A Practical Balancing for a Random Sample from a Finite Population by Systematic Selection

Hee-Choon Shin[1] and Jibum Kim[2]

[1] National Center for Health Statistics*, 3311 Toledo Road, Hyattsville, MD 20782
[2] Sungkyunkwan University, Faculty Hall, #513, 53 Myeongnyun-dong 3-ga, Jongno-gu, Seoul, 110-745, Korea

Abstract

The main objective of sampling is to obtain a representative sample for an unbiased and efficient estimate within a budget constraint. In a balanced sample, according to Yates' definition (Yates, 1971), the mean value of the balanced factor in the sample is equal to the mean of the factor in the population. In this study, a balanced sample is not a purposively selected sample but a randomly selected one. Another important reason for a balanced sample is to protect the inference against a model misspecification (Royall & Herson, 1973a; Royall & Herson, 1973b). In this work, we propose and demonstrate a practical balancing method which would be a small modification to currently practiced design-based sampling procedures for small and large-scale surveys. We demonstrate practicality of our approach with a simulation of sample selection from 3,143 U.S. Counties for estimates of the total and mean population sizes in 2010 with Census 2000 count and State indicator as auxiliary variables. Our simulation study indicated that a balanced sample was good for reducing bias regardless of the particular sorting method. Rather than selecting a random sample from an ordered frame, we should try to find a balanced sample for an unbiased estimate.

Key Words: Balanced Sample, Bias, Variance, Mean Squared Error.

1. Introduction

The main objective of sampling is to obtain a representative sample for an unbiased and efficient estimate within a budget constraint. In a balanced sample, according to Yates' definition (Yates, 1971), the mean value of the balanced factor in the sample is equal to the mean of the factor in the population. In this study, a balanced sample is not a purposively selected sample but a randomly selected one. Another important reason for a balanced sample is to protect the inference against a model misspecification (Royall & Herson, 1973a; Royall & Herson, 1973b). Several approaches including a systematic selection method for a balanced sample have been proposed and practiced (Deville & Tille, 2004; Valliant, Dorfman, & Royall, 2000).

* The findings and conclusions stated in the manuscripts are solely those of the authors. They do not necessarily reflect the views of the National Center for Health Statistics or the Centers for Disease Control and Prevention.

2. Research Objective

In this work, we propose and demonstrate a practical balancing method which would be a small modification to currently practiced design-based sampling procedures for small and large-scale surveys. Consider the problem of selecting size n sample from a population of N elements. There would be 49,950 ways to select a size 2 sample from a population of 1,000 elements. The simple random sampling procedure would give an equal chance to each of the potential 49,950 samples. With a systematic selection procedure (Cochran, 1977; Kendall, Stuart, & Ord, 1983; Sarndal, Swensson, & Wretman, 1992), there would be 500 possible samples of size 2 from a population of 1,000 ordered elements. There would be 6.3851×10^{139} potential simple random samples of size 100 from a population of 1000 elements! Meanwhile, there would be only 10 possible systematic samples of size 100 from a population of 1,000 ordered elements. It would be a daunting task to evaluate all the sample properties for 6.3851×10^{139} potential simple random samples even with a super-fast modern computing machine. Evaluating sample properties of the 10 possible systematic samples and determining a best sample would be a relatively easier task.

The current general approach for selecting a sample from a finite population is to randomly select a single set of sample units from the possible samples by systematic selection with the help of auxiliary variables, and release the selected sample for field work without evaluating the selected sample for balancing. All the estimators and their accuracy depend on representativeness of the selected and released sample. What if the selected sample is an unrepresentative and skewed sample? We argue that we could choose a "balanced" random sample by utilizing available auxiliary variables

3. Simulation

We demonstrated the practicality of our approach with a simulation of sample selection from 3,143 U.S. Counties for an estimate of the total population in 2010, with Census 2000 count and State indicator as auxiliary variables.

Let a be the integer sampling interval and m be the integer part of N/a, where N is the number of U.S. Counties (3,143). Then,

$$N = ma + c,$$

where the integer c is $0 \leq c < a$. The sample size (n) is either m or $m + 1$, depending on the random start. If $c = 0$, then n would be m. Details of the systematic sampling method can be found in standard sampling textbooks (Cochran, 1977; Sarndal, Swensson, & Wretman, 1992). For our simulation, each sample consists of 49 or 50 Counties, and there are 63 unique sets of sample Counties from the *ordered* frame of U.S. Counties.

