# Comparing Different Sampling Schemes to Estimate the Population Mean: A Simulation Study

I.Sarkar[1], P.Shrestha[2], C. Rosenbaum[1], E. Oral[1]*

[1]LSUHSC School of Public Health Biostatistics Program, [2]LSUHSC School of Public Health Epidemiology Program

LSU Health NEW ORLEANS School of Public Health

*Louisiana State University Health Sciences Center School of Public Health*

*Contact: eoral@lsuhsc.edu

## Motivation and Objectives

The most common methods of probability sampling are Simple Random Sampling (SRS), Stratified Random Sampling (StRS) and Cluster Sampling (CS). This study compares the empirical bias and variance of the sample mean from SRS, StRS and CS via an extensive Monte-Carlo simulation study.

## Background

To date there has been no study that compared the empirical biases and variances from SRS, StRS and CS. In a **SRS**, every member of the population is chosen randomly and has an equal chance of being selected. Sampling frame should include the whole population. **StRS** is used when the sampling units associated with the population can be separated into two or more homogeneous groups (strata) where the within-stratum response variation is less than the variation within the entire population. After defining the strata, SRS is applied separately to each stratum. In **CS**, each sampling unit is a collection of subjects. It is more convenient for geographically dispersed populations and effective when sampling frame is not available. It is the least representative sampling scheme among the three. It is less expensive than SRS but provides less precise estimates than SRS of same size.

## Methods

We conducted a Monte Carlo study using $k$=10,000 iterations as follows:

1. Define the population size as $N$=10,000
2. Generate strata membership as $X_1$ and cluster membership as $X_2$, for each $j \in N$ as

   $X_1 \sim discrete\ uniform(1,4)$,

   $X_2 \sim discrete\ uniform(1,100)$
3. Use these covariates in a linear regression framework with coefficients $\alpha$ (baseline population mean), $\beta_1$ (strata effect), $\beta_2$ (cluster effect), and $\epsilon$ (error) to generate population values $Z_i$ $(i = 1, \dots, N)$ and calculate population mean

   $Z_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon_i$, for $i = 1, \dots, N$ and $\epsilon \sim N(0,1)$

   $\bar{Z} = \sum_{i=1}^{N} Z_i / N$
4. Take a sample, $n = 1{,}000$, $k$ times according to the previously discussed sampling strategies. Calculate the sample mean from each sampling method:

   **SRS:** $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$; $\widehat{var(\bar{y})} = \frac{N-n}{Nn} s^2$

   where $s^2 = \frac{1}{n-1}\sum (y_i - \bar{y})^2$

   **StRS:** $\bar{y} = \frac{1}{N}\sum_{j=1}^{h} N_j \bar{y}_j$; $\widehat{var(\bar{y})} = \sum_{j=1}^{h} w_j^2 \left(\frac{N_j - n_j}{N_j n_j}\right) s_j^2$

   where $s_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$, $w_j = \frac{N_j}{N}$, $y_{ij}$ =value of $i^{th}$ unit in $j^{th}$ stratum, $j$=1,2,…$h$

   **CS:** $\bar{y} = \frac{1}{n}\sum_{j=1}^{M} \bar{y}_j$; $\widehat{var(\bar{y})} = \frac{N-n}{Nn} s_b^2$

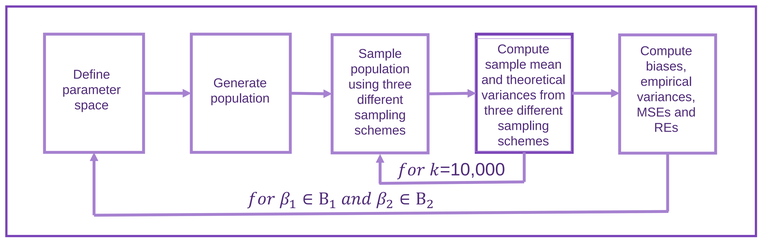
   where $s_b^2 = \frac{1}{n-1}\sum_{j=1}^{n} (\bar{y}_j - \bar{y})^2$, $\bar{y}_j = \frac{1}{M}\sum_{i=1}^{M} y_{ij}$, $y_{ij}$ =value of $i^{th}$ unit in $j^{th}$ cluster, $j$=1,2,…$M$
5. Compute the bias, theoretical and empirical variances, MSEs and relative efficiencies from each sampling method:

