Investigating the Efficiency of Stratified Ranked Set Sampling Using Nonparametric Bootstrap Estimation

Kevin Carl P. Santos
University of the Philippines Diliman

Jenniebie C. Salagubang
Brady Philippines Direct Marketing Inc.

This paper aims to compare stratified random sampling and stratified ranked set sampling. A simulation study was conducted to evaluate the performance of the parameter estimates on both sampling techniques. Population sizes, sampling rates, stratum sizes, and correlation of the target variable and concomitant variable were varied, nonparametric bootstrap was then used in estimating the mean and its standard error. The coefficient of variation (CV) and the bias of the bootstrap estimates were compared. Stratified ranked set sampling generally outperforms stratified random sampling in terms of bias most especially for small populations. The two sampling designs were used in estimating the average mango production per *barangay* in the country.

Keywords: ranked set sampling, nonparametric bootstrap estimation, stratification, simple random sampling

1. Introduction

Sampling has been very essential in different areas of discipline. However, it becomes critical when the target population is very large and heterogeneous with respect to the variable of interest. To address the problem of heterogeneity, the population is subdivided into non-overlapping groups called strata. Then, independent samples are obtained in each stratum. These subsamples comprise the sample from the population of interest.

Stratification is a common strategy for large and heterogeneous populations. One reason for this as stated by Cochran (1977) is that gain in precision in the estimates can be obtained from stratification. The idea behind this technique is to divide the population into strata such that each stratum is homogenous within but they should be heterogeneous across each other. Thus, it is expected that precise estimates will be computed in each stratum and combining these weighted estimates across strata will give a better estimate for the parameter of interest.

Another reason in resorting to stratification as Kish (1965) mentioned is that strata may serve as domains of the study. This means that there can be acceptable estimates not only in the whole population but also in each stratum. Lohr (1999) added that stratification may lead to convenience in administering the survey and may result to lower costs.

As Kish (1965) pointed out, strata may be formed to employ different methods and procedures within them. The most common sampling design used in obtaining sample in each stratum is simple random sampling without replacement (SRSWOR). This is called Stratified Random Sampling (SSRS).

Furthermore, the most crucial issue in stratification is the choice of the stratification variable. This variable should stratify the population into strata which are homogenous within but heterogeneous across each other. According to Cochran (1977), the ideal variable to be used in stratification is the variable of interest itself. He also pointed out that using the variable of interest itself as stratification variable would result to non-overlapping strata and the variability within strata would be low. However, the target variable is usually unknown in practice. A solution to this problem is to look for an auxiliary variable, which is readily available and is believed to be correlated with the variable of interest, that will be a good stratification variable.

Another issue in stratification is the allocation of the sample sizes in the strata. There are three common allocation rules, namely: equal, proportional, and optimal allocations. Equal allocation is rarely used because of its impracticality. Proportional allocation takes into account the stratum sizes in allocating sample sizes while optimal allocation ensures that the variability and cost in each stratum would be included in the sample size determination. It was shown by Cochran (1977) that if the reciprocals of the stratum sizes are ignored relative to unity, the variance of the estimator of the mean under optimal allocation is the lowest compared to that under proportional allocation and under SRSWOR.

Another statistical technique for data collection that is of interest in this paper is Ranked Set Sampling. This was first introduced by McIntyre (1952) when it was difficult to take the actual measurements for sample observations. He wanted to gain precision in estimating the average yield from the large plots of arable crops. The measurements he needed were costly and tedious to collect. This led him to this approach in data collection.

McIntyre was not able to provide mathematical proofs of the optimal properties of the estimator of the population mean using RSS. Takahasi and Wakimoto (1968) provided the mathematical foundations of RSS showing that its estimator for the mean is unbiased even with different distributional assumptions under the assumption of perfect ranking. In addition to that, the RSS estimator of the mean is more efficient than that of SRS. It was shown that the relative precision (RP) of RSS compared to SRS estimator of the population mean is:

$$1 \leq RP = \frac{Var(\overline{X}_{SRS})}{Var(\overline{X}_{RSS})} \leq \frac{k+1}{2},$$

where k is the set size. This means that as the set size k increases, the RP also increases. Hence, increasing the set size will result to a more reliable estimate for RSS compared to that of SRS. Nevertheless, it should be noted that taking a large set size would entail higher costs in obtaining samples. This is the reason why k is not usually high in practice.

