

Consequences of Violating Randomization In Cluster Sampling

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 subject 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 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 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 a 1648 study done by Van Helmont.

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 and 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].

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 will 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 were created, each with 100 individual responses randomly assigned to them, generated from a normal (Gaussian) pseudo-random number generator using the IMSL RNNOR subroutine [14].

After the scores were assigned, each cluster had their mean computed. These 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.

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]:

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

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 the ρ can be determined, the ratio of sampling error between cluster sampling and simple random sampling can be computed as followed:

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

Figure 1 indicates the approximate value of rho using the number of participants in each cluster. For example, a cluster containing 10 participants and a ρ value of .2 would have a 1.18 sampling error compared to that of 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

$\bar{n}$	$\rho = .01$	$\rho = .02$	$\rho = .03$	$\rho = .04$	$\rho = .05$
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.

Figure 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.

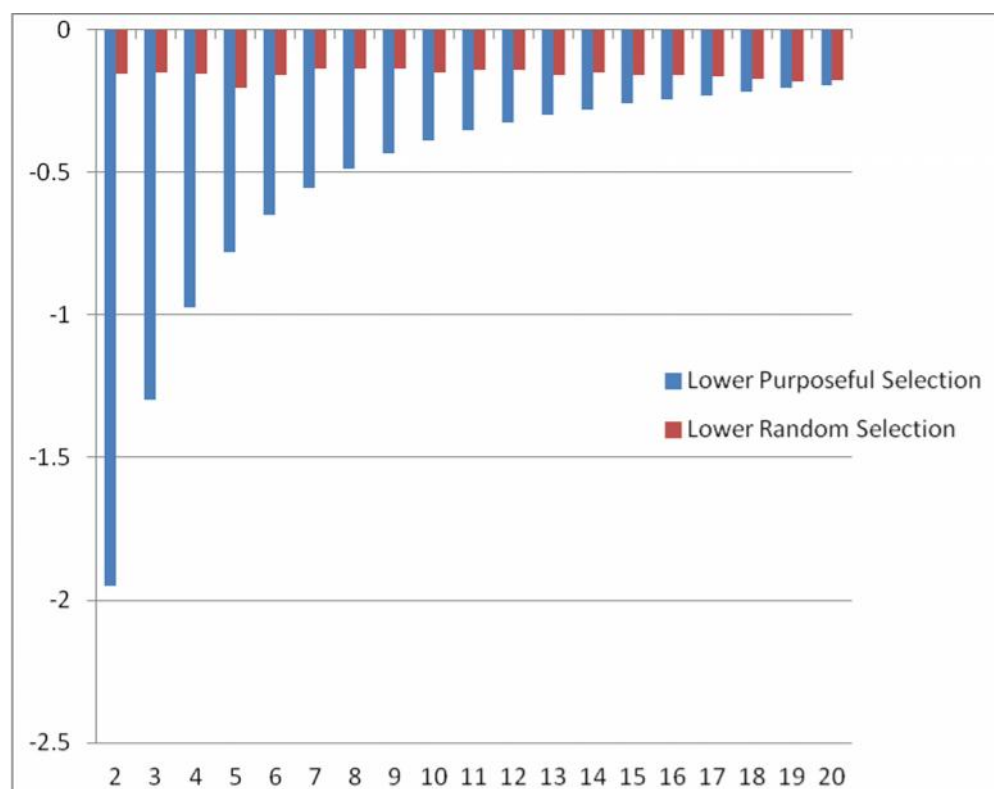


Figure 1: Random vs. Purposeful Lower Limit Cluster Results

Table 2 contains the 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).

