British Journal of Mathematics & Computer Science
4(6): 841-848, 2014

SCIENCEDOMAIN *international*
www.sciencedomain.org

Consequences of Violating Randomization in Cluster Sampling

Jason Parrott[1*] and Shlomo Sawilowsky[1]

[1]*Wayne State University, Detroit, Michigan, USA.*

Original Research Article

Received: 17 September 2013
Accepted: 21 November 2013
Published: 16 January 2014

Abstract

Although not as efficient as simple random sampling, cluster sampling has been regarded as a valid sampling technique when the researcher is attempting to save cost. In order to do so, it is necessary that random selection occurs in all stages of sampling. This simulation study examines purposeful selection of cluster sampling in the second stage of a two stage cluster design.

Keywords: *Cluster Sampling; Random Sampling; Monte Carlo Methods*.

1 Introduction

Randomization is the process that gives each element in a population a non-zero chance of being selected. It is an essential step in ensuring that variability is equally distributed between the treatment and control/comparison group (or between a treatment 1 and treatment 2), diminishing the possibility that any characteristic of the population will be overrepresented [1]. It is essential in both the selection and assignment processes, because random selection is a necessary precursor to random assignment in experimental design [2].

A study involving the sampling from a large population, such as an entire state or country, can be more efficiently conducted by defining a sampling frame of groups or clusters, and then selecting participants from those clusters. If a simple random sampling is applied to those clusters, the resulting estimated mean will be unbiased, although the sampling variance will be greater than that obtained via simple random sampling. Hansen and Hurwitz [3] noted the increase in variance due to clustering can be quite substantial, even if the correlation among clusters is small. It can be shown to best offset this increase in sampling variance, it is imperative that the clusters are formed in such as fashion that the subjects "within a cluster vary as much as possible" [4].

It is rarely possible to gerrymander the subjects within the clusters to produce this variation. Hence, in practice, it should be expected that the savings in time and cost by clustering comes at

**Corresponding author: parrottj@royaloakschools.com;*

the loss of efficiency, as Stuart [4] noted, "it usually leads to a substantial loss in precision; and it hardly needs saying that we only use this method when there are compensating advantages of cost" (p. 63). Therefore, Kish [5] suggested that when the lower cost per element of sampling outweighs the increase in variance, and the problems associated with statistical analysis, then cluster sampling is becomes an acceptable choice.

Cornfield [6] suggested that sample size could be inflated to account for the loss in precision. He opined cluster sampling should not be discouraged if it is reasonable to increase the sample size, but this is obviously an inefficient approach to sampling.

Cluster sampling has been slower to develop due to the added design and analysis requirements that it entails. Donner and Klar [7] stated that initial studies of cluster randomization can be traced back to Van Helmont in 1648. In this study, participants were assigned in lots to either the experimental group which received the treatment of bloodletting or to the control group.

Lindquist [8] noted when employing cluster sampling in educational research, for example, that there is the possibility of a large systematic difference from school to school which could account for variability. Glass and Hopkins [9] noted a common flaw in educational research is to select schools or classes at random and then students from those schools or classes. This violates the assumption of interdependence and can't be considered a true random sample. See Simpson et al. [10], and Donner and Klar [7], for a review of applied studies where researchers had to wrestle with this problem. Although currently hierarchical linear modeling is recommended for this type of research layout (e. g., [11]), cluster (and hierarchical cluster) sampling can ameliorate this design complexity [12,13].

Purposeful selection is the process by which predetermined clusters are chosen, whereas a completely random selection of a two stage cluster sample requires random selection at both stages of the selection. The question arises if using nonrandom, preselected clusters (i.e. large counties of a state or province that happen to be geographically contiguous) would be valid if the individual participants were subsequently randomly selected from the clusters. In other words, at Stage 1 of the sampling plan the clusters were nonrandomly selected, but at Stage 2, the subjects within the clusters were randomly selected. Theoretically, if at any stage of a sampling plan the principles of randomization are violated, that sampling frame wills no longer representative of the population. However, it is not known if this merely violates a technical cannon of sampling theory. Therefore, the purpose of this study is to demonstrate how much violation of randomization in selection affects the representation of the population.

