MACHINE INTELLIGENCE RESEARCH

ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

MULTISTAGE SAMPLING STRATEGY FOR IMPROVED EFFICIENCY ESTIMATION IN OBSERVATIONAL STUDIES

Md Abdul Qudus Sheikh
Research Scholar, Department Of Mathematics,
Faculty of School of Teaching department, Dr. A. P. J.Abdul Kalam University Indore, MP,
India, maqudussheikh@gmail.com

Dr. Anjna Rajoria
Research Guide, Department Of Mathematics, Faculty of School of Teaching department, Dr.
A. P. J.Abdul Kalam University Indore, MP, India, anjna_rajoria@rediffmail.com

Abstract

Important data in many fields of study often comes from observational studies. However, possible biases and sampling mistakes might make accurate assessment of demographic characteristics difficult in such investigations. To better estimate efficiency in observational investigations, this academic article introduces a cluster multistage sampling approach. Incorporating several sampling phases, the suggested method allows for the selection of varied and representative samples, which improves the precision and accuracy of parameter estimation.

Introduction

The goal of the data collection method known as multistage sampling, which also goes by the name multistage cluster sampling, is to efficiently collect information from big populations that are spread out throughout a wide area. Each stage of the process entails a further subdivision of the population into smaller and more manageable clusters or groupings.

In the first phase of multistage sampling, the total population is divided into larger clusters, often according to geographical areas or other distinguishable traits. States, cities, neighborhoods, or any other logical divisions might serve as these clusters to make the sampling process easier.

The second stage involves a randomly selected subset of these bigger clusters. This method of random selection minimizes bias and guarantees a representative sample by giving each cluster an equal chance of being included in the study.

The selected clusters are split into smaller units in the third stage. Households or people within a predetermined urban cluster, for instance, can be randomly selected to take part in the survey.

Researchers are able to overcome the difficulties of surveying large and/or varied populations by employing a multistage sampling process. It's a cost-effective way to get information without having to personally contact every member of the population.

When a population is geographically dispersed and hard to reach, multistage sampling is an effective method to get representative data. The survey can focus on various locations by selecting clusters in the first and second stages, guaranteeing a more accurate representation of the total population.

Methodology

There is a cost to collecting data, thus it may be more efficient to merely collect a subset of it. To do this, we employ a k-stage sampling approach, where we start with the complete dataset, represented by Zk = Z. Next, we generate a series of Zk_1, Zk_2,..., Z reduced sampling spaces. Here, Zj is the sample space for all Stage 1 through Stage j information. Starting at Stage 1 and working their way up, people often advance through these stages in a linear fashion.

A discrete random variable J = 1, 2,..., k, where k is the highest stage from which data of a person is taken, is used to determine which stages we should sample each individual from. The likelihood of obtaining data for a specific data point z Z up to Stage j may be calculated with the use of this random variable. This probability represents the likeliness of witnessing data from various phases for a certain person, which helps optimize the sampling process and save expenses.

Researchers and analysts can use the k-stage sampling model to weigh the expense of observation against the requirement for thorough data before deciding which stages to prioritize. Saving money without sacrificing useful information is possible with careful planning of which phases to sample. This method allows for more efficient and cost-effective data analysis and research, which is especially helpful when working with enormous datasets or limited resources.

Methodology

Concept and Method for Multi Stage Cluster Sampling The population is split up into subsets, or "clusters," in a multi-step process in multistage cluster sampling. Usually, there are the following stages:

First, PSUs (primary sampling units) must be chosen. Based on administrative units or geographic areas, the population is grouped into clusters. Primary sampling units (PSUs) for the survey are these clusters.

The Second Step: Choosing Subsequent Samples A sample of secondary sampling units (SSUs) is randomly selected from each cluster of primary sampling units (PSU). Smaller statistical units (SSUs) might be geographical regions or populations of individuals.

Third Step: Choosing the Final Samples Once secondary sampling units (SSUs) have been selected, a random sample of individuals or households is picked from each SSU to form the final sampling units for data collection.

Fundamental to all sampling strategies, the concept of cluster sampling rests on the idea that a population may be broken down into manageable chunks, or "sampling units." The smallest elements, or population units, are these sampling units. By following a predetermined set of criteria, elements in cluster sampling are put together into distinct clusters.

When a complete list of individual elements within the population is unavailable or difficult to gather, cluster sampling becomes useful due to its primary advantage. The emphasis moves from picking individual elements to selecting whole clusters as the primary sampling units.

The cluster sampling process can be outlined as follows:

The entire population is split up into smaller groups called "clusters" based on a set of predetermined characteristics. To best depict the population's variety, the clusters should be internally varied yet outwardly homogeneous.

