samples from each row (within cycle) of a RSS.

3. Simulation

In this section, we conducted the simulation study to estimate the multivariate outcome mean and the performance of the hypothesis testing by RSS scheme. We also studied the performance of testing hypothesis of equality of multivariate outcome means for two groups. For estimation of α of testing $H_o : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ vs. $H_a : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$, we considered four multivariate outcomes Y_i $(i = 1, 2, 3, 4)$ with $\boldsymbol{\mu} = [0.3, 0.3, 0.3, 0.2]$, variances as $\sigma^2_1 = \sigma^2_2 = \sigma^2_3 = \sigma^2_4 = 4$ and covariances as $\sigma_{12} = 2.39, \sigma_{13} = 1.59, \sigma_{14} = 2.83, \sigma_{23} = 3.19, \sigma_{24} = 1.18$, and $\sigma_{34} = 2.24$. The auxiliary covariate (X) was simulated with mean 0 and variance $\sigma^2_x = 1$. For this simulation study, we considered unstructured covariance among multivariate outcome Y_i as shown below. Moreover, we used autoregressive covariance structure between auxiliary variable X and Y_i with correlation parameter ρ.

$$\mathrm{Cov}(X, Y_i) = \begin{bmatrix} 1 & 2\rho & 2\rho^2 & 2\rho^3 & 2\rho^4 \\ 2\rho & 4 & 2.39 & 1.59 & 2.83 \\ 2\rho^2 & 2.39 & 4 & 3.19 & 1.18 \\ 2\rho^3 & 1.59 & 3.19 & 4 & 2.24 \\ 2\rho^4 & 2.83 & 1.18 & 2.24 & 4 \end{bmatrix}.$$

The RSS for of X and Y_i were simulated from multivariate normal with mean $\boldsymbol{\mu}$ and above variance covariance matrix by following the steps as described in Section 2.1. For comparisons of estimation of α for SRS and RSS, different sample sizes $(n = rm)$ were evaluated by varying the ρ, set size and cycle size. This entire process was repeated 2,000 times. For details of the parameter values, referred

Table 2: Estimation of the α of testing $H_o : \boldsymbol{\mu} = 0$ vs. $H_a : \boldsymbol{\mu} \neq 0$

ρ	Cycle	Set = 3			Set = 4			Set = 5		
		SRS	RSS	BS^a	SRS	RSS	BS^a	SRS	RSS	BS^a
−0.8	5	0.0460	0.1375	0.0215	0.0375	0.1065	0.0475	0.0455	0.0795	0.0605
	10	0.0455	0.0755	0.0450	0.0460	0.0530	0.0495	0.0410	0.0535	0.0620
	20	0.0520	0.0545	0.0535	0.0485	0.0445	0.0545	0.0560	0.0410	0.0605
	30	0.0555	0.0455	0.0560	0.0495	0.0430	0.0600	0.0480	0.0360	0.0530
−0.6	5	0.0475	0.1580	0.0210	0.0480	0.1070	0.0470	0.0550	0.0710	0.0600
	10	0.0460	0.0720	0.0455	0.0465	0.0590	0.0555	0.0490	0.0405	0.0500
	20	0.0520	0.0475	0.0460	0.0450	0.0460	0.0520	0.0605	0.0395	0.0550
	30	0.0450	0.0415	0.0500	0.0470	0.0385	0.0555	0.0505	0.0415	0.0595
−0.4	5	0.0460	0.1400	0.0225	0.0540	0.1055	0.0440	0.0515	0.0840	0.0585
	10	0.0580	0.0690	0.0395	0.0515	0.0555	0.0515	0.0450	0.0530	0.0655
	20	0.0495	0.0500	0.0525	0.0495	0.0415	0.0510	0.0520	0.0360	0.0560
	30	0.0595	0.0460	0.0530	0.0520	0.0360	0.0530	0.0560	0.0290	0.0515
0.4	5	0.0515	0.1530	0.0290	0.0615	0.0975	0.0425	0.0570	0.0800	0.0640
	10	0.0445	0.0730	0.0405	0.0560	0.0495	0.0485	0.0495	0.0445	0.0540
	20	0.0520	0.0420	0.0430	0.0505	0.0430	0.0575	0.0495	0.0310	0.0505
	30	0.0525	0.0495	0.0570	0.0405	0.0385	0.0580	0.0495	0.0310	0.0505
0.6	5	0.0520	0.1545	0.0255	0.0545	0.0900	0.0380	0.0465	0.0785	0.0610
	10	0.0540	0.0840	0.0525	0.0595	0.0660	0.0620	0.0475	0.0535	0.0595
	20	0.0440	0.0495	0.0490	0.0470	0.0430	0.0555	0.0525	0.0400	0.0555
	30	0.0555	0.0455	0.0535	0.0475	0.0340	0.0510	0.0555	0.0295	0.0465
0.8	5	0.0575	0.1365	0.0195	0.0465	0.0970	0.0450	0.0470	0.0840	0.0580
	10	0.0520	0.0750	0.0455	0.0520	0.0730	0.0680	0.0495	0.0470	0.0580
	20	0.0495	0.0590	0.0620	0.0535	0.0360	0.0470	0.0580	0.0350	0.0505
	30	0.0560	0.0450	0.0520	0.0495	0.0405	0.0580	0.0500	0.0400	0.0570

