A Simulation Study to Compare the Bias and Efficiency of Sample Mean Obtained from Probability and Nonprobability Samples

Harun Mazumder[1], Sarah Grunblatt[2], Chaoyi Zeng[1], E. Oral[1]*

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

Louisiana State University Health Sciences Center School of Public Health

**Contact: eoral@lsuhsc.edu*

Motivation and Objectives

Many different sampling methods exist, and each has its own merits. Typically, nonprobability sampling is cheaper and easier to implement. In this study, we compare what effect nonprobability sampling (particularly convenience sampling) will have on the sample mean, especially in regards to bias and efficiency, as opposed to probability sampling. This study aims to better understand if there is truly a significant enough difference in the results to justify implementing one method over another. By designing a Monte-Carlo simulation, this study compares the bias and efficiency of the sample mean obtained from convenience sampling (CS) and simple random sampling (SRS).

Background

Nonprobability sampling methods do not involve random selection of the population members. Therefore, not all members of the population have a chance of being in the sample. As such, it is difficult to know if the sample reflects the distribution of the larger population, creating the issue of non-generalizability of research findings. Thus, probability sampling is superior to nonprobability sampling. However, in some circumstances, probability sampling may not be feasible, practical, or theoretically sensible. In such cases, nonprobability sampling methods are implemented. They are cost- and time-effective, easy to use, and can be utilized with a small population.

CS, which is also known as Haphazard or Accidental sampling, involves collecting a sample via the easiest possible manner, which often occurs at one location and time with whomever is available. Some potential respondents may accidentally be missed or unconsciously avoided. There is a high chance that a convenient sample could under- or over-represent the target population.

Methods

We conducted a Monte Carlo simulation with 10,000 iterations to compare the bias and relative efficiency (RE) of the sample mean obtained from SRS vs CS. For the CS, we defined

$$s_j = \begin{cases} 1, \text{if subject } j \text{ is selected} \\ 0, \text{if subject } j \text{ is not selected} \end{cases}$$

so that the selection probability becomes $P(S_j = 1) = p_j$ for j=1,2,…,N. For population size, we considered N=5,000. We assumed that the relationship between the outcome Z and the covariate X is given by the equation

$$Z_j = \theta + \beta X_j + \epsilon_j \qquad (1)$$

where the covariate and the error term were generated from $X \sim N(0,1)$ and $\epsilon \sim N(0,1)$, and θ was set at 10. We generated the population from (1). Then for each iteration we selected a simple random sample and a convenience sample. To select the convenience sample, we assumed that the relationship between p_j and the covariate X is given by

$$\ln\left(\frac{p_j}{1-p_j}\right) = \alpha + \gamma X_j$$

and considered various values for the parameters $\alpha, \beta, and\ \gamma$. We calculated the selection probability by using the formula

$$p_j = \frac{e^{\alpha+\gamma X_j}}{1+e^{\alpha+\gamma X_j}}$$

and generated s_j ~ Bernouilli(p_j). If $s_j = 1$ then the subject Z_j was selected to the convenience sample, otherwise the subject Z_j was not selected (for j=1,2,…,N). We calculated the sample means for SRS ($\bar{y}_{Sk}$) and convenience sample ($\bar{y}_{Ck}$) in each iteration. The empirical biases and variances, as well as the empirical mean square errors (MSEs) and relative efficiencies were calculated from the following formulas

$$Bias(SRS) = \frac{\sum_{k=1}^{10,000} |\bar{y}_{Sk} - \bar{Z}|}{10,000} \qquad Bias(Conv) = \frac{\sum_{k=1}^{10,000} |\bar{y}_{Ck} - \bar{Z}|}{10,000}$$

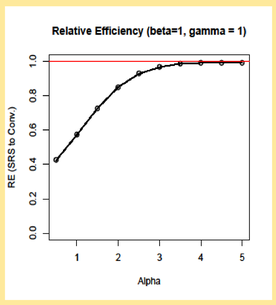
$$EVar(SRS) = \frac{\sum_{k=1}^{10,000} (\bar{y}_{Sk} - \bar{Z})^2}{10,000} \qquad EVar(Conv) = \frac{\sum_{k=1}^{10,000} (\bar{y}_{Ck} - \bar{Z})^2}{10,000}$$

