AN ANALYSIS OF THE CLUSTER SAMPLING METHODS IN SOCIAL RESEARCH

DR. SAROJ KUMAR SINGH

Abstract: Cluster sampling is a probability sampling procedure in which elements of the population are randomly selected in naturally occurring aggregates or clusters. The aim of this study is to assess the sampling methods in social research. The role of research in several fields of applied economics, whether related to business or to be economy as a whole has greatly increased in modern times. The researcher must decide the way of selecting a sample from a given population and probability samples are those based on simple random sampling, systematic sampling, stratified sampling, cluster/area sampling whereas non-probability samples are those based on convenience sampling, judgment sampling, and quota sampling techniques. Thus, this paper attempts evaluating information from sampling and assesses its merits and demerits so that appropriate conclusion and suggestions could be made.

Keywords: Cluster Sampling, Probability Sampling, Social Research, Sampling Methods.

Introduction: Cluster sampling is a probability sampling procedure in which elements of the population are randomly selected in naturally occurring groupings (clusters). In the context of cluster sampling, a "cluster" is an aggregate or intact grouping of population elements. Element sampling is the selection of population elements individually, one at a time. On the other hand, cluster sampling involves the selection of population elements not individually, but in aggregates. The sampling units or clusters may be space-based, such as naturally occurring geographical or physical units (e.g., states, counties, census tracts, blocks, or buildings); organization-based, such as such units as school districts, schools, grade levels, or classes; or telephone-based, such as area codes or exchanges of telephone numbers. For the most part, the cluster sample designs described in this chapter are space-based or area-based sampling procedures.

The heterogeneity of the cluster is central to a good cluster sample design. Ideally, the within-cluster differences would be high, and the between-cluster differences would be low. The clusters should be like each other. On the other hand, the elements within each cluster should be as heterogeneous as the target population. Ideally, the clusters would be small but not so small as to be homogeneous.

Objectives of the Study: The objectives of this paper are as follows:

1. To find out different subtypes of cluster probability sampling methods.
2. To evaluate its importance in social research.
3. To assess the cluster sampling methods in social research.
4. To draw appropriate conclusion and suggest measures to reduce errors in sampling.

The main purpose of this paper is to assess the sampling methods in social research. For this purpose, describe sampling as a method of data collection. This paper will focus on sampling as a method to select participants for surveys; more specifically interviewing and self-administered questionnaires. Probability and non-probability sampling as well as the surrounding validity issues will be discussed.

What Are the Steps in Selecting a Cluster Sample? There are six major steps in selecting a cluster sample:

1. Define the target population.
2. Determine the desired sample size.
3. Identify an existing sampling frame or develop a new sampling frame of clusters of the target population.
4. Evaluate the sampling frame for under-coverage, over-coverage, multiple coverage, and clustering, and make adjustments where necessary. Ideally, the clusters would be as heterogeneous as the population, mutually exclusive, and collectively exhaustive. Duplication of elements in the sample may result if population elements belonged to more than one cluster. Omissions will result in coverage bias.
5. Determine the number of clusters to be selected. This may be done by dividing the sample size by estimated average number of population elements in each cluster. To the extent the homogeneity and heterogeneity of the clusters are different from that of the population, as cluster number increases, precision increases. On the other hand, as differences between clusters increases, precision decreases.
6. Randomly select the targeted number of clusters.

Subtypes of Cluster Sampling: Two major dimensions are used to classify different types of cluster sampling. One is based on the number of stages in the sample design, and the other is based on the proportional representation of the clusters in the total sample.

Subtypes Based on Number of Stage: Often cluster sampling is carried out in more than one "stage." A stage is a step in the sampling process in which a sample is taken. Considering the number of stages in

the design, there are three major subtypes of cluster sampling: single-stage cluster sampling, two-stage cluster sampling, and multistage cluster sampling.

Single-stage cluster sampling. In a single-stage cluster sample design, sampling is done only once. As an example of single-stage cluster sampling, let us say one is interested in studying homeless persons who live in shelters. If there are five shelters in a city, a researcher will randomly select one of the shelters and then include in the study all the homeless persons who reside at the selected shelter. A market researcher might choose to use a single-stage cluster sample design. Say a researcher was interested in test marketing a product. The researcher may randomly select zip codes; send samples of the product together with a mail-back evaluation questionnaire to each address within the selected clusters.