SRS = simple random sample; RSS = ranked set sampling; BS^a = Bootstrap α.

to Table 2. Table 2 results demonstrate that we can achieve nominal value for α by using RSS with moderate to large samples, however, for smaller sample bootstrap RSS sampling can achieve nominal value for α.

For estimation of the power of testing $H_o : \boldsymbol{\mu} = 0$ vs. $H_a : \boldsymbol{\mu} \neq 0$, similar simulation settings were considered as described above except with $\boldsymbol{\mu} = [0.6, 0.6, 0.6, 0.4]$. In addition to that bootstrap power was also calculated by taking 1,000 bootstrap samples for each simulated RSS. Furthermore, MSE of SRS, MSE of RSS and the multivariate naive estimator efficiency were calculated. Table 3 reports the simulation results for estimating the power of testing hypothesis under various simulation settings. We can also report that the power of the test increases as the set size increases with RSS, however, for testing hypothesis RSS gives more power than SRS. As expected, Table 3 also shows that RSS provides more efficient estimates of the multivariate naive estimator in terms of smaller MSEs.

Furthermore, the performance of testing hypothesis of equality of multivariate outcome means for two groups, we simulated two groups with multivariate outcome (Y_i) $(i = 1, 2, 3, 4)$ with means for the first group $\mu_1 = [0.3, 0.3, 0.3, 0.2]$ and mean for the second group $\mu_2 = [0.6, 0.6, 0.6, 0.4]$ with similar covariance matrix of Y as described above $(\text{Cov}(X, Y_i))$. Table 4 represents the results of estimation of power of the testing hypothesis $H_o : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ vs. $H_a : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ with various parameter values of ρ, set size and cycle sizes. Overall, from Table 4, we can conclude that RSS is more powerful for testing hypothesis of equality of multivariate outcome means for two groups compared to SRS.

We also conducted a simulation study to show that the multivariate regression estimator for RSS is more efficient than SRS. We considered multivariate outcomes $\boldsymbol{Y}$ with mean $\boldsymbol{\mu} = (0.3, 0.3, 0.3, 0.2)$ and the variance-covariance matrix $(\text{Cov}(X, Y_i)))$ as described above in this section. We also simu-

Table 3: Estimation of power of testing $H_o : \boldsymbol{\mu} = 0$ vs. $H_a : \boldsymbol{\mu} \neq 0$

