Nonparametric Bootstrap Estimation of the Population Ratio Using Ranked Set Sampling

Kevin Carl P. Santos, Charisse Mae I. Castillo
Reyna Belle d.S. de Jesus, Niña B. Telan and
Crystal Angela P. Vidal
University of the Philippines Diliman

Ranked Set Sampling (RSS) yields unbiased and more reliable estimators of the population mean and proportion while keeping low costs. Using nonparametric bootstrap estimation, the efficiency of the ratio estimates using RSS with Simple Random Sampling (SRS) are compared. A simulation study accounting for the sampling rate, population size, population variance and correlation with the concomitant variable was conducted to compare RSS and SRS in estimating ratios. When ranking was done on the numerator characteristic, RSS generally performs better than SRS in terms of their relative bias. Likewise, in terms of precision, RSS generally produces better estimates when ranking was done on the numerator characteristic. On homogeneous populations, contrary to what was expected, RSS performed better over SRS. On heterogeneous populations, on the other hand, the two sampling designs are generally comparable.

Keywords: Population ratio, ranked set sampling, simple random sampling, nonparametric bootstrap estimation

1. Introduction

Ever since its conception, Ranked Set Sampling (RSS) has been gaining popularity in literature as a sampling technique that can potentially increase efficiency while keeping reasonable costs. Introduced by McIntyre (1952), this sampling scheme was first implemented to estimate the average pasture and forage yields by ranking a small number of fields with respect to the characteristic of interest via judgment. Since then, RSS has been used in the fields of agriculture, environmental studies, ecology and many others.

RSS is especially designed for situations when the variable of interest is difficult or costly to measure.Take for instance the mean height of trees in a certain forest as the characteristic of interest. Instead of getting a large sample size to ensure that the population is represented well, a possible approach is to do

ranking. Taking first three random trees, each is categorized as small, medium or large based on their height, making three subsets. The process is repeated until the three subsets are well populated depending on the number which was initially set by the researcher. It is then that measurements will be averaged for every subset. Finally, the average measurement of the three subsets becomes the RSS estimate. Consider toxic contaminant in soil areas as another example. As it is costly and quite labor intensive to the said feature, some other highly correlated characteristics such as amount of surface staining or degree of soil discoloration which is cheaper to obtain may be used as a ranking variable (Patil, 1995).In some cases, subjective information is often available which can be used to make an artificial stratification of the data as mentioned by Chen et al. (2004). However, ranking by judgment is not always recommended because it may result to ranking errors. Stoke (1977) found out that if there is an auxiliary variable available, it can be used to "judgment order" the variable of interest. The gains in precision, of course, depend on the correlation of the two variables.

It has been recognized that the efficiency of the RSS relative to SRS, assuming equal sample sizes, depends on the correlation of the ranking variable with the target variable. According to Chen et al. (2006), the ranking variable becomes more useful as the degree of association between this variable and the variable of interest increases.

In most times, RSS estimators perform better than their SRS counterparts for they have smaller standard errors, that is, they have better precision. The advantage of RSS over SRS can be more clearly seen when it is easier to rank the sampling units than to obtain their actual measures. Indeed, RSS has emerged as an alternative to SRS.

Most of RSS literatures available focused only on finding estimates of the mean. Finding an estimate of the population ratio and assessing its performance against other sampling designs, however, has not been given as much attention.

In the study by Samawi and Muttlak (1996), they used a population with bivariate elements and showed how RSS can give a better estimator for the population ratio. They concluded that the efficiency of the estimator of the population ratio is dependent on the ranking variable and on the size of the population. Furthermore, it will increase even in the presence of ranking errors.

It has been established that RSS gives a better estimate of the population mean and proportion according to Takahasi and Wakimoto (1968), Wolfe (2004) and Chen et al. (2005). This paper aims to compare the bias and precision of estimates of the population ratio using RSS and SRS.

Since then, a lot of literature has been devoted to RSS and its advantages over other sampling designs. Many researchers have come up with modifications

of the technique to better suit some situations and to improve the efficiency of its estimators. However, majority of these take interest only in the estimation of the population mean. Rarely do we see literature that utilizes RSS to estimate population ratios and assess how it performs compared to other techniques. The objective of this study is to actually explore more on the behaviour of the estimates using RSS compared to that of SRS.