The ranking criterion is the most important issue in RSS. Before, ranking was done visually or through eye inspection but because of ranking errors, RSS might be worse than other sampling designs. Similar to sampling with probability proportional to size (PPS), an auxiliary or concomitant variable can be used to rank the sampling units. This auxiliary variable should be highly and directly associated with the characteristic of interest or should be a frugal measure of the variable of interest. RSS was extended to ranking using a concomitant variable by Stokes (1977). He concluded that the reliability of the estimates depends on the degree of the relationship of the two variables.

Chen (2007) enumerated applications of RSS in many areas of discipline such as agriculture, environment and ecology, medicine, and genetics. Samawi and Muttlak (1996) introduced the so-called Extreme RSS. This has been applied in genetics for quantitative trait loci (QTL) mapping to measure obesity and cholesterol level. Chen and Wang (2004) used RSS in studying lung cancer. They investigated how lung cancer is affected by smoking through the use of bio-markers. Another novel application of RSS is in comparing treatments in experiments which include many clinical trials.

Since it was shown that RSS is better than SRS in estimating the population mean in terms of efficiency, this paper aims to determine whether Stratified Ranked Set Sampling (SRSS) will perform better than Stratified Random Sampling (SSRS). SRSS performs ranked set sampling in each stratum independently and combining these RSSs will comprise the sample from the whole population.

Ibrahim et al. (2010) already compared SRSS, RSS and SSRS with Stratified Median Ranked Set Sampling (SMRSS). They have shown that SMRSS estimator is an unbiased estimator of the population mean when the population is symmetric and is more efficient in estimating the population mean compared to the other abovementioned sampling designs. However, their simulation study did not consider different stratum sizes in the population and different degrees of correlation of the auxiliary variable used in ranking. Additionally, their sample sizes are very small (*n* = 7, 12, 14, 15, 18). These sample sizes will still be allocated to 2-3 strata in their study. This study wants to investigate these things since Ibrahim et al. did not take these into account. Another difference of this paper and theirs is the assumption of the distribution of the population in the simulation study. This simulation study is limited to normal populations only while Ibrahim et al. investigated other distributions as well.

Different scenarios in estimating the population mean are considered in the simulations done in this study. Section 2 discusses the details of the different sampling designs. Section 3 elaborates on the different cases taken into account in the simulation. Section 4 shows the results of the simulations while Section 5 includes conclusions and directions for future research.

2. Sampling Designs

This section gives a discussion on the sampling schemes that were used in the paper.

2.1. Simple Random Sampling without Replacement (SRSWOR)

Simple Random Sampling (SRS) is the simplest form of probability sampling design. A simple random sample of size *n* is taken from the population wherein all possible samples of size *n* are given the same chance of selection. SRS can be done with replacement or without replacement.

In Simple Random Sampling without replacement (SRSWOR), each possible combination of *n* different elements out of *N* has the same chance of being selected in the sample. To obtain an SRSWOR, each element in the sampling frame will be assigned with a unique number. Afterwards, *n* distinct numbers will be drawn using a random process. The elements associated with the distinct numbers will comprise the sample. The table below shows how the mean is estimated using SRSWOR.

Table 2.1 SRSWOR Estimator of Population Mean, its Variance and Estimator of the Standard Error

Estimator	Variance	Estimated Standard Error
$\bar{y}_{SRSWOR} = \frac{1}{n}\sum_{i=1}^{n} y_i$	$\frac{S^2}{n}\frac{N-n}{N}$	$\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N}}$ where s^2 is the sample variance

The sample mean $\bar{y}_{SRSWOR}$ is an unbiased and consistent estimator of the population mean. The estimator of the variance of the sample mean is also unbiased. However, their square roots or the estimators of the standard errors are slightly biased for the true standard error of the sample mean. This sampling design is used in obtained independent samples in each stratum in SSRS.

2.2. Ranked Set Sampling

Ranked Set Sampling (RSS) requires drawing a simple random sample (SRS) of size k from a sampling frame. Afterwards, these k elements will be ranked using either visual inspection or a readily available concomitant variable. As a consequence, the sample frame should also include auxiliary information that will be used to rank the elements. It must be noted that these k units are not necessarily obtained physically. For instance, a mapping can be done initially on the concomitant variable before actually measuring the variable of interest.

The first order statistic from the ranked units will be the first unit in the sample. Then, another SRS of size k will be drawn and these will be ranked again. The second order statistic from this set will be the second unit in the sample. This is repeated until the k^{th} order statistic is obtained for the k^{th} batch of SRS. This whole process is called a cycle and k is called the set size. If a cycle would be repeated m times, the total number of units in the sample is $n = mk$.