Set	$\boldsymbol{\rho}$	Cycle	SRS Power	RSS Power	Bootstrap Power	SRS MSE	RSS MSE
3	0.4	5	0.0900	0.2255	0.0445	3.88E−05	2.07E−05
		10	0.1645	0.2120	0.1425	2.53E−06	1.37E−06
		20	0.3445	0.3490	0.3510	1.79E−07	8.11E−08
		30	0.4880	0.5280	0.5630	3.07E−08	1.69E−08
	0.6	5	0.0850	0.2260	0.0440	4.93E−05	2.58E−05
		10	0.1400	0.1990	0.1345	2.97E−06	1.61E−06
		20	0.2945	0.2980	0.3030	1.94E−07	9.52E−08
		30	0.4300	0.4540	0.4860	4.31E−08	2.30E−08
	0.8	5	0.0995	0.2560	0.0505	2.28E−05	1.18E−05
		10	0.2055	0.2640	0.1735	1.48E−06	8.17E−07
		20	0.4140	0.4415	0.4465	1.17E−07	4.64E−08
		30	0.5950	0.6805	0.7080	1.96E−08	9.07E−09
4	0.4	5	0.1100	0.1895	0.0950	1.16E−05	5.45E−06
		10	0.2000	0.2365	0.2255	7.51E−07	3.29E−07
		20	0.4490	0.4580	0.5125	5.32E−08	2.22E−08
		30	0.6165	0.6600	0.7265	1.00E−08	4.16E−09
	0.6	5	0.1015	0.1770	0.0865	1.53E−05	6.32E−06
		10	0.1875	0.2020	0.1850	1.08E−06	4.01E−07
		20	0.3565	0.3560	0.4000	6.00E−08	2.54E−08
		30	0.5515	0.6285	0.6960	1.10E−08	4.65E−09
	0.8	5	0.1125	0.2255	0.1060	8.08E−06	3.59E−06
		10	0.2500	0.2940	0.2820	5.32E−07	2.15E−07
		20	0.4950	0.5620	0.6150	2.61E−08	1.27E−08
		30	0.7435	0.8290	0.8630	6.97E−09	2.63E−09
5	0.4	5	0.1440	0.1920	0.1475	5.22E−06	1.86E−06
		10	0.2575	0.2830	0.3080	3.01E−07	1.13E−07
		20	0.5325	0.5810	0.6565	2.10E−08	7.11E−09
		30	0.7355	0.7910	0.8455	4.41E−09	1.54E−09
	0.6	5	0.1200	0.1690	0.1285	6.41E−06	2.36E−06
		10	0.2110	0.2140	0.2430	3.87E−07	1.45E−07
		20	0.4800	0.4930	0.5820	2.70E−08	9.50E−09
		30	0.6445	0.6960	0.7795	5.27E−09	1.63E−09
	0.8	5	0.1505	0.2160	0.1625	2.96E−06	1.20E−06
		10	0.3080	0.3310	0.3625	2.05E−07	6.88E−08
		20	0.6425	0.7230	0.7935	1.28E−08	4.78E−09
		30	0.8360	0.9060	0.9490	2.44E−09	8.07E−10
3	−0.4	5	0.1360	0.3185	0.0790	2.34E−04	1.35E−04
		10	0.2635	0.3865	0.2970	1.35E−05	8.05E−06
		20	0.6155	0.6740	0.6760	9.20E−07	5.14E−07
		30	0.8125	0.8450	0.8625	1.94E−07	1.01E−07
	−0.6	5	0.1210	0.3260	0.0815	6.66E−04	3.72E−04
		10	0.2670	0.3685	0.2845	4.59E−05	2.30E−05
		20	0.5460	0.5945	0.5950	2.55E−06	1.39E−06
		30	0.7690	0.7840	0.8100	5.50E−07	2.61E−07
	−0.8	5	0.1235	0.3075	0.0815	1.20E−03	6.60E−04
		10	0.2550	0.3735	0.2915	7.74E−05	4.49E−05
		20	0.5505	0.5840	0.5835	4.62E−06	2.18E−06
		30	0.7840	0.8160	0.8375	9.25E−07	4.42E−07
4	−0.4	5	0.1830	0.3020	0.1775	8.33E−05	3.32E−05
		10	0.4050	0.4740	0.4600	5.39E−06	2.19E−06
		20	0.7870	0.8130	0.8365	3.10E−07	1.33E−07
		30	0.9150	0.9505	0.9610	5.80E−08	2.81E−08