2. Sampling Designs

This section gives a brief background on the different sampling techniques that would be utilized to obtaining sampling units in this study.

2.1. Simple Random Sampling

SRS is widely used because of its simplicity and intuitive appeal compared to other probability sampling designs. Simple random sampling with replacement (SRSWR) and simple random sampling without replacement (SRSWOR) are the two sampling procedures under SRS. In SRSWR, distinct element can appear more than once in a sample from a population where each one of the number of the population raised to number of samples i.e. N^n, has the same probability of selection. In SRSWOR, a distinct element can appear only once in a sample from a population wherein each one of the combinations of elements in the sample has the same chances of being selected in the sample.

Aside from means and proportions, ratios are usually of interest. For instance, estimating the income per capita and proportion of expenditures allotted for food are estimation problem of ratios. Income per capita is simply the total income divided by the number of household members. Proportion of expenditures allotted for food is the ratio of the food expenditures over the total expenditures of a given household.

In estimating the population ratio, there are two variables of interest denoted by Y and X, called the numerator and denominator characteristics, respectively. Let the population values be represented by $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)\}$ and the sample of size n obtained be represented by $\{(x_1, y_1), (x_2, y_2),\ldots, (x_n, y_n)\}$. Table 2.1 shows the estimator of the population ratio under SRSWOR, its approximate variance and estimated standard error as cited by Cochran (1977). In the absence of the population mean of X, μ_x , its sample mean $\overline{x}$ can be used instead. The estimator of the ratio is biased but the bias becomes negligible if the sample size is sufficiently large. Moreover, on the same assumption, its variance is approximately equal to its mean-squared error (MSE).

Table 2.1 SRSWOR Estimator of the Ratio, Its Approximate Variance and Estimated Standard Error

Parameter	Estimator	Approx. Variance	Estimated Standard Error
$R = \frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} X_i} = \frac{\mu_y}{\mu_x}$	$\hat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}}$	$\frac{1-f}{n\mu_x^2} \frac{\sum_{i=1}^{N} (Y_i - RX_i)^2}{N-1}$	$\frac{\sqrt{1-f}}{\sqrt{n}\mu_x} \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{R}x_i)^2}{n-1}}$

2.2. Ranked Set Sampling

Introduced by McIntyre (1952), RSS gives better estimates of the population mean and proportion in terms of efficiency and costs. This method ranks a small set of samples from the population using a concomitant variable or via judgment and eye inspection, and some elements from these sets are then selected in the sample. McIntyre (1952) recommended this procedure when there are complexities in determining the measurements of the samples. Recently, studies on RSS and its relative estimation performance compared with other sampling techniques are increasing. According to Patil (1995), Balanced RSS is a more precise method compared to SRS. When judgment ranking is perfectly done, RSS gives an unbiased estimator of the mean as cited by Wolfe (2004). Furthermore, its variance is always smaller than that of SRS.

The original scheme of the Balanced RSS technique is as follows. A set of m^2 units is randomly drawn from the population of interest. These units will be divided into m groups. Each group will be ranked according to a concomitant variable which is assumed to be highly linearly correlated with the target variable. The first order statistic will be taken from the 1st ranked set. The second order statistic will be taken from the 2nd ranked set and so on until the mth order statistics is taken from the m^{th} ranked set. The other m-1 units in each set will not be considered anymore. They are just used to determine the rank of the elements in each set. These steps are repeated r times to obtain the desired $n=mr$ elements in the sample that will be used to estimate the parameter of interest.

The ranking procedure is the most crucial part in this sampling design. It is strongly advisable to use a covariate or a concomitant variable that is highly positively correlated with the variable of interest, say Z, since it is difficult to rank by observations as noted by Patil et al. (1995). Even when there are ranking errors, RSS assures that estimator of the mean is still unbiased while the relative precision is dependent on how ranking is properly managed according to Chen (2004).