2 Methodology

Monte Carlo methods were used to represent responses in a two stage cluster design. The simulation was generated on a WINTEL compatible personal computer using Fortran. A population of 100 clusters of equal size was created, each with 100 individual responses randomly assigned to them. It was generated from a normal (Gaussian) pseudo-random number generator using the IMSL RNNOR subroutine [14].

Cluster means were computed after the scores were assigned. The clusters were rank ordered from highest to lowest to determine purposeful selection later in the study. The initial number of clusters was set to two. Individual scores from each cluster were randomly chosen, representing

the second stage. After their confidence intervals were computed, they were stored and the simulation was repeated for 10,000 repetitions per experiment. At the conclusion, an overall mean was computed for the upper and lower limit of the confidence intervals and recorded.

The simulation was repeated, this time using the two highest clusters that were available according to their mean. This process was repeated 10,000 times and the overall mean of these confidence intervals were computed and stored. The upper limit and lower limit of the confidence interval of the randomly selected group and the purposefully selected group was compared and the difference was computed. The width of the confidence interval for the random selection was also computed and compared to the width of the confidence interval for the purposeful selection using a proportion. This process was completed 19 times, increasing the number of clusters by one (i. e., 2 clusters, 3 clusters, 4 clusters, etc.) until 20 of 100 random and purposeful clusters were chosen and compared.

Usually, in applied studies the researcher must consider the possibility of extraneous or confounding variables. One of the advantages of using a Monte Carlo design is that the study operates in a controlled environment, obviating unknown external influences that can influence the outcome of the study.

3 Results and Discussion

Ceteras paribus, equal cluster sampling already lacks the power available in a simple random sampling. The inefficiency can be quantified by using the following formula from [15] where S^2 *cluster* variance in the cluster sample and S^2 random equals the variance in the random sample:

Formula 1. Rho (ρ)

$$\rho = \frac{s^2 cluster - s^2 random}{(\overline{n} - 1)s^2 random}.$$

This will produce *rho* (ρ) for a cluster sample compared to a simple random sample. The magnitude of ρ can be computed in this manner or referenced from previous studies (e.g., [15]). After ρ is determined the ratio of sampling error between cluster sampling and simple random sampling can be computed as followed:

Formula 2. Ratio of sampling error between cluster sampling and simple random

$$\frac{s^2 cluster}{s^2 random} = 1 + \rho(\overline{n} - 1)$$

Fig. 1 indicates the approximate value of *rho* using the number of participants in each cluster. For example, a cluster containing 10 participants and ρ=.2 would have a 1.18 sampling error compared with the same size simple random sample. Clearly, using a cluster sample compared with a simple random sample affects the integrity of the study, which may only be acceptable if considerations of cost saving is paramount.

Table 1. Estimated ρ Values

$\bar{n}$	ρ =.01	ρ =.02	ρ =.03	ρ =.04	ρ =.05
1	1	1	1	1	1
2	1.01	1.02	1.03	1.04	1.05
3	1.02	1.04	1.06	1.08	1.1
4	1.03	1.06	1.09	1.12	1.15
5	1.04	1.08	1.12	1.16	1.2
6	1.05	1.1	1.15	1.2	1.25
7	1.06	1.12	1.18	1.24	1.3
8	1.07	1.14	1.21	1.28	1.35
9	1.08	1.16	1.24	1.32	1.4
10	1.09	1.18	1.27	1.36	1.45

The confidence interval was analyzed for both random and purposeful clusters at each sample size of clusters. Graphs and tables were developed for the lower limit and the upper limit of each confidence interval. The width of random and purposeful confidence intervals were also compared using a proportion. First, the lower limit of the confidence intervals for both random and purposeful selection will be analyzed.