Table 2.2 shows the RSS estimator of the mean. It is an unbiased estimator of the population mean and its variance is smaller than that of SRS. The estimator of the variance of the sample mean under RSS is biased but the bias is a function of the set size and number of cycles. Hence, as the set size or number of cycles increases, the bias is expected to be negligible as stated by Chen et al. (2004).

Table 2.2 RSS Estimator of Population Mean, its Variance and Estimator of the Standard Error

Estimator	Variance	Estimated Standard Error
$\bar{y}_{RSS} = \frac{1}{mk}\sum_{r=1}^{k}\sum_{i=1}^{m} y_{[r]i}$	$\frac{\sigma^2}{mk} - \frac{1}{m^2 r}\sum_{i=1}^{m}(\mu_{[i:m]} - \mu)^2$	$\sqrt{\frac{1}{mk-1}\sum_{r=1}^{k}\sum_{i=1}^{m}(y_{[r]i} - \bar{y})^2}$

2.3 Stratified Simple Random Sampling

When the population is heterogeneous with respect to the variable of interest, SRS should not be used because its standard error is expected to be large. The idea is that there exists a variable that can divide the population to obtain l homogeneous sub-groupings. This variable is called the stratification variable. In which case, SRS can be used in each homogenous grouping. This is the so-called Stratified Simple Random Sampling (SSRS) or simply Stratified Random Sampling.

In this sampling design, SRSWOR is conducted in obtaining samples in each stratum. Sampling is done independently across strata. The estimator of the mean is simply the weighted mean of the stratum means where the weight is the ratio of the stratum size to the population size. Due to the independence of sampling across strata, the variance of the estimator under SSRS is simply the weighted variances of each stratum under SRSWOR as shown in Table 2.3. It is expected that the variance of $\overline{y}_{SSRS}$ is lower than that of SRS when the population is very heterogeneous and an appropriate stratification variable is chosen.

Table 2.3 SSRS Estimator of Population Mean, its Variance and Estimator of the Standard Error

Estimator	Variance	Estimated Standard Error
$\overline{y}_{SSRS} = \sum_{h=1}^{l} \frac{N_h}{N} \overline{y}_h$	$\sum_{h=1}^{l} W_h^{\ 2} \frac{S_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right)$	$\sqrt{\sum_{h=1}^{l} W_h^{\ 2} \frac{s_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right)}$
	where $W_h = \frac{n_h}{N_h}$	

Another issue in stratification is the sample size determination in each stratum. Some of the common allocation rules are equal, proportional, and optimal allocation. The allocation rule used in this study is limited to proportional allocation. This method allocates the samples proportional to size of the stratum. In addition to that, it was shown in Cochran (1977) that the variance of the estimator of the sample mean under proportional allocation is smaller than that using SRSWOR.

2.4 Stratified Ranked Set Sampling

Since stratification is used to homogenize the population into strata and, as shown by Patil (2002) and Wolfe (2004), the variance of the estimator of the mean under RSS is lower than SRS, the authors wanted to investigate the reliability of the estimates when RSS is used to obtain samples in each stratum instead of using SRSWOR.

The process discussed in Section 2.2 is performed in each stratum and the process is repeated *m* times. The derivations of the estimator of the mean and its variance are similar to that of SSRS. The estimator of the mean is simply the weighted RSS estimator in each stratum where the weight is the ratio of the stratum size to the population size. The variance of $\bar{y}_{SRSS}$ and its estimator are derived similarly by substituting the variance of $\bar{y}_{RSS}$.

Table 2.4 SRSS Estimator of Population Mean, its Variance and Estimator of the Standard Error

Estimator	Variance	Estimated Standard Error
$\bar{y}_{SRSS} = \frac{1}{mk}\sum_{h=1}^{l}\sum_{r=1}^{k}\sum_{i=1}^{m}\frac{N_h}{N} y_{[r]ih}$	$\sum_{h=1}^{l} W_h^{\;2}\frac{\sigma_h^2}{m_h k_h} - \sum_{h=1}^{l}\sum_{i=1}^{m}\frac{W_h^2}{m_h^2 k_h}(\mu_{[i:m]h}-\mu_h)^2$ where $W_h = \frac{n_h}{N_h}$	$\sqrt{\sum_{h=1}^{l}\sum_{r=1}^{k}\sum_{i=1}^{m}\frac{W_h^2}{m_h k_h - 1}(y_{[r]ih}-\bar{y})^2}$

3. Simulation Studies

Different scenarios were considered to represent real-life situations and evaluate the behavior of the parameter estimates based on these scenarios using the proposed sampling design. Two sampling designs, namely, Stratified Simple Random Sampling (SSRS) and Stratified Ranked Set Sampling (SRSS) would be compared in this study. Table 3.1 below shows the different cases considered.