Estimating population ratios is quite different because there are two variables of interest. The ranking procedure becomes more complicated. The elements can either be ranked based on X or on Y. Samawi and Muttlak (1996) studied how to estimate ratios using RSS. They proposed estimators of the population ratio using RSS as shown in Table 2.2. The estimator of the ratio and its approximate variance remain the same whether the ranking is based on Y or X. Moreover, assuming that the bias of the estimators can be ignored, they showed that $Var(\hat{R}_{RSS}) \leq Var(\hat{R}_{SRS})$

Table 2.3 RSS Estimator of the Ratio, its Approximate Variance and Estimated Standard Error

Parameter	$R = \dfrac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} X_i} = \dfrac{\mu_y}{\mu_x}$
Estimator	$\hat{R}_{RSS} = \dfrac{\overline{Y}_{[n]}}{\overline{X}_{[n]}}$ where $\overline{Y}_{[n]} = \dfrac{1}{n}\sum_{i=1}^{n} Y_{(i)}$ and $\overline{X}_{[n]} = \dfrac{1}{n}\sum_{i=1}^{n} X_{(i)}$
Approx. Variance	$\dfrac{R^2}{n}\left\{V_x^2 + V_y^2 - 2\rho_{xy} V_x V_y - \left[\dfrac{\sum_{i=1}^{n} T^2_{x[i]}}{n\mu_x^2} + \dfrac{\sum_{i=1}^{n} T^2_{y[i]}}{n\mu_y^2} - 2\dfrac{\sum_{i=1}^{n} T^2_{xy[i]}}{n\mu_x^2\mu_y^2}\right]\right\}$ where $T_{x[i]} = \mu_{x[i]} - \mu_x$ $T_{y[i]} = \mu_{y[i]} - \mu_y$ $T_{xy[i]} = (\mu_{x[i]} - \mu_x)(\mu_{y[i]} - \mu_y)$ $\rho_{xy} = \dfrac{\sum_{i=1}^{N}(X_i - \mu_x)(Y_i - \mu_y)}{N\sigma_x\sigma_y}$
Standard Error	$\sqrt{\dfrac{1}{n\overline{X}^2_{[n]}}\left[s^2_{y[n]} + \hat{R}^2_{RSS} s^2_{x[n]} - 2\hat{R}_{RSS}\hat{\rho} s_{y[n]} s_{x[n]}\right]}$ (rank based on Y) $\sqrt{\dfrac{1}{n\overline{Y}^2_{[n]}}\left[s^2_{x[n]} + \hat{R}^2_{RSS} s^2_{y[n]} - 2\hat{R}_{RSS}\hat{\rho} s_{y[n]} s_{x[n]}\right]}$ (rank based on X) where $s^2_{y[n]}$ and $s^2_{x[n]}$ are the sample variances of Y and X, respectively and $\hat{\rho}$ is the sample correlation coefficient.

In their study, they assumed that the elements came from a bivariate normal distribution. On the assumption that there exists a linear relationship between X and Y and it is easier to rank based on X, they concluded that ranking based on the denominator characteristic is always recommended in estimating the ratio using RSS. However, in their simulation study, the sample sizes used were very small, i.e. n= 3, 5, 7, 9, 15, and 21. Furthermore, the paper failed to address the relationship of the concomitant variable to the two variables of interest. This study primarily aims to investigate the behaviour of the RSS and SRS for small and large population and sample sizes and to determine the effect in the efficiency of the RSS estimators if the concomitant variable is not strongly correlated with either of the two target variables.

3. Simulations

In this study, different simulation scenarios are considered in order to compare the performance of RSS and SRS in estimating the population ratio as shown in Table 3.1.

Table 3.1 Simulation Scenarios

Cases Considered	Scenarios	
Population Sizes (N)	(1) 10,000 (2) 7,000 (3) 5,000	
Sampling Rate (n)	(1) 1% of N (2) 3% of N (3) 5% of N	
Correlation of Concomitant Variable with the Target Variable	Low Variance	(1) Rank based on X (1.1) High i.e. Z=5*X+20+e, e ~N(0,15) (1.2) Moderately High i.e. Z=2+1.5*X+5*e, e ~N(0,15) (2) Rank based on Y (2.1) High i.e. Z = Y + 5*e, e ~N(0,1) (2.2) Moderately High i.e. Z=5+Y+10*e, e ~N(0,1)
	High Variance	(1) Rank based on X (1.3) High i.e.Z=2*X+5+e, e ~N(0,120) (1.4) Moderately High i.e. Z=X+5*e, e ~N(0,120) (2) Rank based on Y (2.3) High i.e. Z=5+Y+e, e ~N(0,1) (2.4) Moderately High i.e. Z=10+10*Y+4*e, e ~N(0,80)