The Cluster Sampling Method: The clusters themselves serve as the sampling units in cluster sampling. Researchers randomly choose clusters based on specified sampling techniques rather than picking individual elements.

Various sampling strategies, such as simple random sampling, stratified sampling, and systematic sampling, are used to choose a representative sample of clusters from the population. To ensure the reliability of the sample, the selected clusters should be fairly representative of the population at large.

Numbering Chosen Groups: A full enumeration is performed inside each selected cluster after the clusters have been selected. To put it another way, information is gathered from all of the sampling units found inside the selected clusters.

In situations when acquiring a comprehensive list of elements would be resource-intensive or time-consuming, such as when dealing with vast or geographically distributed populations, cluster sampling is very helpful. When researchers use clusters as the primary sampling units, they are able to collect data from a more manageable portion of the population with less time and money spent on the process.

To choose n clusters from N clusters, we use a sampling technique called Simple Random Sampling Without Replacement (SRSWOR). After obtaining these n clusters, we determine the mean for each cluster individually, accounting for all the units contained inside each selected cluster. Let's refer to the averages of these clusters as $y_1, y_2, ..., y_n$.

To estimate the population mean, we take the average of all these cluster means ($y_1, y_2, ..., y_n$). This overall average of the cluster means serves as an estimator of the population mean.

Using this approach, we can obtain a representative estimate of the population mean without having to calculate the mean for all the units individually. By using SRSWOR to select clusters and then finding the mean within each cluster, we can efficiently estimate the population mean based on a smaller subset of the data. This is particularly useful when dealing with large populations, where calculating the mean for all units could be time-consuming and resource-intensive.

It's important to note that the accuracy of this estimator depends on the quality of the sampling method and the representativeness of the selected clusters. A well-designed sampling strategy ensures that the estimator provides a reliable approximation of the population mean.

To drive the equation for the variance of cluster mean, we'll follow similar steps as in deriving the variance of the sample mean in Simple Random Sampling without Replacement (SRSWOR).

In SRSWOR, the sampling units are denoted as 1, 2, ..., N, where N is the population size.

For cluster sampling, the sampling units are grouped into clusters. Each cluster contains multiple sampling units. Let there be M clusters, and each cluster "i" contains ni sampling units. The total number of sampling units in the population is still N, which can be expressed as:

$N = n_1 + n_2 + ... + n_M$

Now, we want to estimate the population mean using cluster sampling. The cluster mean of the i-th cluster is denoted by Yi, and we have M cluster means in total.

The formula for the variance of the cluster mean (cl_y) can be derived as follows:

Step 1: Define the cluster mean (Yi) and the cluster total (Ti):

- Yi: The mean of the sampling units in cluster i.
- Ti: The sum of all sampling units' values in cluster i.

Step 2: Calculate the overall population mean (Y) and the total sum of squares (SS_total):

- Y: The overall population mean, which is the sum of all cluster totals ($T_1 + T_2 + ... + T_M$) divided by *N*.
- SS_total: The total sum of squares, which is the sum of the squared differences between each sampling unit and the overall population mean (Y).

Step 3: Calculate the sum of squares between clusters (SS_between):

- SS_between: The sum of squares between clusters, which represents the variation between the cluster means and the overall population mean (Y). It is calculated as the sum of the squared differences between each cluster mean (Yi) and the overall population mean (Y), each multiplied by the number of sampling units in the corresponding cluster (n_i).

Step 4: Calculate the variance of the cluster mean (cl_y):

- $Var(cl_y) = SS\_between / (N - M)$

Now, let's derive the formula using the provided notation:

Step 1: Define the cluster mean (Y_i) and the cluster total (T_i):

- Yi: Cluster mean for cluster i.
- Ti: Cluster total for cluster i.

Step 2: Calculate the overall population mean (Y) and the total sum of squares (SS_total):

- Y: The overall population mean is given by the formula: Y = ($T_1 + T_2 + \dots + T_M$) / N
- SS_total: The total sum of squares is given by the formula: SS_total = Σ_i(Ti - Y)^2, where Σ_i represents the sum over all clusters i.

Step 3: Calculate the sum of squares between clusters (SS_between):

- SS_between: The sum of squares between clusters is given by the formula: SS_between = Σ_i(ni * (Yi - Y)^2), where Σ_i represents the sum over all clusters i.