Table 3.1 Cases Considered in the Simulation Study

Cases Considered	Scenarios
Population Sizes (N)	(1) 1,200 (2) 12,000
Stratum Sizes	(1) 33-33-33% (balanced) (2) 20-30-50% (moderately unbalanced) (3) 15-25-60% (unbalanced)
Sampling Rate (n)	(1) 1% of N (2) 3% of N (3) 5% of N
Correlation of Concomitant Variable with the Target Variable	(1) High i.e. X = 2+ 5*Y + e, e ~N(0,1) (2) Moderately High i.e. X = 2 + 5*Y + 5*e, e ~N(0,1)

Only two population sizes were considered. The large populations are represented by N=12,000 while the small populations are represented by N=1,200. Usually, SRS performs well in large populations while RSS has an edge in small populations. This paper would like to determine if this will still hold true when stratification is conducted in the population.

To investigate the performance of the two sampling designs, the stratum sizes are also varied. The number of strata is limited to 3 only. Sampling rates are 1%, 3%, and 5% of the population sizes. This is to determine the behavior of the two sampling designs when the sample size increases.

In addition, the set size *k* is limited to 3-5 only to avoid propagation of ranking error. The set sizes were varied depending on the sample size being divisible by 3, 4 or 5.

This paper also investigates the effect of the correlation of the concomitant variable and the target variable. Since the auxiliary variable is used in ranking, there may be an effect in the results of RSS estimator. In Stratified Ranked Set Sampling, the correlation between target variable and concomitant variable has two cases: high and moderately high. The concomitant variable should be strongly correlated with the target variable so the case of low correlation is not considered in this study.

The study assumed normally distributed simulated populations and applied proportional allocation scheme for the three strata. Two kinds of data set were used. The first one assumes that the elements within strata were homogeneous while the variances within and across strata were varied for the other data set. Descriptions of the simulated data are given below in Table 3.1.

Table 3.1 Simulated Finite Population Means, Variances, and CVs

Data	Stratum sizes	Mean	Variance	CV (in %)
Data Set 1	All	25	10	12.6
		50	10	6.3
		75	10	4.2
Data Set 2	33-33-33	25	100	40
		50	400	40
		75	900	40
	20-30-50	25	156.25	50
		50	225	30
		75	126.5625	15
	15-25-60	25	225	60
		50	506.25	45
		75	56.25	10

As shown in Table 2.4, the estimated standard error of SRSS is quite complicated because it requires the computation of the averages of the order statistics across all cycles, which is needed to be done across the strata. This is why there are three indices in the formula of the estimated standard error. Moreover, the inclusion probabilities would be very hard to compute since RSS is done across strata. Because of these, the authors used variance estimation method. Linearization using Taylor series expansion would be difficult to perform so Nonparametric Bootstrap Method was used instead. This is a re-sampling method to determine the empirical distribution of the estimator. The goal of this paper is to determine the empirical or sampling distribution of the sample mean. The bootstrap method was performed for both SSRS and SRSS for comparison purposes. Along with this method is the estimation of the variance of the bootstrap estimate of the mean.

The nonparametric bootstrap uses simple random sampling with replacement (SRSWR) in re-sampling. Glivenko-Cantelli Lemma justifies the use of nonparametric bootstrap. Nonparametric bootstrap estimation has not yet been used in RSS based on literature.

The procedure for the estimation procedure is as follows: Obtain a sample in each stratum using SRSWOR for SSRS and RSS for SRSS. Only 3 strata were considered in this study. Then, once the samples are obtained, nonparametric bootstrap will be performed 200 times in each stratum. Then, the weighted mean will be computed in each bootstrap sample. In this case, there will be 200 weighted means where the weight is $W_h = N_h/N$. Afterwards, the arithmetic mean of these 200 weighted means will be calculated. This will be the first bootstrap estimate of the mean.

Subsequently, the Monte Carlo variance of these 200 weighted means will be computed. This variance will be the estimated variance of the 1st bootstrap mean. This entire process will be repeated 100 times. After which, there will be 100 bootstrap estimates of the mean and of the variance of the sample mean. The average of these 100 bootstrap estimates will be the estimated mean and its estimated variance using nonparametric bootstrap.

4. Results and Discussion

This section presents the results of the simulation study. The tables include the Monte Carlo variance, bias of the estimate and the coefficient of variation, expressed as percentage, of the bootstrap estimate. The bias and CV of the bootstrap estimate were used as measures of validity and reliability of the estimates. Tables 4.1-4.4 show the results for the first kind of data set which assumes homogeneity of the elements within each stratum. Tables 4.5-4.8 present the results for data sets wherein the variances among strata were varied.