The generated population consists of Y (numerator characteristic), X (denominator characteristic), and Z (concomitant variable). Population sizes and sampling rates are varied to determine the behaviour of the RSS and SRS estimates for small and large populations and sample sizes. RSS is expected to perform well in small populations while SRS in large populations.

Moreover, as mentioned earlier, the correlation of the concomitant variable is important in achieving greater efficiency of estimates. It must be noted that the available auxiliary variable is not always strongly correlated with the target variable so the authors investigated the case when the correlation is high and moderately high. Furthermore, since there are two variables of interest, ranking is done based on the numerator (Y) and denominator (X) characteristic. That is, the ranking in one of the two target variables is perfect while the second with ranking errors.

Also, SRS is known to work well on homogeneous populations while RSS on heterogeneous populations. This paper aims to determine how the two sampling design would behave depending on variability of the measurements in the population; hence, varying the population variance of the simulated data sets.

Nonparametric Bootstrap approach is used in this study in obtaining estimates of the ratios and their standard errors. Bootstrap is a resampling method used for determining the sampling distribution of an estimator. From an original sample, SRSWR is done repeatedly from which bootstrap estimates are taken.

In conducting nonparametric bootstrap estimation, 100 initial samples were selected from a population using either RSS or SRS. The size of the sample depends on the sampling rate considered, such as 1%, 3% or 5% of the population size. After obtaining the initial samples, SRSWR is performed in each sample to get 200 resamples. In each resample, the ratio is computed. Afterwards, the average of these resamples is calculated. Then, the arithmetic mean of the 100 average ratios of the resamples is already the bootstrap estimate of the population ratio and the standard deviation of the ratios is the estimated standard error of the bootstrap estimate.

4. Results and Discussion

This section presents the results of the simulation study. The bias is used to measure the validity of the RSS and SRS estimates. It is expressed in percent and computed using the formula given below:

$$\text{Relative Bias} = \left| \frac{\hat{\theta} - \theta}{\theta} \right| \times 100\%$$

Results presented on Tables 4.1-4.4 are for those scenarios wherein the population variance is low or, in other words, the population is homogeneous with respect to the variables of interest. On the other hand, Tables 4.5-4.8 show the results for heterogeneous populations.

Table 4.1 shows that for homogeneous populations with a concomitant variable that is highly correlated to the numerator characteristic (Y), regardless of the sample size and sampling rate, RSS performs better than SRS in terms of having smaller biases. Also, all RSS estimates of the ratios are more precise than those of SRS surprisingly. It must be noted as well that as the sampling rate increases, both the bias and the standard error become smaller for RSS.

Table 4.1 Bootstrap Estimates of Ratios, their Standard Errors and Biases for Different Population Sizes by Ranking on Numerator Characteristic (High Correlation)

Ranking on Y		RSS		SRS	
Population Size	Sampling Rate	Standard Error	Bias	Standard Error	Bias
10000	5%	0.029	0.02	0.0518	1.93
	3%	0.0389	0.06	0.059	2.8
	1%	0.0612	0.25	0.1038	2.5
7000	5%	0.0367	0.05	0.058	1.96
	3%	0.0491	0.21	0.0706	1.89
	1%	0.0755	0.30	0.1253	1.11
5000	5%	0.0419	0.07	0.0743	1.55
	3%	0.0525	0.25	0.0802	0.88
	1%	0.0896	0.30	0.1414	1.93

Table 4.2 shows that for homogeneous populations with a concomitant variable that is highly correlated to the denominator characteristic (X), in spite of the population size and sampling rate, SRS generally performs better than RSS. For all sample sizes, SRS estimates of the ratio are more precise since their standard errors are smaller than that of RSS. In terms of bias, only for the N=10000 with sample size sample size n=300 does RSS perform better.