Two-stage cluster sampling: A two-stage cluster sample design includes all the steps in single-stage cluster sample design with one exception, the last step. Instead of including all the elements in the selected clusters in the sample, a random sample (either a simple random sample, stratified sample, or systematic sample) is taken from the elements in each selected cluster. Sampling beyond the first stage is sometimes referred to as subsampling. Generally, unless the clusters are homogeneous, a two-stage cluster sample design is better than a one-stage cluster sample design. A self-weighting sample will result if at the first stage sampling is conducted with probability proportional to size (see below). Using the example of the study of homeless persons described above, instead of selecting all the persons who reside at the selected shelter for inclusion in the study, the researcher would randomly select a subset of the residents of the shelter.

Multistage cluster sampling: Surveys of large geographical areas require a somewhat more complicated sample design than those described up to this point. Typically, a multistage cluster sample design must be used. Multistage cluster sampling involves the repetition of two basic steps: listing and sampling. Typically, at each stage, the clusters get progressively smaller in size; and at the last stage element sampling is used. Sampling procedures (simple random sampling, stratified sampling, or systematic sampling) at each stage may differ. It is not necessary that the sampling procedures at each stage be the same. The number of stages that are used is often determined by the availability of sampling frames at different stages.

Special terminology is used to refer to the different sampling units. The sampling unit that is used in the first stage is referred to as the primary sampling unit. The units of subsequent sampling are referred to as the secondary sampling unit, tertiary sampling units, etc., until one gets to the "final" or "ultimate" sampling unit.

Typically, as the sampling process moves from the selection of primary sampling unit to the other sampling stages, the sampling units become more homogeneous. The large clusters tend to be more heterogeneous than small clusters. Because of the greater heterogeneity of the primary sampling unit, sampling error is minimized if one sample has more primary sampling unit than secondary sampling unit, more secondary sampling unit than tertiary sampling units, and so forth.

Subtypes Based on the Proportional Representation of Clusters in Sample: Clusters may be selected in such a way that it is an EPSEM sampling procedure; that is, every element in the population would have an equal chance to be included in the sample. If the clusters sampled are roughly the same size, the sample design may be considered an EPSEM sample design. If the clusters have unequal sizes, an EPSEM sample design may be achieved by using a probability proportionate to size (PPS) selection procedure. The probability of selecting a cluster is dependent on the proportional distribution of its elements in the target population. Using PPS, a self-weighting sample is obtained. Probability disproportional to size sampling involves selecting clusters without considering the proportional distribution of the elements in the target population.

Respondent Selection Procedures: Typically, in household surveys employing a two-stage cluster sample design or a multistage cluster sample design, individual elements are selected at the last stage of the sample design. If the household contains more than one member of the target population, one element must be selected. Both nonprobability and probability procedures are used to select the element from whom to collect data.

Two principal nonprobability household respondent selection procedures are used: head of household selection and first-adult selection. In using the head of household selection the researcher simply asks to speak to the head of household. One may alternatively ask for the male and female heads of household. The first-adult approach involves the selection of the first adult contacted, providing he/she is a member of the target population. These procedures are easy to administer, do not take much time, and are not intrusive. However, they incur selection bias, and are likely to oversample females, as they are more likely than males to be available to be interviewed. The head of household method tends to oversample women, especially in urban areas, due to the greater number of single-parent female-headed households than single-parent male-headed households. The first-adult selection method tends

to oversample women because women are more likely to be at home. These respondent selection procedures do not give every member of the target population a chance to be included in the sample. Combining the probability selection of clusters with the nonprobability selection of household members makes the sampling procedure a mixed-methods procedure. Mixed-methods sampling procedures are described in more detail in the next chapter.

There are several probability household respondent selection procedures. The most frequently used probability approaches are the Kish tables, the Troldah-Carter-Bryant tables, the Hagan and Carter selection method, and the last/next birthday method (Binson, Canchola, & Catania, 2000). These procedures reflect a struggle among researchers to minimize systematic error. Typically, the introduction to the interview is lengthened as they involve two consents: the initial consent from the first contact in the household and second from the person selected to be interviewed. This has the effect of decreasing undercoverage bias but increasing refusal rates. Moreover, if the selected person is not at home, the interviewer is restricted from selecting someone else in the household. Callbacks must be made. The success of the callbacks affects the study's unit nonresponse bias.