   $Bias$: $\sum \frac{|\bar{y} - \bar{Z}|}{K}$, $Tvar$: $\sum \frac{var(\bar{y})}{K}$, $Evar$: $\sum \frac{(\bar{y} - \bar{Z})^2}{K}$, $MSE$: $Bias^2 + Evar$, $RE$= $\frac{MSE_i}{MSE_{i \neq j}}$

## Methods (Cont.)

6. Repeat this process for various values of $\beta_1$ and $\beta_2$, where $\beta_1$ represents the relationship with strata and $\beta_2$ represents the relationship with the clusters.

### Simulation Structure




## Results

### Simulation 1: $\alpha = 0; \beta_1 \in [0,3.5]; \beta_2 = 0$

$\alpha$=0, $\beta_2$=0, No of strata=$h$=4, No of clusters=$M$=100, No of clusters selected=10

| | $\beta_1$: | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 | 3 | 3.5 |
|---|---|---|---|---|---|---|---|---|---|
| SRS | Bias | 0.024 | 0.027 | 0.036 | 0.047 | 0.0587 | 0.071 | 0.084 | 0.096 |
| | Theo Var | 0.001 | 0.001 | 0.002 | 0.003 | 0.0054 | 0.008 | 0.011 | 0.015 |
| | Emp Var | 0.001 | 0.001 | 0.002 | 0.004 | 0.0054 | 0.008 | 0.011 | 0.015 |
| StRS | Bias | 0.024 | 0.024 | 0.024 | 0.024 | 0.0238 | 0.024 | 0.024 | 0.024 |
| | Theo Var | 0.001 | 0.001 | 0.001 | 0.001 | 0.0009 | 0.001 | 0.001 | 0.001 |
| | Emp Var | 0.001 | 0.001 | 0.001 | 0.001 | 0.0009 | 0.001 | 0.001 | 0.001 |
| CS | Bias | 0.025 | 0.028 | 0.032 | 0.048 | 0.0597 | 0.070 | 0.079 | 0.094 |
| | Theo Var | 0.001 | 0.001 | 0.002 | 0.004 | 0.0056 | 0.008 | 0.010 | 0.014 |
| | Emp Var | 0.001 | 0.001 | 0.002 | 0.004 | 0.0055 | 0.008 | 0.010 | 0.014 |
| *RE* | StRS vs SRS | 0.992 | 0.766 | 0.441 | 0.259 | 0.1646 | 0.117 | 0.084 | 0.061 |
| | CS vs SRS | 1.047 | 1.016 | 0.804 | 1.034 | 1.0265 | 0.966 | 0.892 | 0.957 |
| | CS vs StRS | 1.056 | 1.327 | 1.822 | 3.990 | 6.2349 | 8.285 | 10.593 | 15.823 |

### Simulation 2: $\alpha = 0; \beta_1 = 0.5; \beta_2 \in [0,3.5]$

$\alpha$=0, $\beta_1$=0.5, No of strata=$h$=4, No of clusters=$M$=100, No of clusters selected=10

| | $\beta_2$: | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 | 3 | 3.5 |
|---|---|---|---|---|---|---|---|---|---|
| SRS | Bias | 0.027 | 0.349 | 0.698 | 1.034 | 1.371 | 1.732 | 2.088 | 2.434 |
| | Theo Var | 0.001 | 0.189 | 0.751 | 1.688 | 3.003 | 4.690 | 6.752 | 9.187 |
| | Emp Var | 0.001 | 0.189 | 0.771 | 1.674 | 2.970 | 4.709 | 6.856 | 9.258 |
| StRS | Bias | 0.024 | 0.349 | 0.694 | 1.055 | 1.391 | 1.726 | 2.093 | 2.426 |
| | Theo Var | 0.001 | 0.189 | 0.751 | 1.687 | 3.003 | 4.687 | 6.753 | 9.183 |
| | Emp Var | 0.001 | 0.192 | 0.759 | 1.744 | 3.035 | 4.680 | 6.844 | 9.237 |
| CS | Bias | 0.023 | 3.459 | 7.003 | 10.286 | 13.791 | 17.410 | 21.250 | 24.559 |
| | Theo Var | 0.001 | 18.977 | 75.383 | 170.602 | 302.857 | 473.080 | 679.903 | 927.811 |
| | Emp Var | 0.001 | 18.635 | 76.029 | 165.264 | 295.483 | 467.648 | 699.916 | 939.905 |
| *RE* | StRS vs SRS | 0.759 | 1.011 | 0.986 | 1.042 | 1.025 | 0.994 | 1.001 | 0.996 |
| | CS vs SRS | 0.717 | 98.572 | 99.409 | 98.843 | 100.171 | 99.972 | 102.672 | 101.636 |
| | CS vs StRS | 0.945 | 97.087 | 101.010 | 95.238 | 98.039 | 101.010 | 103.093 | 102.041 |