For homogeneous populations with an auxiliary variable that has moderately high correlation with the numerator characteristic (Y), most of the biases of the estimates in RSS are smaller than to those of SRS as shown in Table 4.3. Only on the largest population size, N=10000, regardless of the sampling rate, does SRS perform better in terms of smaller bias. In terms of precision of estimates, RSS has the advantage.

Table 4.2 Bootstrap Estimates of Ratios, their Standard Errors and Biases for Different Population Sizes by Ranking on Denominator Characteristic (High Correlation)

Ranking on X		RSS		SRS	
Population Size	Sampling Rate	Standard Error	Bias	Standard Error	Bias
10000	5%	0.0591	0.63	0.0339	0.30
	3%	0.0772	0.14	0.0409	0.42
	1%	0.1187	0.56	0.067	0.01
7000	5%	0.0661	0.58	0.0394	0.26
	3%	0.0885	0.25	0.052	0.14
	1%	0.1589	1.42	0.0945	0.07
5000	5%	0.0823	0.31	0.0445	0.19
	3%	0.1012	1.21	0.0564	0.54
	1%	0.1468	2.78	0.1049	1.01

Table 4.3 Bootstrap Estimates of Ratios, their Standard Errors and Biases for Different Population Sizes by Ranking on Numerator Characteristic (Moderately High Correlation)

Ranking on Y		RSS		SRS	
Population Size	Sampling Rate	Standard Error	Bias	Standard Error	Bias
10000	5%	0.0402	0.36	0.0442	0.07
	3%	0.0541	0.44	0.0625	0.29
	1%	0.0584	1.24	0.1056	0.66
7000	5%	0.0952	1.38	0.0561	1.92
	3%	0.0657	0.72	0.0619	1.37
	1%	0.1124	0.33	0.1205	2.05
5000	5%	0.0608	0.16	0.0692	1.81
	3%	0.0797	0.51	0.0889	0.86
	1%	0.1171	0.52	0.1597	2.45

As shown in Table 4.4, SRS produced more estimates with smaller bias when the auxiliary variable used has moderately high correlation with the denominator variable. Furthermore, when the correlation becomes weaker, all SRS estimates are more precise.

Table 4.5 shows that when the population becomes heterogeneous in which the ranking variable used is highly correlated with the numerator characteristic, RSS and SRS almost have the same performance in terms of the bias of the estimates. The same performance can be noticed in terms of precision though SRS has more precise estimates for higher sampling rates compared to RSS.

Table 4.4 Bootstrap Estimates of Ratios, their Standard Errors and Biases for Different Population Sizes by Ranking on Denominator Characteristic (Moderately High Correlation)

Ranking on X		RSS		SRS	
Population Size	Sampling Rate	Standard Error	Bias	Standard Error	Bias
10000	5%	0.0406	0.3641	0.0307	0.0238
	3%	0.0789	0.7764	0.0414	0.0254
	1%	0.1192	0.2769	0.0768	0.0876
7000	5%	0.0715	0.0819	0.0385	0.2798
	3%	0.094	0.5057	0.0587	0.5667
	1%	0.1826	1.6995	0.09	0.542
5000	5%	0.0956	1.1064	0.0574	0.3168
	3%	0.1046	0.9193	0.0708	0.3042
	1%	0.1931	0.2392	0.1111	9.1247

Table 4.5 Bootstrap Estimates of Ratios, their Standard Errors and Biases for Different Population Sizes by Ranking on Numerator Characteristic (High Correlation)

Ranking on Y		RSS		SRS	
Population Size	Sampling Rate	Standard Error	Bias	Standard Error	Bias
10000	5%	0.0706	2.50	0.0417	13.80
	3%	0.0343	0.60	0.0417	6.09
	1%	0.0741	11.44	0.1078	9.88
7000	5%	0.0803	7.39	0.0454	4.72
	3%	0.0426	3.54	0.0651	9.17
	1%	0.0723	9.86	0.1175	2.95
5000	5%	0.121	21.35	0.0629	16.67
	3%	0.0501	0.78	0.0879	16.26
	1%	0.2056	53.57	0.1506	18.18

For heterogeneous populations with a highly correlated concomitant variable, SRS estimates are better for higher population sizes in terms of bias. RSS, on the other hand, works better for smaller population size. However, as shown in Table 4.6, SRS estimates are generally more reliable. Only two cases for RSS produced estimates with smaller standard errors, that is, for the small and medium population sizes, with a sampling rate of 1%.