136
137
138
139

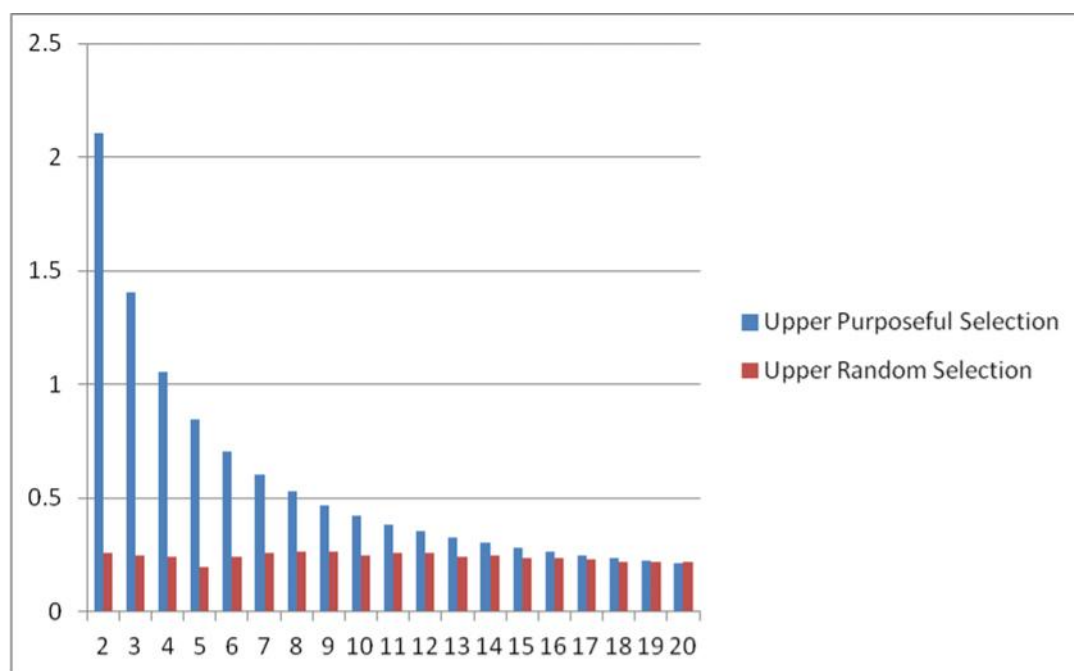
Table 2: Random vs. Purposeful Lower Limit Cluster Results

Cluster Size	Lower Purposeful Selection	Lower Random Selection	Difference
2	-1.9485	-0.1534	1.7951
3	-1.2990	-0.1503	1.1488
4	-0.9743	-0.1549	0.8194
5	-0.7794	-0.2029	0.5765
6	-0.6495	-0.1606	0.4889
7	-0.5567	-0.1368	0.4199
8	-0.4871	-0.1362	0.3509
9	-0.4330	-0.1348	0.2982
10	-0.3897	-0.1487	0.2410
11	-0.3543	-0.1401	0.2142
12	-0.3248	-0.1413	0.1834
13	-0.2998	-0.1581	0.1417
14	-0.2784	-0.1495	0.1289
15	-0.2598	-0.1603	0.0995
16	-0.2436	-0.1587	0.0849
17	-0.2292	-0.1636	0.0657
18	-0.2165	-0.1740	0.0425
19	-0.2051	-0.1805	0.0246
20	-0.1949	-0.1774	0.0174

140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161

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

162
163



164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179

Figure 2: Random vs. Purposeful Upper Limit Cluster Results

The results are compiled in Table 3. 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).

Table 3: Random vs. Purposeful Upper Limit Cluster Results

Cluster Size	Upper Purposeful Selection	Upper Random Selection	Difference
2	2.1066	0.2564	1.8502
3	1.4044	0.2476	1.1568
4	1.0533	0.2409	0.8124
5	0.8427	0.1949	0.6478
6	0.7022	0.2384	0.4638
7	0.6019	0.2584	0.3435
8	0.5267	0.2603	0.2663
9	0.4681	0.2641	0.2041
10	0.4213	0.2487	0.1726
11	0.3830	0.2592	0.1239
12	0.3511	0.2564	0.0947
13	0.3241	0.2388	0.0853
14	0.3009	0.2471	0.0539
15	0.2809	0.2349	0.0460
16	0.2633	0.2370	0.0263

17	0.2478	0.2291	0.0188
18	0.2341	0.2203	0.0138
19	0.2218	0.2172	0.0046
20	0.2107	0.2185	-0.0078

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 Figure 3 and Table 4.