Continued

Set	ρ	Cycle	SRS Power	RSS Power	Bootstrap Power	SRS MSE	RSS MSE
4	−0.6	5	0.1675	0.3075	0.1735	2.34E−04	8.93E−05
		10	0.3410	0.4435	0.4255	1.41E−05	6.17E−06
		20	0.7095	0.7420	0.7725	8.19E−07	3.46E−07
		30	0.9015	0.9105	0.9335	1.66E−07	7.02E−05
	−0.8	5	0.1565	0.3065	0.1860	3.50E−04	1.65E−04
		10	0.3635	0.4240	0.4095	1.95E−05	1.05E−05
		20	0.7605	0.7605	0.7935	1.41E−06	6.53E−07
		30	0.8985	0.8970	0.9210	2.95E−07	1.25E−07
5	−0.4	5	0.2555	0.3640	0.3100	3.33E−05	1.20E−05
		10	0.5135	0.5645	0.5920	2.18E−06	8.53E−07
		20	0.8590	0.8865	0.9185	1.08E−07	4.18E−08
		30	0.9775	0.9785	0.9865	2.34E−08	8.02E−09
	−0.6	5	0.2120	0.3210	0.2605	9.69E−05	3.36E−05
		10	0.4630	0.5105	0.5400	5.91E−06	2.21E−06
		20	0.8155	0.8235	0.8565	3.88E−07	1.31E−07
		30	0.9530	0.9500	0.9650	7.24E−08	2.69E−08
	−0.8	5	0.2115	0.3435	0.2920	1.49E−04	5.84E−05
		10	0.4595	0.5270	0.5595	9.85E−06	3.53E−06
		20	0.8080	0.8135	0.8575	6.00E−07	2.14E−07
		30	0.9485	0.9525	0.9680	1.23E−07	4.38E−08

SRS = simple random sample; RSS = ranked set sampling; MSE = mean square error.

Table 4: Estimation of power of testing $H_o : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ vs. $H_a : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$

ρ	Cycle	Set = 3			Set = 4			Set = 5		
		SRS	RSS	BS^a	SRS	RSS	BS^a	SRS	RSS	BS^a
−0.4	10	0.3055	0.3740	0.4058	0.3190	0.3545	0.4227	0.3905	0.4015	0.4354
	20	0.3975	0.3990	0.4021	0.4665	0.4852	0.4973	0.4675	0.4840	0.5175
	30	0.4785	0.4885	0.4800	0.5000	0.5245	0.5127	0.5865	0.6035	0.6131
	40	0.5710	0.5605	0.5824	0.6150	0.6505	0.6421	0.6480	0.6570	0.6491
−0.6	10	0.2710	0.3610	0.3812	0.3470	0.3630	0.408	0.3639	0.3900	0.4128
	20	0.3950	0.4160	0.4210	0.4240	0.4125	0.4357	0.4515	0.4970	0.5087
	30	0.4570	0.4470	0.4424	0.4825	0.4915	0.4879	0.5285	0.5675	0.5564
	40	0.5555	0.5535	0.5542	0.5745	0.6210	0.6321	0.6180	0.6475	0.6427
−0.8	10	0.2820	0.3550	0.3829	0.3160	0.3565	0.4186	0.3670	0.3855	0.4210
	20	0.3785	0.3990	0.4021	0.4135	0.4500	0.4610	0.4475	0.5035	0.5142
	30	0.4675	0.4625	0.4610	0.4810	0.5270	0.5287	0.5165	0.5525	0.5641
	40	0.5275	0.5280	0.5195	0.5635	0.6035	0.5987	0.6040	0.6335	0.6289
0.4	10	0.3485	0.4480	0.4845	0.4025	0.4445	0.4975	0.4715	0.4775	0.5012
	20	0.5075	0.5580	0.5641	0.5700	0.6305	0.6441	0.6475	0.6625	0.6951
	30	0.6520	0.6520	0.6641	0.7325	0.7825	0.7888	0.7430	0.8065	0.8125
	40	0.6895	0.7265	0.7248	0.8105	0.8605	0.8589	0.8600	0.9210	0.9287
0.6	10	0.3305	0.4130	0.4965	0.4055	0.4270	0.5102	0.4380	0.4680	0.5354
	20	0.4745	0.5020	0.5214	0.5360	0.6040	0.6214	0.5985	0.6610	0.6698
	30	0.5970	0.6265	0.6369	0.6620	0.7135	0.7125	0.7420	0.8105	0.8214
	40	0.6585	0.7145	0.7235	0.7495	0.8020	0.7985	0.8300	0.8920	0.8879
0.8	10	0.3700	0.4490	0.5035	0.4665	0.5155	0.5210	0.5140	0.5450	0.5621
	20	0.5495	0.5865	0.6089	0.6490	0.7020	0.7124	0.7275	0.7975	0.8213
	30	0.7040	0.7415	0.7358	0.7745	0.8555	0.8614	0.8395	0.9255	0.9159
	40	0.8055	0.8530	0.8521	0.8755	0.9295	0.9124	0.9320	0.9725	0.9800