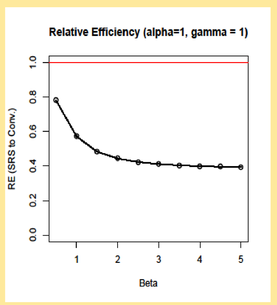
$$MSE(SRS) = Bias(SRS)^2 + EVar(SRS) \quad MSE(SRS) = Bias(SRS)^2 + EVar(SRS)$$

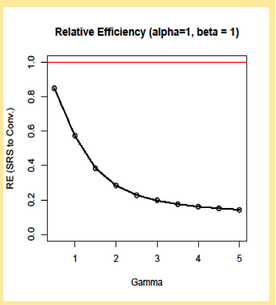
$$RE = \frac{MSE(SRS)}{MSE(Conv)}$$

Results

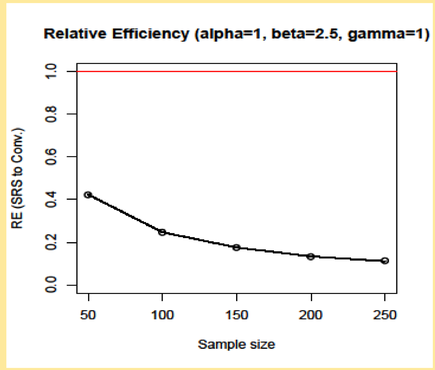
We considered several scenarios to examine the effects of parameters on RE. We first fixed the sample size at 50, and changed other parameters to reflect the change in the degree of relationship between the predictor and outcome. From the plots below, we could see that as α







increases, given γ and β are fixed, the RE increases, implying the gradual improvement in the performance of CS. However, SRS is still better since RE < 1 for all values of alpha. When we increase β, while keeping α and γ fixed, we see the gradual improvement in the performance of SRS over CS. This is expected since when β increases, the relationship between the outcome and the covariate increases. Similarly, when we increase γ while keeping α and β fixed, the RE values get smaller, indicating that SRS is performing better (or, equivalently indicating that CS is doing worse). Next, we fixed all the parameters ($\alpha, \beta, and\ \gamma$) and increased the sample size to see the effect of sample size.



From the plot given on the left we see that as the sample size increases, the RE values still decrease, i.e., even if we enlarge the sample size, the results from the CS are still not accurate. In other words, CS still does not provide an accurate reflection of the true population; SRS still provides smaller MSE as the sample size increases. Note that, although not presented here for conciseness, we observed similar results when we compared biases from SRS and CS: as α increases, there is a gradual improvement in the bias from CS. However, when β or γ are increased, the biases from SRS are much smaller than the biases from CS.

Discussion

While many flaws of CS are already known, this method is still frequently implemented due to its low cost and ease of implementation. Our study confirms that the bias and the MSE of the sample mean from CS is higher than that of SRS. As a result, CS should perhaps be limited to pilot studies where very little on the study topic is known. The results could then serve as preliminary data, and help to determine better sampling designs for future studies. As we showed here, CS is not a reliable sampling method to be depended upon for large-scale research studies where statistically significant results and major decisions will be made. Due to its un-probabilistic nature of the data collection process, CS results in biased estimates with large variances.

References

1. Boonstra, Philip S.; Little, Roderick JA; West, Brady T.; Andridge, Rebecca R.; and Alvarado-Leiton, Fernanda, "A simulation study of diagnostics for bias in non-probability samples" (March 2019). *The University of Michigan Department of Biostatistics Working Paper Series*. Working Paper 125.
2. Wiśniowski, A., Sakshaug, J. W., Perez Ruiz, D. A., & Blom, A. G. (2020). Integrating probability and nonprobability samples for survey inference. Journal of Survey Statistics and Methodology, 8(1), 120-147.
3. West, B. T., Little, R. J., Andridge, R. R., Boonstra, P. S., Ware, E. B., Pandit, A., & Alvarado-Leiton, F. (2020). Measures of Selection Bias in Regression Coefficients Estimated from Non-Probability Samples. arXiv preprint arXiv:2004.06139.