Table 4.1 shows that for the small populations (N=1200), regardless of the stratum sizes, most of the SRSS estimates of the means are closer to the population means. The highlighted biases are those cases in which the SRSS has smaller bias than that of SSRS. However, for all the nine cases, SSRS estimates are more precise since they have smaller CVs as compared to SRSS estimates.

Table 4.11 Bootstrap Estimates of the Mean, their Variances, Biases and Coefficients of Variation (CV) for Small Population (High Correlation)

Population Size 1,200		SSRS			SRSS		
High Correlation							
Stratum Sizes	Sampling Rate	Var	Bias	CV	Var	Bias	CV
33-33-33%	1%	5.98	1.60	**4.97**	8.15	**0.93**	5.76
	3%	2.67	**0.79**	**3.29**	2.74	1.01	3.34
	5%	1.68	1.05	**2.62**	1.70	**1.01**	2.63
20-30-50%	1%	2.65	**0.28**	**2.82**	5.41	0.66	4.06
	3%	2.35	0.23	**2.67**	2.52	**0.11**	2.75
	5%	1.59	0.23	**2.19**	1.65	**0.00**	2.23
15-25-60%	1%	1.63	0.15	**2.07**	5.92	**0.07**	3.96
	3%	2.62	**0.07**	**2.63**	2.69	0.09	2.67
	5%	1.68	0.20	**2.10**	1.70	**0.08**	2.12

Table 4.2 shows that for small population and having an auxiliary variable with moderately high correlation with the target variable; SSRS yield CVs which are lower than that of SRSS. Thus, in general, SSRS still performs a little better than SRSS in terms of reliability of its estimates. Note that their CVs are not so different from each other. In terms of bias, SRSS has an edge over SSRS because SRSS bootstrap estimates have smaller bias in 7 out of 9 scenarios. Even if the CVs of SRSS are a little larger than that of SSRS, the values of its bootstrap estimates are much closer to the true value of the parameter.

In addition, the CVs for SRSS are smaller than that of SSRS when the sampling rate is 5% in the two unequal stratum sizes cases. When the population is small and the stratum sizes are all equal, the bootstrap estimates of the mean are less biased under SSRS if the sampling rates are 1% and 3%.

Table 4.2 Bootstrap Estimates of the Mean, their Variances, Biases and CVs for Small Population (Moderately High Correlation)

Population Size 1,200		SSRS			SRSS		
High Correlation							
Stratum Sizes	Sampling Rate	Var	Bias	CV	Var	Bias	CV
33-33-33%	1%	6.63	**0.45**	**5.13**	7.68	1.05	5.60
	3%	2.55	**1.14**	**3.23**	2.70	1.37	3.33
	5%	1.65	1.22	**2.60**	1.73	**1.06**	2.66
20-30-50%	1%	3.75	2.92	**3.27**	7.27	**0.07**	4.68
	3%	2.72	3.59	2.97	2.50	**0.01**	**2.75**
	5%	1.81	0.23	2.33	1.67	**0.02**	**2.25**
15-25-60%	1%	4.47	0.38	**3.43**	6.04	**0.19**	3.99
	3%	2.43	1.85	**2.49**	2.77	**0.09**	2.71
	5%	1.94	0.58	2.28	1.67	**0.50**	**2.12**

Table 4.3 shows that almost all of the SRSS estimates have smaller biases, hence closer to population means, than their SSRS estimates counterparts. The only two scenarios in which SSRS estimates are closer to the population mean is the equal stratum sizes case with sampling rates of 3% and 5%. On the other hand, for all the stratum sizes considered, SRSS have lower CVs at the 5% sampling rate.

As shown in Table 4.4, only 3 out of 9 CVs under SRSS are lower than that of SSRS. But, it can be noticed that their CVs are not so different from each other. There is just a minimal difference in their CVs. In terms of the bias of the bootstrap estimates, SRSS has lesser bias compared to SSRS in general. Thus, the SRSS bootstrap estimates are closer to the true value of the population mean.

Moreover, large bias, in general, was observed for the equal stratum sizes case for the three sampling rates for both sampling designs. It can be noticed as well that in the two unbalanced cases, the bias of the estimates under SRSS is very small.