SRS = simple random sample; RSS = ranked set sampling; BS^a = Bootstrap α.

lated correlated auxiliary covariate (X) with mean 0 and variance 1. Table 5 shows that for various parameter settings, the RSS is more efficient than SRS in estimating multivariate regression estimator.

Table 5: Estimation of multivariate regression estimator

ρ	Cycle	Set = 3		Set = 4		Set = 5	
		MSE SRS	MSE RSS	MSE SRS	MSE RSS	MSE SRS	MSE RSS
0.4	5	0.0078	0.0010	0.0024	0.0002	0.0010	6.49E−05
	10	0.0004	4.40E−05	0.0001	1.22E−05	4.51E−05	3.29E−06
	20	2.35E−05	2.73E−06	5.23E−06	6.57E−07	2.46E−06	2.30E−07
	30	4.02E−06	5.61E−07	1.23E−06	1.26E−07	5.43E−07	4.17E−08
0.6	5	0.0098	0.0011	0.0032	0.0003	0.0011	7.69E−05
	10	0.0004	5.77E−05	0.0001	1.16E−05	6.31E−05	4.72E−06
	20	2.61E−05	3.60E−06	7.21E−06	7.73E−07	3.39E−06	2.90E−07
	30	4.92E−06	6.95E−07	1.52E−06	1.72E−07	5.99E−07	7.68E−08
0.8	5	0.0109	0.0006	0.0036	0.0001	0.0012	5.70E−05
	10	0.0005	2.96E−05	0.0002	7.87E−06	7.10E−05	2.62E−06
	20	2.98E−05	1.74E−06	1.08E−05	4.59E−07	3.68E−06	1.57E−07
	30	5.27E−06	3.37E−07	1.64E−06	8.36E−08	6.44E−07	2.75E−08
−0.4	5	0.0133	0.0057	0.0039	0.0013	0.0014	0.0005
	10	0.0006	0.0003	0.0002	5.74E−05	8.44E−05	2.53E−05
	20	3.64E−05	1.52E−05	1.25E−05	4.58E−06	4.25E−06	1.58E−06
	30	7.16E−06	3.44E−06	2.69E−06	9.54E−07	7.72E−07	3.07E−07
−0.6	5	0.0270	0.0145	0.0070	0.0035	0.0028	0.001197
	10	0.0013	0.0007	0.0003	0.0002	0.0001	7.06E−05
	20	6.85E−05	4.62E−05	1.82E−05	1.51E−05	8.30E−06	4.00E−06
	30	1.15E−05	9.00E−06	4.13E−06	2.45E−06	1.63E−06	8.03E−07
−0.8	5	0.0485	0.0273	0.0114	0.0074	0.0038	0.0024
	10	0.0017	0.0015	0.0005	0.0004	0.0002	0.0001
	20	7.85E−05	7.46E−05	2.69E−05	2.11E−05	1.17E−05	8.28E−06
	30	1.85E−05	1.73E−05	5.90E−06	4.77E−06	2.42E−06	1.32E−06