Assessment of Cluster Sampling: Cluster sampling has the strengths and weaknesses associated with most probability sampling procedures when compared to nonprobability sampling procedures. However, it has several special strengths and weaknesses when compared to other probability sampling procedures, such as simple random sampling. Some of the strengths of cluster sampling when compared to simple random sampling are:

- If the clusters are geographically defined, cluster sampling requires less time, money, and labor than simple random sampling. It is the most cost-effective probability sampling procedure.
- For the same level of costs, cluster sampling with a higher sample size may yield less sampling error than that resulting from simple random sampling with a smaller sample size.

Cluster sampling permits subsequent sampling because the sampled clusters are aggregates of elements.

- Unlike simple random sampling, cluster sampling permits the estimation characteristics of subsets (clusters) as well as the target population.
- Single-stage cluster sampling requires a sampling frame of the clusters only, and two-stage cluster sampling and multistage cluster sampling require a sampling frame of the elements of the population only for the clusters sampled at the last stage of the process.

Cluster sampling is much easier to implement than simple random sampling.

Some of the weaknesses of cluster sampling when compared to simple random sampling include:

- The sampled clusters may not be as representative of the population as a simple random sample of the same sample size.
- Combining the variance from two separately homogeneous clusters may cause the variance of the entire sample to be higher than that of simple random sampling.
- Cluster sampling introduces more complexity in analyzing data. Inferential statistical analysis of data collected via cluster sampling is more difficult to compute and interpret results than inferential statistical analysis of data collected via simple random sampling.
- The statistical software used to analyze the data collected must use formulas that take into account the use of a cluster sample design. Many statistical software programs utilize formulas for simple random sampling and, as a result, overestimate levels of significance.
- The more stages there are in a cluster sample design, the greater overall sampling error.
- If clusters are not similar to each other, the fewer the number of clusters, the greater the sampling error.
- Cluster sampling yields larger sampling errors for samples of comparable size than other probability samples. If the clusters are similar to each other, this error is minimized. Moreover, these errors can be reduced by increasing the number of clusters. Note, this has the effect of increasing data collection costs.
- The more clusters one selects, the less the difference in data collection costs between cluster sampling and simple random sampling.

Since elements within a cluster tend to be alike, we receive less new information about the population when we select another element from that cluster rather than from another cluster. This lack of new information makes a cluster sample less precise than a simple random sample.

Table: Strengths and Weaknesses of Cluster Sampling Compared to Simple Random Sampling

Difference Between Cluster Sampling and Stratified Sampling

Strengths	**Weaknesses**
Compared to simple random sampling:	**Compared to simple random sampling:**
If the clusters are geographically defined, cluster sampling requires less	A cluster sample may not be as representative of the population as a simple random sample of the

time, money, and labor.	same sample size.
Cluster sampling permits subsequent sampling because the sampled clusters are aggregates of elements.	Variances of cluster samples tend to be much higher than variances of simple random samples.
One can estimate characteristics of the clusters as well as the population.	Cluster sampling introduces more complexity in analyzing data and interpreting results of the analyses.
Cluster sampling does not require a sampling frame of all of the elements in the target population.	Cluster sampling yields larger sampling errors for samples of comparable size than other probability samples.

Cluster sampling is similar to stratified sampling in that both involve separating the population into categories and then sampling within the categories. Both sampling procedures permit analysis of individual categories (strata or clusters) in addition to analysis of the total sample. However, there are important differences. Some of these differences include:

In stratified sampling, once the categories (strata) are created, a random sample is drawn from each category (stratum). On the other hand, in cluster sampling, elements are not selected from each cluster. In single-stage cluster sampling, once the categories (clusters) are created, a random sample of cluster is drawn. All elements in the selected cluster are included in the sample. In two-stage cluster sampling and multi-stage cluster sampling, a random sample of cluster is drawn and then elements are randomly selected from the selected clusters.