Table 4.3 Bootstrap Estimates of the Sample Mean, their Variances, Biases and CVs for Large Population (High Correlation)

Population Size 1,200		SSRS			SRSS		
Stratum Sizes	High Correlation Sampling Rate	Var	Bias	CV	Var	Bias	CV
33-33-33%	1%	0.83	1.10	1.84	0.83	**0.95**	**1.84**
	3%	0.28	**0.92**	**1.06**	0.28	1.04	1.06
	5%	0.17	**0.93**	0.82	0.17	0.98	**0.82**
20-30-50%	1%	0.82	0.12	**1.57**	0.84	**0.11**	1.59
	3%	0.28	0.06	0.93	0.28	**0.01**	**0.92**
	5%	0.17	0.03	0.71	0.17	**0.02**	**0.71**
15-25-60%	1%	0.81	0.09	**1.46**	0.85	**0.03**	1.50
	3%	0.28	0.11	**0.86**	0.28	**0.05**	0.86
	5%	0.17	0.11	0.67	0.17	**0.01**	**0.67**

Table 4.4 Bootstrap Estimates of the Mean, their Variances, Biases and CVs for Large Population (Moderately High Correlation)

Population Size 1,200		SSRS			SRSS		
Stratum Sizes	High Correlation Sampling Rate	Var	Bias	CV	Var	Bias	CV
33-33-33%	1%	0.80	1.22	**1.81**	0.83	**1.10**	1.84
	3%	0.28	1.12	1.07	0.28	**0.86**	**1.07**
	5%	0.16	1.09	**0.82**	0.17	**0.90**	0.82
20-30-50%	1%	0.82	0.25	**1.57**	0.85	**0.12**	1.60
	3%	0.28	0.15	**0.91**	0.28	**0.06**	0.92
	5%	0.15	0.16	**0.66**	0.17	**0.07**	0.72
15-25-60%	1%	0.84	0.23	1.49	0.82	**0.06**	**1.48**
	3%	0.28	**0.03**	**0.87**	0.29	0.11	0.8
	5%	0.17	0.05	**0.67**	0.17	**0.01**	**0.66**

The succeeding tables show the results when the variances across the strata differed from each other. When the concomitant variable is strongly correlated with the target variable, the bias of SRSS mean estimator is smaller than that of SSRS in general for small populations. The SSRS estimates for the population mean have smaller CVs compared to that of SRSS. This result is consistent with the previous results. Table 4.6 shows that SRSS is still better in terms of the bias of the bootstrap estimates most especially in balanced and unbalanced cases. The result is quite different in the moderately unbalanced case. In general, SRSS performs better in terms of lower CVs compared to SSRS.

Moreover, it can also be noticed that the bias of the estimates in the unbalanced cases decreases in SRSS while it increases in SSRS as the sample size blows up. The moderately unbalanced case has a different behavior in SRSS but in SSRS, the bias declines as the sample size increases.

Table 4.5 Bootstrap Estimates of the Mean, their Variances, Biases and Coefficients of Variation (CV) for Small Population (High Correlation)

Population Size 1,200		SSRS			SRSS		
High Correlation Stratum Sizes	Sampling Rate	Var	Bias	CV	Var	Bias	CV
33-33-33%	1%	23328.78	**0.95**	**302.87**	21434.24	23.15	381.37
	3%	8820.80	17.20	**160.41**	7719.61	**1.92**	172.57
	5%	23294.07	80.21	1543.96	4719.82	**5.98**	**146.27**
20-30-50%	1%	1262.82	7.62	**56.69**	1264.09	**0.03**	61.03
	3%	640.19	**4.98**	**41.38**	719.74	5.10	48.54
	5%	458.50	4.31	38.42	451.66	**1.55**	**37.06**
15-25-60%	1%	3695.28	**0.40**	**94.32**	6859.75	25.17	102.26
	3%	1703.53	4.40	**61.09**	1925.82	**1.15**	67.05
	5%	957.09	62.32	**29.45**	2891.76	**0.31**	83.36

Table 4.6 Bootstrap Estimates of the Mean, their Variances, Biases and CVs for Small Population (Moderately High Correlation)

Population Size 1,200		SSRS			SRSS		
High Correlation Stratum Sizes	Sampling Rate	Var	Bias	CV	Var	Bias	CV
33-33-33%	1%	21160.46	53.09	620.71	2635.98	**5.05**	**97.83**
	3%	7985.29	13.68	**157.36**	8962.15	**1.66**	192.71
	5%	5317.99	10.91	**163.85**	5472.06	**9.86**	164.27
20-30-50%	1%	1277.33	11.71	69.50	1408.97	**3.68**	**62.16**
	3%	675.81	**0.09**	44.60	684.87	0.98	**44.50**
	5%	443.10	**2.38**	**37.02**	487.48	5.01	39.91
15-25-60%	1%	4278.22	21.52	128.80	5213.31	**4.11**	**107.17**
	3%	1866.69	1.79	**67.99**	1951.74	**0.20**	68.14
	5%	1180.76	2.00	52.06	1148.62	**1.87**	**51.41**

Table 4.7 shows that the CVs of the bootstrap estimates under SSRS are generally lower than that of SRSS when population is large. This is not surprising since in large and heterogeneous populations, SSRS works well. In terms of the bias, the two sampling designs are almost comparable.