MSE = mean square error; SRS = simple random sample; RSS = ranked set sampling.

4. Application to China Health and Nutrition Survey data

In this section, we illustrate the efficient ranked set sampling method via ranking on baseline covariate to estimate the multivariate outcome mean, investigate the performance of the hypothesis testing for two groups and estimation of multivariate regression estimator by using the China Health and Nutrition Survey (CHNS) for year 2009. The CHNS is the only large-scale household based survey in China (Yan *et al.*, 2003). As a part of the survey, anthropometry were collected on 10,242 children and adults aged ≥ 7 in year 2009 along with other demographic information. Only 9,986 individuals agreed to provide the fasting blood samples which were evaluated for many biomarkers of diabetes and cardio-metabolic risk factors. For illustration purposes, we focused on the variables such as age of the individuals as our ranking auxiliary variable, and cardio-metabolic biomarkers, for example, Apolipoprotein A, Total cholesterol and Hemoglobin A1c. We treated the survey data as a population and selected the range of RSS (N = set $*$ cycle) as shown in Table 6 by ranking on the baseline covariate age. SRS of similar size N was also selected from CHNS data to evaluate the performance of the hypothesis testing and the efficiency of the sampling procedure compared to RSS in estimating the multivariate outcome mean. The correlations (ρ) between age and biomarkers Apolipoprotein A, Total cholesterol and Hemoglobin A1c are 0.12, 0.32, and 0.22 respectively. The mean for Apolipoprotein A, Total cholesterol and Hemoglobin A1c are 1.14 (g/L), 4.78 mmol/L and 5.67 mmol/L respectively, and for comparison purposes, they can be treated as the true parameters. Table 7 represents the power comparison of RSS with SRS for multivariate means of males and females. Table 7 represents that we can achieve more power with RSS compared to SRS with similar sample sizes. Table 8 shows the results for multivariate regression estimation for biomarker data. We also took 1,000 samples of SRS

Table 6: Multivariate mean estimation and MSEs for China Health and Nutrition Survey data

Set	Cycle	SRS MSE	RSS MSE	Efficiency
3	5	3.07E−05	3.04E−05	1.01
	10	3.88E−06	3.41E−06	1.14
	20	4.66E−07	4.43E−07	1.05
	30	1.38E−07	1.26E−07	1.09
4	5	1.31E−05	1.17E−05	1.12
	10	1.65E−06	1.47E−06	1.12
	20	2.01E−07	1.81E−07	1.11
	30	5.80E−08	5.43E−08	1.07
5	5	6.67E−06	5.85E−06	1.14
	10	8.00E−07	7.08E−07	1.13
	20	1.01E−07	9.06E−08	1.11
	30	2.94E−08	2.57E−08	1.14

SRS = simple random sample; RSS = ranked set sampling; MSE = mean square error.

Table 7: Estimation of power of testing for Biomarker data for gender

Cycle	Set = 3		Set = 4		Set = 5	
	SRS	RSS	SRS	RSS	SRS	RSS
10	0.2531	0.3414	0.2875	0.3397	0.3155	0.3625
20	0.3353	0.3663	0.3722	0.3968	0.4017	0.4265
30	0.3893	0.4066	0.4243	0.4526	0.4691	0.4890
40	0.4324	0.4425	0.4823	0.5017	0.5374	0.5415

SRS = simple random sample; RSS = ranked set sampling.