Sorting. Before implementing systematic selection, the sampling frame was ordered. The following five sorting methods were applied:

a. *Random*: A random number was generated for each County from a uniform distribution between 0 and 1, and the whole frame was sorted by the random numbers.

b. *State and Random*: The frame was sorted by State and the random numbers generated as in a.
c. *Ascending*: The frame was sorted by Census 2000 County population counts in ascending order.
d. *State and Ascending*: The frame was sorted by State and Census 2000 County population counts in ascending order.
e. *State and Serpentine*: The frame was sorted by State and Census 2000 County population counts in a serpentine order. Alternating the ascending and descending order was sequentially applied to the list of States. Specifically, the Counties in the first State were sorted in ascending order and the Counties in the second State were sorted in descending order, and so on.

4. Results

As discussed, there were 63 sets of samples of size 49 or 50 from an *ordered* set of 3,143 Counties. The magnitude of balancing was measured by

$$|\mu - \bar{x}|,$$

where μ is the average of 3,143 County Census 2000 population counts and $\bar{x}$ is the average of sampled Counties' Census 2000 population counts. A smaller value would indicate that the sample is more balanced.

The objective of sampling was to estimate the total U.S. country population size and corresponding mean. The Horvits-Thompson estimator (Horvitz & Thompson, 1952) of the total is:

$$\hat{t}_{HT} = \frac{N}{n}\sum_{i=1}^{n} y_i,$$

where y_i is the 2010 population count of the i^{th} sampled County. The variance estimator was estimated by Cochran's method (Cochran, 1946):

$$Var(\hat{t}_{HT}) = \frac{N^2}{n(n-1)}\left(1-\frac{n}{N}\right)\sum_{i=1}^{n}\left(y_i - \frac{\hat{t}_{HT}}{N}\right)^2.$$

Corresponding Horvits-Thompson estimators of the mean and its variance are

$$\hat{\mu}_{HT} = \frac{\hat{t}_{HT}}{N}, \text{and}$$

$$Var(\hat{\mu}_{HT}) = \frac{1}{N^2}Var(\hat{t}_{HT}).$$

Bias. First, we looked at the relationship between sample balancing and the bias of an estimate. Each graph in Figure 1 shows the relationship between sample balancing and bias of estimates. Bias is defined as the difference between the actual population mean and the

estimate ($\hat{\mu}_{HT}$) from a sample. With regards of the specific sorting method, all 5 graphs indicate that better balancing is associated with a smaller bias.

Estimated Variance. Figure 2 shows the relationship between balancing and the estimated variance of the U.S. 2010 population mean. In general, a better balanced sample is not necessarily related to a lower estimated variance with the exception of sample selection from the frame with ascending sorting order. The best balanced sample generated slightly larger variance for samples from the frame with ascending sorting order.

Mean Squared Error (MSE). Figure 3 shows the relationship between balancing and the mean squared error. MSE is defined as the sum of variance and biased squared. We looked at the MSE's since the bias is exactly known in this simulation. Basically, relationships between balancing and MSE mirrored those between balancing and variance.

5. Concluding Remarks

Our simulation study indicates that a balanced sample is good for reducing bias regardless of the particular sorting method. Rather than selecting a random sample from an ordered frame, we should try to find a balanced sample for an unbiased estimate. However, balanced samples are not necessarily related to smaller variances and MSE's. We should note that a sample is a balanced one if the sample mean of an auxiliary variable is equivalent to the population mean. There could be many ways to obtain the same sample mean. The estimated variance of a sample with the same mean but similar values would be smaller than the estimated variance of a sample with the same mean but vastly differing values. Therefore, it is not surprising to see a larger variance for balanced sample. The interesting relationship between balancing and variance and its functional form in the samples from the frame with an ascending order of Census 2000 population size should be examined further. Further research will be pursued applying other balancing methods (e.g., over-balancing) and/or utilizing auxiliary variable (Census 2000 count in this simulation) as a measure of size for unequal probability samples (Royall & Herson, 1973a; Royall & Herson, 1973b).