Table 4.6 Bootstrap Estimates of Ratios, their Standard Errors and Biases for Different Population Sizes by Ranking on Denominator Characteristic (High Correlation)

Ranking on X		RSS		SRS	
Population Size	Sampling Rate	Standard Error	Bias	Standard Error	Bias
10000	5%	0.0738	0.47	0.0451	1.13
	3%	0.0877	2.06	0.0538	1.79
	1%	0.1408	16.48	0.1042	2.25
7000	5%	0.0538	4.92	0.0529	1.13
	3%	0.1011	12.31	0.0716	3.54
	1%	0.1101	0.33	0.1144	12.46
5000	5%	0.094	1.20	0.0675	4.50
	3%	0.0849	6.29	0.0806	12.34
	1%	0.1388	10.81	0.1575	14.86

Table 4.7 shows that, in general, RSS produced better estimates for a heterogeneous population with moderately high correlation between the numerator characteristic and the auxiliary variable. Reliability of estimates favoured RSS for larger populations while SRS estimates generally favoured smaller populations.

Table 4.7 Bootstrap Estimates of Ratios, their Standard Errors and Biases for Different Population Sizes by Ranking on Numerator Characteristic (Moderately High Correlation)

Ranking on Y		RSS		SRS	
Population Size	Sampling Rate	Standard Error	Bias	Standard Error	Bias
10000	5%	0.0406	4.22	0.0455	19.05
	3%	0.0477	5.26	0.0557	31.02
	1%	0.0882	11.90	0.1014	0.37
7000	5%	0.0447	3.60	0.0558	5.83
	3%	0.0683	1.56	0.0562	2.90
	1%	0.1176	3.19	0.1163	4.03
5000	5%	0.1181	1.99	0.0607	5.27
	3%	0.1472	17.04	0.0819	0.74
	1%	0.1176	0.58	0.1375	12.64

Table 4.8 shows that biases of the estimates from both of the sampling designs are of almost the same performance. However, all estimates using SRS gave more reliable estimates.

Table 4.8 Bootstrap Estimates of Ratios, their Standard Errors and Biases for Different Population Sizes by Ranking on Denominator Characteristic (Moderately High Correlation)

Ranking on X		RSS		SRS	
Population Size	Sampling Rate	Standard Error	Bias	Standard Error	Bias
10000	5%	0.063	1.77	0.0465	1.84
	3%	0.0888	0.81	0.057	4.90
	1%	0.1753	6.66	0.1178	3.60
7000	5%	0.1002	3.51	0.052	5.29
	3%	0.1128	15.12	0.0691	5.15
	1%	0.197	29.80	0.1213	4.10
5000	5%	0.1082	5.16	0.0559	1.24
	3%	0.123	0.15	0.0892	5.83
	1%	0.1896	0.95	0.1417	11.86

5. Illustration

The 2009 Food Income and Expenditure Survey (FIES) data from the National Statistics Office (NSO) was used to illustrate the comparison of the two sampling designs. We pretend that FIES dataset is a population in the conduct of sampling experiment. The target parameter for this illustration is the ratio of total food expenditure to total income or, simply, the proportion of income devoted to food expenditures. The concomitant variable used for total food expenditure (Y) was total expenditure having high correlation value of 0.82. For total income (X), total electricity expenditure was used as the auxiliary variable with moderately high correlation value of 0.62. Both variables have high degree of variability which means that the population is heterogeneous with respect to the variables of interest. There were a total of 38,400 households with population ratio R=0.3655. This means that 36.55% of the total household income is allotted to food expenditures.Tables 5.1 and 5.2 show the summary of results.