Table 8: Multivariate regression estimation for China Health and Nutrition Survey data

Cycle	Set = 3		Set = 4		Set = 5	
	MSE SRS	MSE RSS	MSE SRS	MSE RSS	MSE SRS	MSE RSS
5	4.87E−05	3.03E−05	2.08E−05	1.89E−05	1.00E−05	5.82E−06
10	5.40E−06	4.84E−06	2.26E−06	1.49E−06	1.12E−06	9.91E−07
20	7.11E−07	5.13E−07	2.56E−07	2.25E−07	1.64E−07	1.09E−07
30	2.07E−07	1.59E−07	7.41E−08	5.58E−08	3.92E−08	3.31E−08

MSE = mean square error; SRS = simple random sample; RSS = ranked set sampling.

and RSS of sample size 80 (set = 4 and cycle = 20) and plotted the confidence regions as shown in Figure 1. From Figure 1, we can see that the confidence region for SRS (blue nets) lies completely outside of the confidence region of RSS (red).

5. Conclusion

In statistics, it is important to have a sampling method which is cost effective. RSS is one the important method which can be used to have a more efficient multivariate mean estimator compared to most commonly used method of SRS. The samples taken by using RSS method are more representative samples due to its inherent structure imposed by ranking based on easy-to-available covariates. In this paper, we demonstrated that the RSS is more efficient in estimating the multivariate mean as well as in hypothesis testing for one and two independent samples. Simulation studies for the performance of hypothesis testing showed that the RSS is more powerful compared to SRS. In general, in estimation of the population mean, RSS improves the precision relative to SRS with the same sample size, n. This is true even if the correlation between the auxiliary variable X and multivariate outcome $\boldsymbol{Y}$ is moderate to high (± 0.4 to ± 0.8). However, when the correlation between X and $\boldsymbol{Y}$ is very low (such

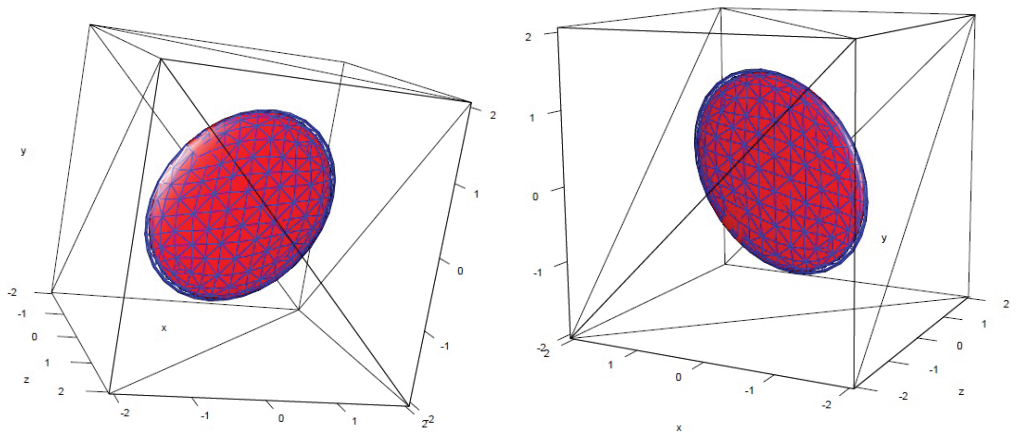

Figure 1: *Confidence region for SRS (blue nets) and RSS (dolid ted) gor China Health and Nutrition Survey data.*

as ± 0.001), RSS is equivalent to SRS and the ranking is not better than random. In practice, the key issue is whether the increase in precision is sufficient to justify the increased costs associated with the ranking process. In contrast, when the correlation between X and $\boldsymbol{Y}$ is very high (±0.9 or higher), the precision in estimating the population mean will be very high as this will improve the ranking of X on Y (Ridout, 2003).