Fig. 1 shows the graphical comparison between the lower limit of the confidence intervals for random cluster selection versus a lower limit of the confidence intervals for a purposeful cluster selection.

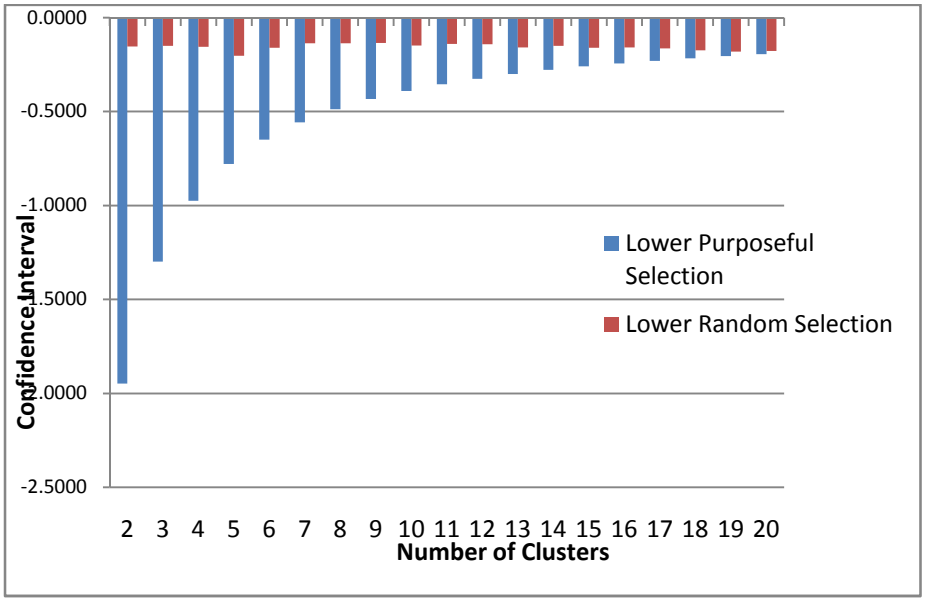

Fig. 1. Random vs. Purposeful Lower Limit Cluster Results

The first observation to note is that the lower limit of the confidence intervals for the random selection of clusters remains consistent independent of the number of clusters chosen. The same

cannot be said for the lower limit of the confidence intervals for the purposeful selection of clusters. The lower limit of the confidence intervals for the purposeful selection of clusters shows variability dependent on the number of clusters being chosen. The largest discrepancy between the purposeful lower limit of the confidence intervals versus the random lower limit of the confidence intervals occurred during the initial selection size of two clusters. In this stage, the difference between the lower limit of the confidence interval for purposeful selection compared to the lower limit of the confidence interval for random selection was 1.8 (-1.948515 to -.15). The next selection size of three clusters showed an improvement in the lower limit of the confidence intervals between purposeful and random clusters to a difference of 1.15 (-1.29901 to -0.1502589). The difference in the lower limit of the confidence intervals between purposeful and random selection of clusters continues to decrease until the last simulation is compiled using a sample size of 20 clusters. At this stage, the difference between the purposeful lower limit of the confidence interval and the random lower limit of the confidence interval was .02 (0.1948515 to -0.1774427).

Similar results were compiled for use of the upper limit of the confidence intervals. Fig. 2 shows a graphical comparison between the upper limit of the confidence intervals for purposeful cluster selection versus random cluster selection.