When the concomitant variable is moderately correlated with the target variable, SRSS performs better than SSRS in large populations as listed in Table 4.8. Generally, the CVs of SRSS are lower than that of SSRS. Furthermore, the bias and CVs of both sampling designs decrease as the sample size increases in general. In the moderately unbalanced (20-30-50%) and unbalanced (15-25-60%) cases, the CVs of the two sampling designs are comparable.

Table 4.7 - Bootstrap Estimates of the Sample Mean, their Variances, Biases and CVs for Large Population (High Correlation)

Population Size 1,200		SSRS			SRSS		
	High Correlation						
Stratum Sizes	Sampling Rate	Var	Bias	CV	Var	Bias	CV
33-33-33%	1%	2683.09	**2.76**	107.14	2772.65	5.08	**100.79**
	3%	908.87	2.49	**59.16**	915.26	**0.78**	61.33
	5%	550.78	**0.82**	**47.60**	536.65	2.65	47.86
20-30-50%	1%	233.44	1.80	**25.66**	240.41	**0.36**	26.60
	3%	80.58	**0.95**	**15.20**	80.33	2.17	15.66
	5%	46.64	**0.77**	**11.77**	48.06	0.81	11.95
15-25-60%	1%	78.87	14.03	**15.64**	134.77	**12.80**	20.15
	3%	22.81	15.73	**8.58**	50.91	**14.48**	12.63
	5%	15.57	**11.35**	**6.74**	33.47	14.97	10.30

Table 4.8 Bootstrap Estimates of the Mean, their Variances, Biases and CVs for Large Population (Moderately High Correlation)

Population Size 1,200		SSRS			SRSS		
	High Correlation						
Stratum Sizes	Sampling Rate	Var	Bias	CV	Var	Bias	CV
33-33-33%	1%	2646.58	4.95	108.87	2709.01	**0.55**	**104.12**
	3%	905.14	3.07	62.43	910.25	**0.03**	**60.71**
	5%	547.20	**9.50**	51.99	531.98	26.99	**36.53**
20-30-50%	1%	231.41	**0.42**	**25.90**	224.52	2.39	26.24
	3%	78.96	2.26	15.54	78.80	**1.34**	**15.38**
	5%	46.87	**0.41**	**11.75**	48.11	1.38	12.02
15-25-60%	1%	572.41	**5.33**	**36.08**	592.98	5.49	36.67
	3%	202.43	4.99	23.79	204.41	**1.60**	**22.35**
	5%	123.23	1.29	17.87	123.50	**0.06**	**17.64**

It should be noted that nonparametric bootstrap was used in the estimation procedure. Thus, the unbiasedness of the estimators under SRSS and SSRS for the population mean mentioned in Section 2 does not apply anymore because that is under design-unbiased estimation. This is why the estimates are biased even though repeated sampling was done.

In general, the CV of SRSS bootstrap estimate of the mean is larger than that of SSRS because the sample obtained using RSS is spread out. SRSS estimates are closer to the actual value producing small bias. In SSRS, the bootstrap estimates of the mean have smaller CVs generally because the measurements obtained using SRSWOR are near the true mean.

5. Illustration

The two sampling designs were illustrated using the 2002 Census of Agriculture data set. The target variable is the production of mangoes in the country. The concomitant variable used is the corn production. This is because mangoes and corns are usually planted together. Instead of using the information per farmer, the variables were aggregated in the barangay level. Thus, the elementary units considered in this illustration are the barangays. The parameter that will be estimated in this case is the average production of mangoes in the country.

Furthermore, the three major islands namely Luzon, Visayas, and Mindanao were defined as the strata. Only those records with complete information on the necessary variables such as stratum, region, province, municipality, barangay, farm areas for mango and corn production were considered in this illustration; hence, there are a total of 1,144 records used.