Missing data is a very common problem in all most every research and can have a very significant impact on the inferences drawn from the collected data such as biased estimation of population parameters and loss of statistical power (Little and Rubin, 2014). The valid statistical analysis which has appropriate missing data mechanisms assumptions (missing completely at random, missing at random, or missing not at random) should be performed in SRS and in RSS. There is an extensive literature available on how to deal with missing data for RSS in auxiliary variable X and univariate response Y (Bouza-Herrera, 2013). However, handling the missing data in multivariate $\boldsymbol{Y}$ with monotone or arbitrary missing pattern is still the active area of research.

References

Bouza-Herrera CN (2013). *Handling Missing Data in Ranked Set Sampling*, Springer, Heidelberg.

Chen Z (1999). Density estimation using ranked-set sampling data, *Environmental and Ecological Statistics*, **6**, 135–146.

Chen Z, Bai Z, and Sinha B (2004). *Ranked Set Sampling: Theory and Applications*, Springer Science & Business Media, New York.

Dell TR and Clutter JL (1972). Ranked set sampling theory with order statistics background, *Biometrics*, **28**, 545–555.

Demir S and Çıngı H (2000). An application of the regression estimator in ranked set sampling, *Hacettepe Bulletin of Natural Sciences and Engineering, Series B*, **29**, 93–101.

Huang Y, Samawi HM, Vogel R, Yin J, Gato WE, and Linder DF (2016). Evaluating the efficiency of treatment comparison in crossover design by allocating subjects based on ranked auxiliary variable, *Communications for Statistical Applications and Methods*, **23**, 543–553.

Jabrah R, Samawi HM, Vogel R, Rochani HD, Linder DF, and Klibert J (2017). Using ranked auxiliary covariate as a more efficient sampling design for ANCOVA model: analysis of a psychological

intervention to buttress resilience, *Communications for Statistical Applications and Methods*, **24**, 241–254.

Jozani MJ and Johnson BC (2011). Design based estimation for ranked set sampling in finite populations, *Environmental and Ecological Statistics*, **18**, 663–685.

Kaur A, Patil GP, Shirk SJ, and Taillie C (1996). Environmental sampling with a concomitant variable: a comparison between ranked set sampling and stratified simple random sampling, *Journal of Applied Statistics*, **23**, 231–256.

Kowalczyk B (2004). Ranked set sampling and its applications in finite population studies, *Statistics in Transition*, **6**, 1031–1046.

Little RJA and Rubin DB (2014). *Statistical Analysis with Missing Data* (2nd ed), John Wiley & Sons, New York.

McIntyre GA (1952). A method for unbiased selective sampling, using ranked sets, *Australian Agricultural Research*, **3**, 385–390.

Modarres R, Hui TP, and Zheng G (2006). Resampling methods for ranked set samples, *Computational Statistics & Data Analysis*, **51**, 1039–1050.

Patil GP, Sinha AK, and Taillie C (1995). Finite population corrections for ranked set sampling, *Annals of the Institute of Statistical Mathematics*, **47**, 621–636.

Ridout MS (2003). On ranked set sampling for multiple characteristics, *Environmental and Ecological Statistics*, **10**, 255–262.

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

Seber GAF (2009). *Multivariate Observations*, John Wiley & Sons, New York.

Takahasi K and Futatsuya M (1998). Dependence between order statistics in samples from finite population and its application to ranked set sampling, *Annals of the Institute of Statistical Mathematics*, **50**, 49–70.

Takahasi K and Wakimoto K (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering, *Annals of the Institute of Statistical Mathematics*, **20**, 1–31.

Yan S, Li J, Li S, Zhang B, Du S, Gordon-Larsen P, Adair L, and Popkin, B (2012). The expanding burden of cardiometabolic risk in China: the China Health and Nutrition Survey, *Obesity Reviews: An Official Journal Of The International Association For The Study Of Obesity*, **13**, 810–821.

Received May 31, 2017; Revised September 27, 2017; Accepted December 08, 2017