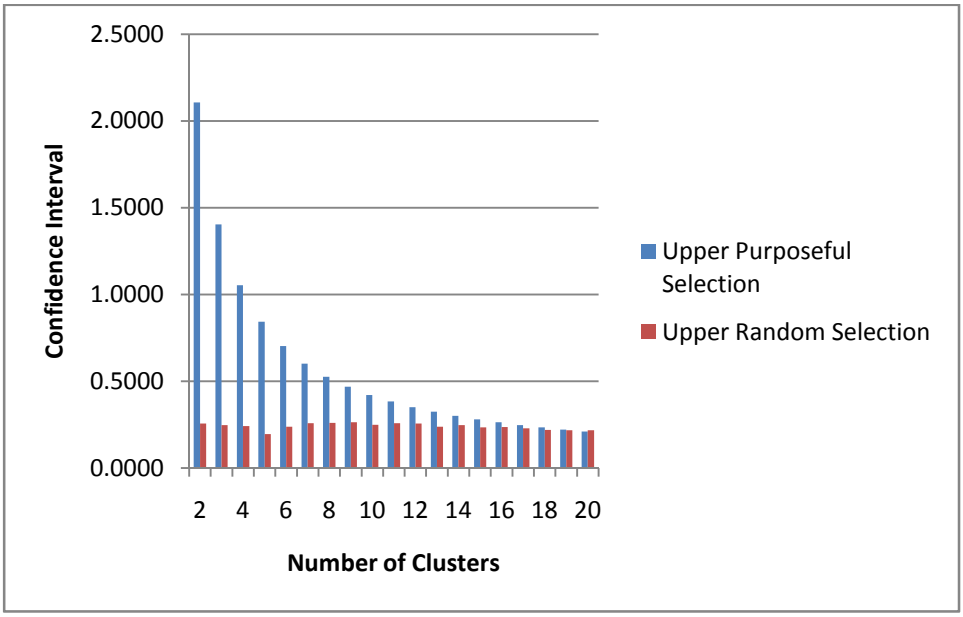

Fig. 2. Random vs. Purposeful Upper Limit Cluster Results

Again, the first observation is that the overall random selection of the upper limit of the confidence interval clusters remained consistent independent of cluster size. The largest discrepancy between the upper limit of the confidence interval for purposeful versus random cluster selection occurred during the initial selection size of two clusters. In this stage, the difference between the upper limit of the confidence interval between purposeful and random selection was 1.85 (2.10664 to 0.256395). The next purposeful selection size of three clusters showed an improvement to a difference of 1.16 (1.404427 to 0.24761) between the upper limit of

the confidence interval of purposeful versus random selection. The difference continues to decrease until the last simulation is compiled using a sample size of 20 clusters. At this stage, the difference between the upper limit of the confidence intervals between purposeful versus random cluster selection was -.01 (0.210664 to 0.218487).

It is evident that there is a difference between purposeful and random selection in both the lower limit and upper limit of the confidence intervals. This difference makes the width of the overall purposeful confidence interval greater than the width of the random confidence interval. The difference in the width of the confidence intervals between the purposeful and random selection is dependent on the number of clusters chosen. The extent of that width was examined in Fig. 3.