## Results (cont.)

In all simulations, we found that the empirical variances and the theoretical variances were approximately equal. This was to be expected and verified that our simulation structure and estimators were properly constructed and calculated, respectively.

In simulation 1, we held $\beta_2$ constant, which can be considered to be the correlation between the study variable and the cluster membership, and increased $\beta_1$ in 0.5 increments, which can be considered to be the correlation between the study variable and the strata membership. From simulation 1 results, we observed that the stratified random sampling out performed both cluster sampling and SRS. The biases and the variances from the StRS were minimum among the three sampling methods once $\beta_1$ >0. This was expected since for $\beta_1$>0 there is a relationship between the study variable and strata. When we compared CS and SRS, we saw that CS and SRS were very close in their biases and variances, and hence the relative efficiency values were close to 1, which was due to the fact that $\beta_2$ was held at 0.

In simulation 2, we held $\beta_1$ constant at 0.5 and increased $\beta_2$ values by 0.5. The remaining parameters were the same values as in simulation 1. From simulation 2 results, we saw that as we increase the correlation between the study variable and the cluster membership, both SRS and StRS provided better estimates than CS. Cluster sampling performed very poorly, especially for high $\beta_2$ values. One explanation might be the fact that since only 10 clusters were randomly selected out of 100, the samples from CS were not representative of the population even for high $\beta_2$ values. When $\beta_2$>0 , SRS and StRS performed very similarly; i.e. the REs ≅1, which is due to holding $\beta_1$ constant at 0.5, where 0.5 represents low correlation with strata membership. We also observed that CS estimates were biased.

## Discussion

The results in simulation 1 were not surprising. This is due to the fact that the true data generation method involved an increasing correlation between strata membership and population mean. We expected to see that StRS outperform the other methods and we did. What was surprising was how well CS did compared to SRS. The explanation could be that each cluster was well mixed when it came to strata membership, or, the subjects were heterogeneous within clusters, thus improving its efficiency in estimating the true population mean.

The results in simulation 2 were more surprising, however. Although it is known that generally estimates from CS are prone to bias and have high variances, we saw that as the correlation between study variable and the cluster membership increased, population mean estimate from CS actually did "worse". We believe this might be explained by the mechanics of CS. If one does not sample the larger cluster membership ($X_2$ is large, $\beta_2$ is large) then one will not get a representative sample of the population and have a large bias for the estimator of the mean as well as a large variance. Further, we observed that StRS and SRS compete for the first place for the best method. The reason for StRS is not better than SRS is that the correlation between strata and study variable is low ($\beta_1 = .5$). Thus the efficiency usually gained by stratification was negligible.

## Conclusion

We conclude that since StRS clearly performs better than both SRS and CS in terms of bias and RE, when researchers have a sampling frame, they should utilize StRS method with appropriately selected strata (e.g., gender, race, age etc.). CS should be avoided unless there is no reliable/available sampling frame. While utilizing CS, researches must be cautious about the fact that, if the subjects are not heterogeneous within the clusters, CS can provide biased estimates of the population mean with large variances.

## References

- Stat Trek,How to Estimate a Mean or Proportion from a Simple Random Sample https://stattrek.com/survey-research/simple-random-sample-analysis.aspx?tutorial=samp
- Leslie Kish, Survey Sampling, John Wiley & Sons