Comparison of Stratified Sampling and Cluster Sampling

- In stratified sampling, in order to minimize sampling error, within-group differences among strata should be minimized, and the strata should be as homogeneous as possible. In cluster sampling, in order to minimize sampling error, within-group differences should be consistent with those in the population, and the clusters should be as heterogeneous as the population. The ideal situation for stratified sampling is to have the homogeneity within each stratum and the strata means to differ from each other. The ideal situation for cluster sampling is to have heterogeneity within the clusters and the cluster means not to differ from each other.
- In stratified sampling, in order to minimize sampling error, between-group differences among strata should be maximized. In cluster sampling, in order to minimize sampling error, between-group differences among the clusters should be minimized.
- In stratified sampling, categories are conceptualized by the researcher. In cluster sampling, the categories are naturally occurring groups.
- In stratified sampling, a sampling frame is needed for the entire target population. In single-stage cluster sampling, a sampling frame is needed only for the clusters. In two-stage cluster sampling and multistage cluster sampling, in addition to a sampling frame of the clusters in the first stage of the process, a sampling frame is needed only for elements of each one of the selected clusters.
- The main purpose of stratified sampling is to increase precision and representativeness. The main purpose of cluster sampling is to decrease costs and increase operational efficiency.
- Compared to simple random sampling, stratified sampling has higher precision and cluster sampling has lower precision. The increase in precision by stratification is not that much. However, clustering can cause a significant decrease in precision.
- The variables used for stratification should be related to the variables under study. The variable used for clustering should not be related to the variables under study.
- Commonly used stratification variables are age, gender, and income. Commonly used classification variables in cluster sampling are geographical area, school, and grade level.
- Stratified sampling requires more prior information than cluster sampling; likewise, cluster sampling requires less prior information than stratified sampling.
- In stratified sampling, the researcher strives to divide the target population into a few subgroups, each with many elements in it. In cluster sampling, the researcher strives to divide the target population into many subgroups, each with few elements in it.

Difference between Multistage Sampling and Multiphase Sampling: Multistage sampling (two-stage cluster sampling and multistage cluster sampling) is often confused with multiphase sampling (also referred to as two-phase sampling, double sampling, and post-stratification sampling). Both sampling procedures involve the multiple sampling at different stages or phases, and in some circumstances may be viewed as mixed-methods sampling. In multistage sampling the sampling units for the different stages are different. On the other hand, in multiphase sampling the same sampling unit is sampled multiple times.

Typically, multiphase sampling is used when one does not have a sampling frame with sufficient auxiliary information to allow for stratification. The first phase is used for screening purposes. Using the available sampling frame, one may proceed as follows:

1. Select an initial sample of elements from the available sampling frame.
2. Conduct a short screening interview to collect the necessary auxiliary information for further sampling and stratification.
3. Post stratify the initial sample into strata using the auxiliary information collected.
4. Using the strata for which one desires to collect additional information, select either all the elements in the strata or a probability sample of the elements in the strata for additional data collection.

Multiphase sampling typically is carried out to increase precision, reduce costs, and reduce nonresponse. As noted earlier, stratified samples have higher levels of precision than simple random samples of the same sample size. However, a sampling frame must include information on the stratification variable(s) for all population elements to employ stratification. Multiphase sampling is an option when a sampling frame does not include such information.

Multiphase sampling may also be employed to reduce data collection costs if it took more time and effort to collect data on some variables than to collect data on other variables. In Phase 1, the easily accessible data may be collected from the entire sample. In Phase 2 and other subsequent phases, if desired or necessary, the data that take greater effort or expense to be collected are collected from a smaller subsample. Data collection costs are minimized. Sampling Essentials Multiphase sampling may also be used to obtain information on non -respondents.

Typically, it costs more to collect data on persons who initially refused to participate in a study and other non-respondents than to collect data from the initial respondents. Such costs might be minimized by employing a multiphase sampling of non-respondents.

Below are descriptions of several popular national surveys that are representative of multistage cluster sampling: the National Home and Hospice Care Survey, the National Ambulatory Medical Care Survey, the National Health and Nutrition Examination Survey, the National Survey of Family Growth, and the National Health Interview Survey.

Conclusion: There are four major choices of probability sample designs: simple random sampling, stratified sampling, systematic sampling, and cluster sampling. The strengths and weaknesses of the above sample designs are compared, and guidelines are presented for their selection.

Simple random sampling is a probability sampling procedure that gives every element in the target population and each possible sample of a given size, an equal chance of being selected. As with other probability sampling procedures, it tends to yield representative samples, and allows the use of inferential statistics to compute margin of errors. However, it tends to have larger sampling errors and less precision than stratified samples of the same sample size. If the target population is widely dispersed, data collection costs might be higher for simple random sampling than those for other probability sample designs, such as cluster sampling.

Stratified sampling is a probability sampling procedure in which the target population is first separated into mutually exclusive, homogeneous segments (strata), and then a simple random sample is selected from each segment (stratum). There are two major subtypes of stratified sampling: proportionate stratified sampling and disproportionate stratified sampling. In proportionate stratified sampling, the number of elements allocated to the various strata is proportional to the representation of the strata in the target population. This condition is not satisfied in disproportionate stratified sampling. In this type of stratification, unequal disproportionate allocation, equal disproportionate allocation, or optimum allocation may be applied.