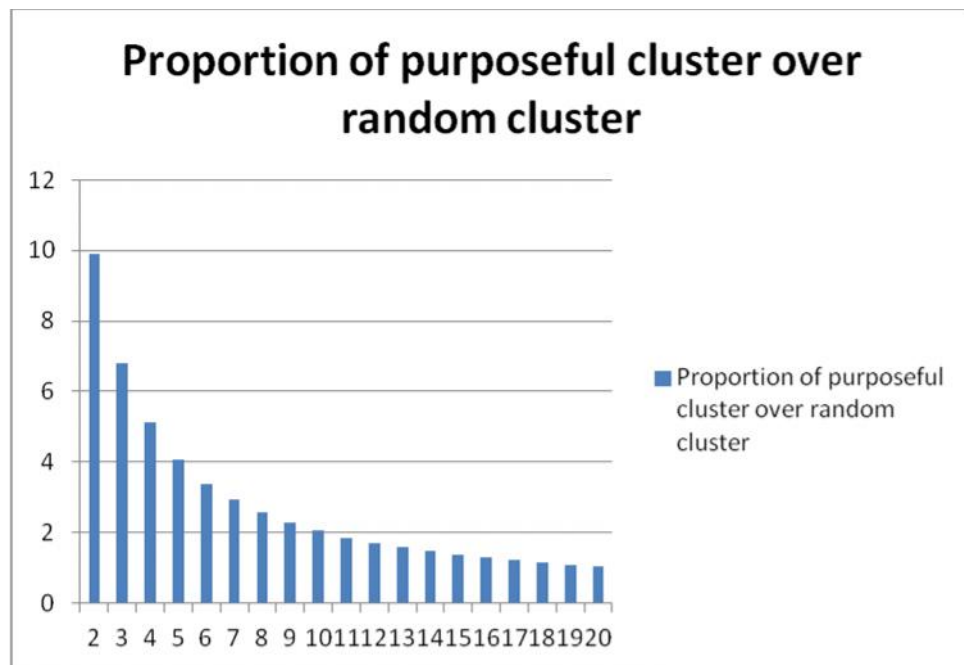


Figure 3: Proportion of purposeful cluster over random cluster

Table 4: Proportion of purposeful cluster over random cluster

Cluster Size	Proportion of purposeful cluster over random cluster
2	9.8952
3	6.7948
4	5.1229
5	4.0782
6	3.3882
7	2.9314
8	2.5566
9	2.2593
10	2.0407
11	1.8468
12	1.6994
13	1.5719
14	1.4608

15	1.3680
16	1.2811
17	1.2150
18	1.1429
19	1.0733
20	1.0242

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

Purposeful selection in the first stage of cluster sampling produces a greater width in confidence intervals at each cluster size compared to random selection. This ratio in the width between purposeful and random selection of confidence intervals decreases as the amount of clusters selected increases showing that overall cluster sample size has an effect on the results of the simulation. Hence, if one invokes nonrandom, or purposeful selection of clusters, the results of the study can become greatly skewed, even under the most optimal conditions. This study determined how much those results would be skewed. 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).

REFERENCES

- [1] Wiesberg, H. F., J. A. Krosnick, & Bowen. (1996). An introduction to survey research, polling, and data analysis. Thousand Oaks, Sage Publications.
- [2] Runyon, R., K. Coleman, et al. (2000). Fundamentals of Behavioral Statistics, Mcgraw-Hill.
- [3] Hansen MH, H. W. (1942). "Relative efficiencies of various sampling units in population inquiries." Journal of the American Statistical Association 37, 89-94.
- [4] Stuart, A. (1984). The ideas of sampling. High Wycombe, UK: Charles Griffin.
- [5] Kish, L. (1965). Survey Sampling. New York, John Wiley and Sons.
- [6] Cornfield, J. (1978). "Randomization by group: A formal analysis." American journal of epidemiology 108, 100-102.
- [7] Donner, A. and N. Klar (2000). Design and Analysis of Cluster Randomization Trials in Health Research. London, Arnold.
- [8] Lindquist, E. F. (1940). Statistical analysis in educational research. Boston, Houghton Mifflin Company.
- [9] Glass, Gene V & Hopkins, Kenneth D. (1996). Statistical Methods in Education Psychology, Third Edition. Boston: Allyn & Bacon.
- [10] Simpson, J. M., N. Klar, et al. (1995). "Accounting for cluster randomization - A review of primary prevention trials, 1990 through 1993." American Journal of Public Health 85(10), 1378-1383.
- [11] Kreft, I., & de Leeuw, J. (1999). Introducing multilevel modeling. London, UK: Sage.

- 247 [12] Bryk, A., Raudenbush, S., & Cheong, Y. F. (2000). Hierarchical linear and nonlinear modeling. Mahwah, NJ:
248 Erlbaum.
- 249 [13] Hastie, T., Tibshirani, R., & Friedman, J. (2009). Data mining, inference, and prediction, 2nd ed. Berlin: Springer-
250 Verlag.
- 251 [14] IMSL (2007). IMSL Fortran Numerical Stat Library. Boulder, CO: Visual Numerics, Inc / Rogue Wave Software.
- 252
- 253 [15] Sudman, S. (1976). Applied Sampling. New York, Academic Press.
- 254
- 255
- 256
- 257