6. Acknowledgements

We thank Alan Dorfman for helpful comments and suggestions.

References

Cochran, W. G. (1946). Relative Accuracy of Systematic and Stratified Random Samples for a Certain Class of Populations. *The Annals of Mathematical Statistics, 17*(2), 164-177.

Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). New York: Wiley.

Deville, J.-C., & Tille, Y. (2004). Efficient Balanced Sampling: The Cube Method. *Biometrika, 91*(4), 893-912.

Horvitz, D. G., & Thompson, D. J. (1952). A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association, 47*(260), 663-685.

Kendall, M., Stuart, A., & Ord, J. K. (1983). *The Advanced Theory of Statistics* (Fourth ed., Vol. 3). London: Griffin.

Royall, R. M., & Herson, J. (1973a). Robust Estimation in Finite Populations I. *Journal of the American Statistical Association, 68*(344), 880-889.

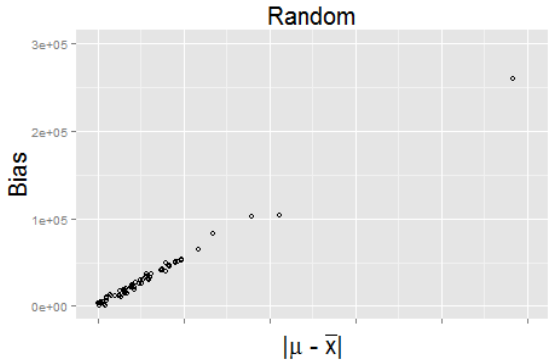
Royall, R. M., & Herson, J. (1973b). Robust Estimation in Finite Populations II: Stratification on a Size Variable. *Journal of the American Statistical Association, 68*(344), 890-893.

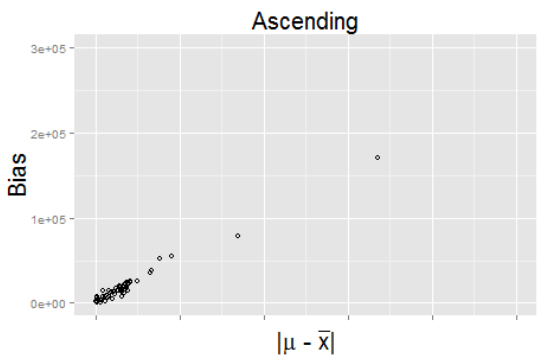
Sarndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling.* New York: Springer.

Valliant, R., Dorfman, A. H., & Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach.* New York: Wiley.

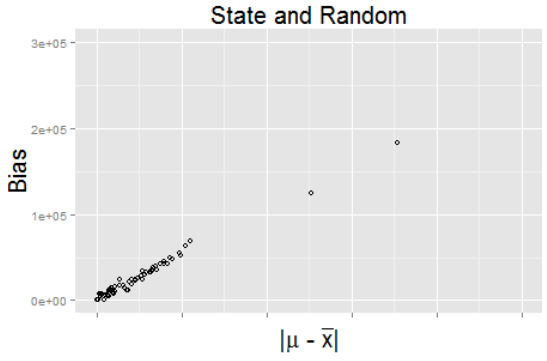
Yates, F. (1971). *Sampling Methods for Censuses and Surveys* (3rd ed.). London: Griffin.