Step 4: Calculate the variance of the cluster mean (cl_y):

- Var(cl_y) = SS_between / (N - M)

The formula for Var(cl_y) is similar to the one you provided in your question:

Var(cl_y) = SS_between / (N - M)

where SS_between is the sum of squares between clusters, N is the total number of sampling units in the population, and M is the total number of clusters in the population.

$$\begin{aligned}\operatorname{Var}\left(\bar{y}_{nM}\right) & =\frac{NM-nM}{NM}\cdot\frac{S^2}{nM} \\ & =\frac{f}{n}\cdot\frac{S^2}{M}\end{aligned}$$

where $f=\frac{N-n}{N}$ and $S^2=\frac{1}{NM-1}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij}-\bar{Y}\right)^2$.
Also

$$\begin{aligned}\operatorname{Var}\left(\bar{y}_c\right) & =\frac{N-n}{Nn}S_b^2 \\ & =\frac{f}{n}S_b^2.\end{aligned}$$

Consider

ISSN:2153-182X, E-ISSN: 2153-1838
Vol. 17 No. 02 (2023)

$$
\begin{aligned}
(NM-1)S^2 &= \sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij}-\bar{Y}\right)^2 \\
&= \sum_{i=1}^{N}\sum_{j=1}^{M}\left[\left(y_{ij}-\bar{y}_i\right)+(\bar{y}_i-\bar{Y})\right]^2 \\
&= \sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{lj}-\bar{y}_i\right)^2+\sum_{i=1}^{N}\sum_{j=1}^{M}(\bar{y}_i-\bar{Y})^2 \\
&= N(M-1)\bar{S}_w^2+M(N-1)S_b^2
\end{aligned}
$$

where

$\bar{S}_{\mathrm{w}}^2=\frac{1}{N}\sum_{i=1}^{N}S_i^2$

$S_i^2=\frac{1}{M-1}\sum_{j=1}^{M}\left(y_{ij}-\bar{y}_i\right)^2$

i^{th} cluster.

$$
\begin{aligned}
E &= \frac{\operatorname{Var}(\bar{y}_{nM})}{\operatorname{Var}(\bar{y}_c)} \\
&= \frac{S^2}{MS_b^2} \\
&= \frac{1}{(NM-1)}\left[\frac{N(M-1)}{M}\frac{\bar{S}_w^2}{S_b^2}+(N-1)\right].
\end{aligned}
$$

Thus the relative efficiency increases when $\bar{S}_w^2$ is large and S_b^2 minimizes to a little size. When clusters are established so that the variation inside the clusters is as high as the variation between clusters, cluster sampling is effective.

Effectiveness in terms of intraclass correlation efficiency Within a cluster, the elements' intraclass correlation can be expressed as

$$
\begin{aligned}
\rho &= \frac{E\left(y_{ij}-Y\right)(y_{iz}-\bar{Y})}{E\left(y_{ij}-\bar{Y}\right)^2};-\frac{1}{M-1}\leq\rho\leq 1 \\
&= \frac{\frac{1}{MN(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k|\neq j)=1}^{M}\left(y_{ij}-\bar{Y}\right)(y_{ik}-\bar{Y})}{\frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij}-\bar{Y}\right)^2} \\
&= \frac{\frac{1}{MN(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k(*,j)=1}^{M}\left(y_{ij}-\bar{Y}\right)(y_{ik}-\bar{Y})}{\left(\frac{MN-1}{MN}\right)S^2}
\end{aligned}
$$

$$\sum_{i=1}^{N} (\bar{y}_i - \bar{Y})^2 = \sum_{i=1}^{N} \left[\frac{1}{M}\sum_{j=1}^{M} (y_{ij} - \bar{Y})\right]^2$$

$$= \sum_{i=1}^{N} \left[\frac{1}{M^2}\sum_{j=1}^{M} (y_{ij} - \bar{Y})^2 + \frac{1}{M^2}\sum_{j=1}^{M} \sum_{k(*j)=1}^{M} (y_{jj} - \bar{Y})(y_{ik} - \bar{Y})\right]$$

$$\Rightarrow \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k(*j)-1}^{M} (y_j - \bar{Y})(y_{ik} - \bar{Y}) = M^2 \sum_{i=1}^{N} (\bar{y}_i - \bar{Y})^2 - \sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{Y})^2$$

or

$$\rho(MN-1)(M-1)S^2 = M^2(N-1)S_b^2 - (NM-1)S^2$$

or $S_b^2 = \frac{(MN- \ )}{M^2(N-1)}[1+\rho(M-1)]S^2$.