Compared to un-stratified sampling, stratified sampling

1. permits the estimation of population parameters and within-strata inferences and comparisons across strata;
2. tends to be more representative of a population;
3. takes advantage of knowledge the researcher has about the population;
4. possibly makes for lower data collection costs; and
5. permits the researcher to use different sampling procedures within the 172 Sampling Essentials different strata.

On the other hand, unlike un-stratified sampling, stratified sampling requires prior information on the stratification variables and more complex analysis procedures.

Systematic sampling is a probability sampling procedure in which a random selection is made of the first element for the sample, and then subsequent elements are selected using a fixed or systematic interval until the desired sample size is reached. Generally, systematic sampling is easier, simpler, less time-consuming, and more economical than simple random sampling. If the ordering is unrelated to the study variables, but randomized, systematic sampling will yield results similar to simple random sampling.

On the other hand, periodicity in the sampling frame is a constant concern in systematic sampling. Cluster sampling is a probability sampling procedure in which elements of the population are randomly selected in naturally occurring aggregates or clusters. Subtypes of cluster sampling may be classified on the basis of the number of sampling events (single-stage cluster sampling, two-stage cluster sampling, and multistage cluster sampling) and on the basis of the proportional representation of the clusters in the sample (probability proportional to size and probability disproportional to size). Some of the strengths of cluster sampling when compared to simple random sampling include requiring less time, money, and labor; and permitting subsequent sampling and the estimation characteristics of clusters as well as the target population. However, cluster sampling when compared to simple random sampling may not be as representative of the population as a simple random sample of the same sample size, and variances of cluster sampling are likely to be higher than those for simple random sampling.

References:

1. Dr.Sangappa.V.Mamanshetty, Current Issues and Challenges in Agriculture; Business Sciences International Research Journal ISSN 2321 – 3191 Vol 2 Issue 1 (2014), Pg 235-238
2. Binson, D., Canchola, J. A., & Catania, J. A. (2000). Random selection in a national telephone survey: A comparison of the Kish, next birthday, and last-birthday methods. Journal of Official Statistics, 16, 53–59.
3. Burnam, M. A., & Koegel, P. (1988). Methodology for obtaining a representative sample of homeless persons: The Los Angeles Skid Row Study. Evaluation Review, 12, 117–152.
4. Hari Sundar.G.Ram, Bibin Markose, Peer Group influence on Purchase Decision Making; Business Sciences International Research Journal ISSN 2321 – 3191 Vol 2 Issue 1 (2014), Pg 247-260
5. Bryant, B. E. (1975). Respondent selection in a time of changing household composition. Journal of Marketing Research, 12, 129–135.
6. Czaja, R., Blair, J., & Sebestik, J. P. (1982). Respondent selection in a telephone survey: A comparison of three techniques. Journal of Marketing Research, 19, 381–385.
7. Hagan, D. E., & Collier, C. M. (1983). Must respondent selection procedures for telephone surveys be invasive? Public Opinion Quarterly, 47, 547–556.
8. Kish, L. (1949). A procedure for objective respondent selection within the household. Journal of the American Statistical Association, 44, 380–387.
9. Siddharth Shastri, Leveraging foreign Direct investment for Economic; Business Sciences International Research Journal ISSN 2321 – 3191 Vol 2 Issue 1 (2014), Pg 261-268
10. Lavrakas, P. J., Bauman, S. L., & Merkle, D. M. (1993). The last-birthday method and within-unit coverage problems. Proceedings of the section on survey research methods, American Statistical Association, 1107–1112.
11. Salmon, C. T., & Nichols, J. S. (1983). The next-birthday method of respondent selection. Public Opinion Quarterly, 47, 270–276.
12. Troldahl, V. C., & Carter, R. E. (1964). Random selection of respondents within households in phone surveys. Journal of Marketing Surveys, 1, 71–76.
13. Beenaprakash, Dr.Saritavichore, To Study Factors Facilitating Disruptive innovation; Business Sciences International Research Journal ISSN 2321 – 3191 Vol 2 Issue 1 (2014), Pg 243-246

***

Dr. Saroj Kumar Singh/ Dept. of Rural Economics/ S. N. S. R. K. S. College/ Saharsa (Bihar)/
A constituent Unit of B. N. M. University/ Madhepura/ Bihar/