Table 5.1 Bootstrap Estimates of the Ratio, their Standard Errors and Biases for the 2009 FIES Data Set by Ranking on Food Expenditure (High Correlation)

Ranking on Total Income		RSS			SRS		
Population	Sampling Rate	Ratio	Standard Error	Bias	Ratio	Standard Error	Bias
38400	5%	0.3625	0.0033	0.8208	0.3708	0.0061	1.4501
	3%	0.3763	0.0086	2.9549	0.3821	0.0043	4.5418
	1%	0.3704	0.0129	1.3406	0.3757	0.0090	2.7907

Table 5.2 Bootstrap Estimates of the Ratio, their Standard Errors and Biases for the2009 FIES Data Set by Ranking on Total Income (Moderately High Correlation)

Ranking on Total Income		RSS			SRS		
Population	Sampling Rate	Ratio	Standard Error	Bias	Ratio	Standard Error	Bias
38400	5%	0.3671	0.0126	0.4378	0.3665	0.0064	0.2734
	3%	0.3686	0.0095	0.8482	0.3683	0.0105	0.7661
	1%	0.3650	0.0143	0.1368	0.3716	0.0104	1.6690

In accordance with the simulation results, RSS produced better estimates in terms of bias when the ranking is done over the numerator characteristic (Y) which is food expenditures in this case. Also, as expected, RSS performed better than SRS since the population is heterogeneous. As shown in Table 5.2, when the concomitant variable is not strongly correlated with the denominator characteristic (X), which is household income in this illustration, RSS is highly likely not to outperform SRS. This further confirms the results of the simulation study presented in the previous section.

6. Conclusions and Directions for Future Research

Based on the results of the simulations, the bootstrap estimates of the ratio using RSS performs better compared to that of SRS in terms of their relative bias when the ranking variable is based on numerator characteristic, Y. It is expected that SRS would perform better when the population of interest is homogenous. But the results of the simulation showed that RSS is superior over SRS when ranked on Y even when the variance is small. This may be due to the ranking variable used. This is also evident in the results using the denominator characteristic, X, as ranking variable where SRS performs better than its RSS counterpart given a homogenous population. These results are true when the ranking variable is highly correlated with the variable of interest. On the other hand, when the correlation is moderately high with a homogeneous population, the bootstrap estimates of SRS and RSS are almost comparable. For heterogeneous populations, the RSS estimates are generally superior over SRS. This is especially apparent when the ranking variable is highly correlated with the variable of interest.

The study is limited to comparing two sampling designs only. To better understand the advantages of RSS as a sampling technique, it would be better to compare its efficiency with other designs aside from the SRS. Furthermore, this study did not explore the possibility that Y is linearly correlated with X, or Y is a linear function of X.

References

CHEN, H., E. STANCY, and D. WOLFE, 2006, An empirical assessment of ranking accuracy in ranked set sampling, Computational Statistics & Data Analysis, 51, 1411-1419.

CHEN, Z. 2006, Ranked set sampling: Its essence and some new applications, Environmental and Ecological Statistics, 14, 355-363.

CHEN, Z. and Y. WANG, 2004, Efficient regression analysis with ranked-set sampling. Biometrics, 60 (4), 997-1004.

CHEN, H., E. STANCY and D. WOLFE, 2005, Ranked set sampling for efficient estimation of a population proportion. Statistics in Medicine, 24 (21), 3319-3329.

COCHRAN, W., 1977, Sampling Techniques, 3rd edition, John Wiley and Sons, New York.

MCINTYRE, G.A., 1952, A method of unbiased selective sampling using ranked sets. Australian Journal of Agricultural Research, 3, 385–390.

PATIL, G.P., 1995, Editorial: Ranked set sampling. Environmental and Ecological Statistics, 2, 271–285.

PATIL, G., A. SINHA and C. TAILLIE, 1995, Finite population corrections for ranked set sampling, Annals of the Institute of Statistical Mathematics, 47 (4), 621-636.

SAMAWI, H. and H. MUTTLAK, 1996, Estimation of ratio using ranked set sampling. Biometric Journal, 38 (6),753-764.

STOKES, S., 1977, Ranked set sampling with concomitant variables, Communications in Statistics - Theory and Methods, A6 (12), 1207–1211.

TAKAHASI, K. and K. WAKIMOTO, 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.

WOLFE, D., 2004, Ranked set sampling: An approach to more efficient data collection, Statistical Science, 19(4):636-643.