The variance of $\bar{y}_d$ now becomes

$$\begin{aligned} \text{Var } (\bar{y}_c) \quad &= \frac{N-n}{Nn} S_b^2 \\ &= \frac{N-n}{Nn}\frac{MN-1}{N-1}\frac{S^2}{M^2}[1+(M-1)\rho]. \end{aligned}$$

For large $N, \frac{MN-1}{MN} \approx 1, N-1 \approx N, \frac{N-n}{N} \approx 1$ and so

$$\text{Var } (\bar{y}_d) \approx \frac{1}{n}\frac{S^2}{M}[1+(M-1)\rho]$$

Result and Discussion

The result of multistage cluster sampling is an estimator for the population mean. By using clusters as primary sampling units, it becomes easier and more cost-effective to estimate the population mean compared to sampling each individual element.

The efficiency of cluster sampling over Simple Random Sampling Without Replacement (SRSWOR) can be assessed using the intra-class correlation (ρ). A larger intra-class correlation indicates higher homogeneity within clusters and increases the efficiency of the estimator. Conversely, a small variance between cluster means contributes to higher efficiency.

Conclusion

Multistage cluster sampling offers a practical and effective approach for estimating population parameters in scenarios where individual sampling is impractical or resource-intensive. By selecting clusters as the primary sampling units and employing suitable sampling techniques, researchers can efficiently obtain representative data from large and geographically dispersed populations. However, the success of cluster sampling relies on careful cluster design and understanding the factors that influence the efficiency of the estimator.

References

Obermeyer Z, Murray CJL, Gakidou E: Fifty years of violent war deaths from Vietnam to Bosnia: analysis of data from the world health survey programme. BMJ. 2008, 336: 1482-1486. 10.1136/bmj.a137.

Roberts L: Commentary: Ensuring health statistics in conflict are evidence-based. Confl Health. 2010, 4: 10-10. 10.1186/1752-1505-4-10.

Mills EJ, Checchi F, Orbinski JJ, Schull MJ, Burkle FM, Beyrer C, Cooper C, Hardy C, Singh S, Garfield R: others: Users' guides to the medical literature: how to use an article about mortality in a humanitarian emergency. Conflict and Health. 2008, 2: 9-10.1186/1752-1505-2-9.

Morris SK, Nguyen CK: A review of the cluster survey sampling method in humanitarian emergencies. Public Health Nurs. 2008, 25: 370-374. 10.1111/j.1525-1446.2008.00719.x.

Checchi F, Roberts L: Documenting Mortality in Crises: What Keeps Us from Doing Better. Plos Med. 2008, 5: e146-10.1371/journal.pmed.0050146.

Burnham G, Lafta R, Doocy S, Roberts L: Mortality after the 2003 invasion of Iraq: a cross-sectional cluster sample survey. The Lancet. 2006, 368: 1421-1428. 10.1016/S0140-6736(06)69491-9.

Roberts L, Lafta R, Garfield R, Khudhairi J, Burnham G: Mortality before and after the 2003 invasion of Iraq: cluster sample survey. The Lancet. 2004, 364: 1857-1864. 10.1016/S0140-6736(04)17441-2.

Spiegel PB, Salama P: War and mortality in Kosovo, 1998–99: an epidemiological testimony. The Lancet. 2000, 355: 2204-2209. 10.1016/S0140-6736(00)02404-1.

Coghlan B, Brennan RJ, Ngoy P, Dofara D, Otto B, Clements M, Stewart T: Mortality in the Democratic Republic of Congo: a nationwide survey. The Lancet. 2006, 367: 44-51. 10.1016/S0140-6736(06)67923-3.

Depoortere E, Checchi F, Broillet F, Gerstl S, Minetti A, Gayraud O, Briet V, Pahl J, Defourny I, Tatay M, Brown V: Violence and mortality in West Darfur, Sudan (2003–04): epidemiological evidence from four surveys. The Lancet. 2004, 364: 1315-1320.

Turner AG, Magnani RJ, Shuaib MA: Not Quite as Quick but Much Cleaner Alternative to the Expanded Programme on Immunization (EPI) Cluster Survey Design. International Journal of Epidemiology. 1996, 25: 198-203. 10.1093/ije/25.1.198.

Rose A, Grais R, Coulombier D, Ritter H: A comparison of cluster and systematic sampling methods for measuring crude mortality. Bull. World Health Organ. 2006, 84: 290-296.

Grais F, Rose A, Guthmann J: Don't spin the pen: two alternative methods for second-stage sampling in urban cluster surveys. 2007

Working group for Mortality Estimation in Emergencies: Wanted: studies on mortality estimation methods for humanitarian emergencies, suggestions for future research. Emerg Themes Epidemiol. 2007, 4: 9-

Spiegel PB, Robinson C: Large-Scale"Expert" Mortality Surveys in Conflicts–Concerns and Recommendations. JAMA. 2010, 304: 567-10.1001/jama.2010.1094.