In the first stratum, Luzon, there are a total of 460 barangays while the second stratum, Visayas, has 210 barangays and lastly, Mindanao, has 474 barangays. Based on these, the weights used per stratum are 40-18-42% respectively. The two sampling designs were used to obtain 1%, 3%, 5%, and 10% of the population as the sample sizes. Based on Table 4.9, it is evident that SRSS works very well as compared to SSRS since it has lower biases and CVs for all cases except only when the sample size is at 3%. This result further confirms the results of the simulation study. That is, SRSS works well with small population and the bootstrap estimate under this sampling design produces smaller bias.

Table 4.9 Bootstrap Estimates of the Mean, their Variances, Biases and CVs for the 2002 Census of Agriculture

Population Size 1,200		SSRS				SRSS			
High Correlation Stratum Sizes	Sampling Rate	Mean	Var	Bias	CV	Mean	Var	Bias	CV
	1%	9.92	27.98	89.34	53.31	3.05	0.80	**41.73**	**29.37**
40-18-42%	3%	2.95	0.24	43.64	**16.51**	6.02	2.65	**14.88**	27.02
	5%	6.86	7.98	30.83	41.20	3.91	0.70	**25.30**	**21.4**
	10%	4.66	2.07	11.09	30.91	5.18	1.05	**1.11**	**19.82**

6. Conclusion and Direction for Future Research

Based on the simulation scenarios, the bias of the bootstrap estimates using Stratified Ranked Set Sampling is generally smaller compared to that of Stratified Random Sampling most especially in small populations regardless of the sample size obtained and the degree of variability in each stratum. This means that whether the stratum sizes are different (balanced, moderately unbalanced or unbalanced) or the elements in each stratum are very heterogeneous, the bias of the bootstrap estimate of SRSS is smaller than that of SSRS. Hence, the bootstrap estimate of the mean under SRSS is expected to be closer to the true mean. The measure of reliability may suffer because the CVs of SRSS are larger compared to that of SSRS. This result is consistent with the fact that the sample obtained using SRSS is spread out in the population giving enough representation of the population of interest.

If the researcher wants an estimated value much closer to the population mean, Stratified Ranked Set Sampling is recommended most especially when the population is small regardless of the variability of strata in the population. This implies that even if the stratification variable were not appropriately chosen, SRSS would more likely to give less biased estimate of the mean provided that the population is small. Furthermore, it was shown that the SRSS gives less biased estimates provided that the correlation of the concomitant variable and target variable is at least moderately strong.

On the other hand, if the researcher puts a heavier weight on the reliability of the estimates, then Stratified Random Sampling works better since SSRS has lower coefficient of variation (CV) in general. This holds true for large populations regardless of the stratum sizes, sampling rates, and the variability of elements in each stratum.

This paper has several limitations. The number of strata is fixed to 3. It might be of interest to determine the effect of the number of strata in the estimation procedure. Moreover, Ibrahim et al. (2010) studied different distributions of the population. It would be appealing to study the performance of SRSS and SSRS under different distribution assumptions other than normal and under different population sizes as well.

Furthermore, the study is limited in comparing the two sampling designs only. It would be better if the usual RSS were compared as well. In practice, it is possible that the auxiliary variable used has a low correlation with the target variable if it is the only information available to the researchers. Perhaps, it is also a good idea if this scenario would be added in the simulation study.

REFERENCES

CHEN, Z., BAI, Z., and SINHA, B., 2003, *Ranked Set Sampling: Theory and Application*, Springer.

CHEN, Z., 2007, Ranked Set Sampling: Its Essence and Some New Applications. *Environ. Ecol. Stat.*, 14:355-363.

CHEN Z. and WANG Y., 2004, Efficient Regression Analysis with Ranked Set Sampling, *Biometrics* 60:997-1004.

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

IBRAHIM, K. et al., 2010, Estimating the Population Mean Using Stratified Median Ranked Set Sampling, *Applied Mathematical Sciences*, 4(47):2341-2354.

KISH, L., 1965, *Survey Sampling*. John Wiley and Sons, New York.

LOHR, S., 1999, *Sampling: Design and Analysis*. Duxbury Press, Pacific Grove, CA.

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.

STOKES, S., 1977, Ranked Set Sampling with Concomitant Variables, *Communications in Statistics - Theory and Methods*, A6 (12): 1207-1211.

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.

SAMAWI, H.M. & MUTTLAK, H.A., 1996, Estimation of Ratio using Rank Set Sampling, *Biomet J* 38:753-764.

WOLFE, D., 2004, Ranked Set Sampling: An Approach to More Efficient Data Collection, *Statistical Science*, 19(4):636-643.

ACKNOWLEDGMENT

The authors would like to express their sincerest gratitude to Joanne Guansing for sharing her expertise in programming and Dr. Erniel Barrios for giving insightful comments which improved this paper.