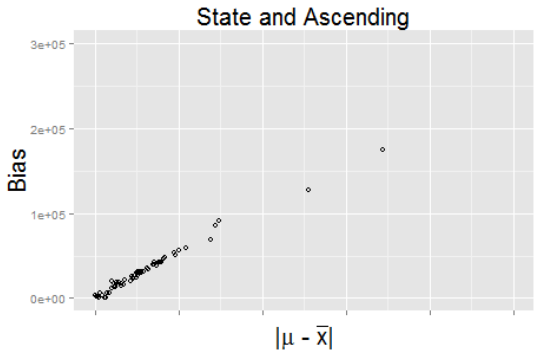
Figure 1. Balancing and Bias





μ :Census 2000 Population Mean (89540.12)
$\hat{x}$:Census 2000 Sample Mean





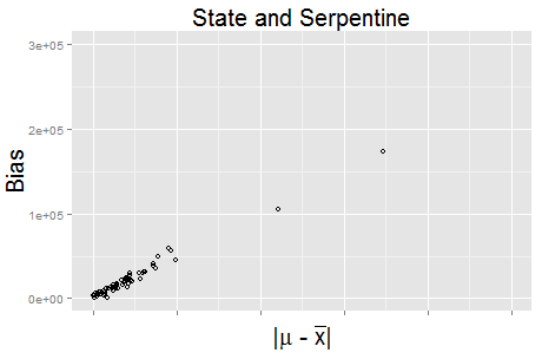
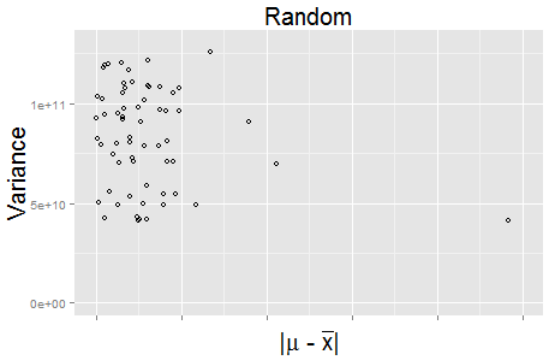
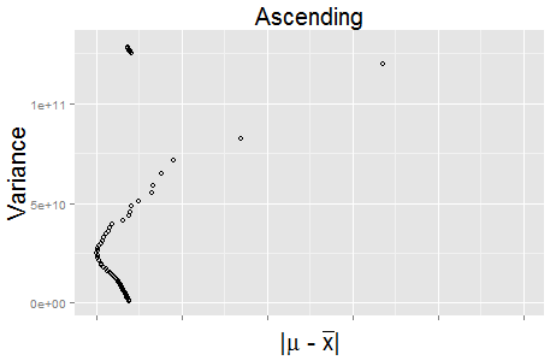
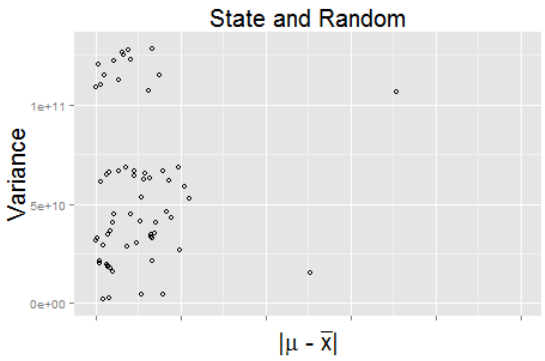


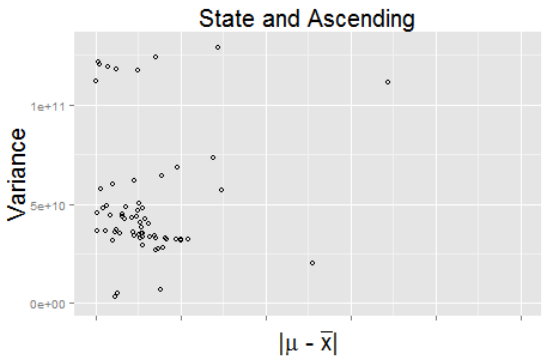
Figure 2. Balancing and Estimated Variance





μ :Census 2000 Population Mean (89540.12)
$\hat{x}$:Census 2000 Sample Mean





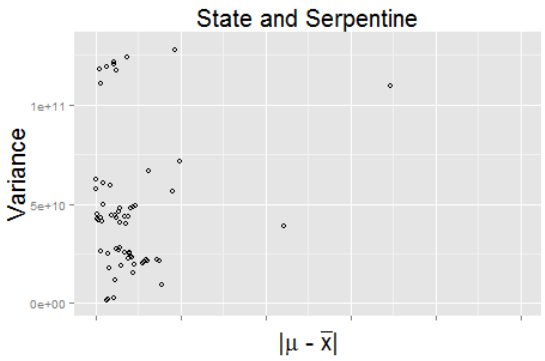


Figure 3. Balancing and Mean Squared Error