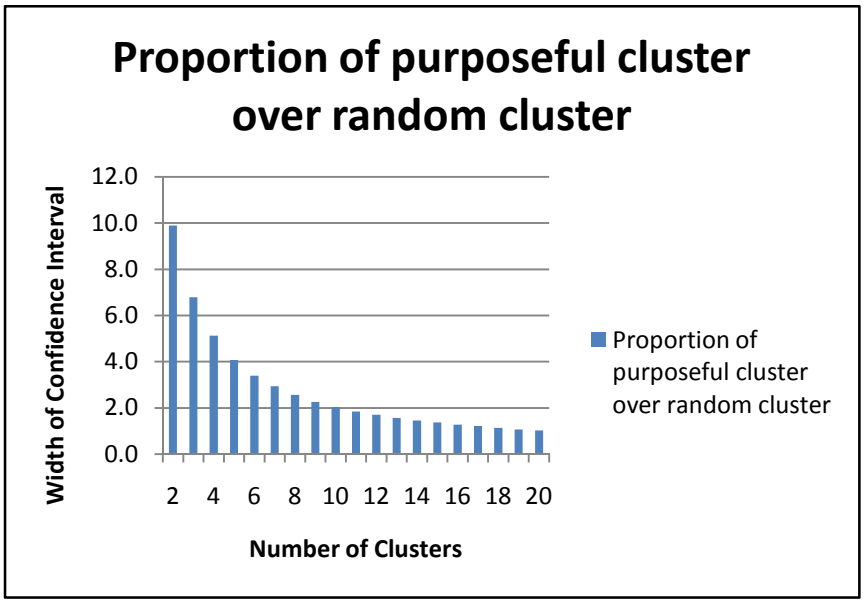

Fig. 3. Proportion of purposeful cluster over random cluster

It is evident that the proportion of the width of the confidence intervals between purposeful clusters selection compared to random clusters is also different. It is similar to the previous two graphs of lower limit and upper limit of the confidence intervals. As the number of clusters increase, the ratio between the purposeful and random samples decreases. For example, the width of the confidence interval for purposeful cluster selection using two clusters was 9.9 times greater than the width of the random cluster selection using two clusters. The width of the confidence interval of a purposeful cluster selection using three clusters is 6.8 times greater than its random cluster selection counterpart. The difference in the overall width of confidence intervals continue to decrease as cluster size increases until its last simulation using a sample size of 20 clusters. Using a sample size of 20 clusters the width of the purposeful selection is 1.02 times greater than the random selection.

4 Conclusion

Cluster sampling can save on time or cost if a population is dispersed in specific areas, although it will lead to a decrease in efficiency of the results. When this is compounded with improper or purposeful selection of clusters, the sampling distribution can become greatly skewed. It was determined that the ratio of the confidence interval for purposeful selection of clusters was almost ten times wider than the confidence interval for random cluster selection using two clusters. On average, there was a width of 0.4 between the upper and lower limit confidence intervals.

Purposeful selection in the first stage of cluster sampling produces a greater width in confidence intervals at each cluster size as compared confidences produced by random selection. Therefore, the researcher should be discouraged from purposefully selecting clusters due cost and time, because, in the words of Cornfield [6], the study will be "an exercise in self deception" (p. 101).

Competing Interests

Authors have declared that no competing interests exist.

REFERENCES

[1] Wiesberg HF, Krosnick JA, Bowen. An introduction to survey research, polling, and data analysis. Thousand Oaks, Sage Publications; 1996.

[2] Runyon RK, Coleman, et al. Fundamentals of Behavioral Statistics, McGraw-Hill; 2000.

[3] Hansen MH, HW. Relative efficiencies of various sampling units in population inquiries. Journal of the American Statistical Association. 1942;37:89-94.

[4] Stuart A. The ideas of sampling. High Wycombe, UK: Charles Griffin; 1984.

[5] Kish L. Survey Sampling. New York, John Wiley and Sons; 1965.

[6] Cornfield J. Randomization by group: A formal analysis. American journal of epidemiology. 1978;108:100-102.

[7] Donner A, Klar N. Design and Analysis of Cluster Randomization Trials in Health Research. London, Arnold; 2000.

[8] Lindquist EF. Statistical analysis in educational research. Boston, Houghton Mifflin Company; 1940.

[9] Glass Gene V, Hopkins, Kenneth D. Statistical Methods in Education Psychology, Third Edition. Boston: Allyn & Bacon; 1996.

[10] Simpson JM, N Klar, et al. Accounting for cluster randomization-A review of primary prevention trials, 1990 through 1993. American Journal of Public Health. 1995;85(10):1378-1383.

[11] Kreft I, de Leeuw J. Introducing multilevel modeling. London, UK: Sage; 1999.

[12] Bryk A, Raudenbush S, Cheong YF. Hierarchical linear and nonlinear modeling. Mahwah, NJ: Erlbaum; 2000.

[13] Hastie T, Tibshirani R, Friedman J. Data mining, inference, and prediction, 2nd ed. Berlin: Springer-Verlag; 2009.

[14] [IMSL IMSL Fortran Numerical Stat Library. Boulder, CO: Visual Numerics, Inc / Rogue Wave Software; 2007.

[15] Sudman S. Applied Sampling. New York, Academic Press; 1976.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
www.sciencedomain.org/review-history.php?iid=370&id=6&aid=3371