Biostatistics and Biometrics
Open Access Journal
ISSN: 2573-2633

Juniper Publishers
key to the Researchers

Review Article
Volume 5 Issue 5 - March 2018
DOI: 10.19080/BBOAJ.2018.05.555673

Biostat Biometrics Open Acc J

A Review on the Recent Development on the Cluster Sampling

Iqbal Jeelani M, Faizan Danish* and Mansha Gul

Department of Statistics and Computer Science, SKUAST-Jammu, Main campus, Chatha-180009 (J & K), India

Submission: December 29, 2017; **Published:** March 28, 2018

***Corresponding author:** Faizan Danish, Division of Statistics and Computer Science, Faculty of Basic Sciences, SKUAST-Jammu, Main campus, Chatha-180009 (J & K), #SKUAST-Kashmir, India, Tel: +919469177753; Email: danishstat@gmail.com

Abstract

Precise testing is a standout amongst the most common sampling technique. The fame of the systematic sampling is fundamentally because of its common sense. Problems of systematic sampling occur more frequently than is generally realized and, since many of the techniques are still far from satisfactory, the situation offers great incentive to further development. A fair amount of research has been done in this area with the main focus being directed to handling the problems that arise when using the cluster sampling design in practice. The main theme of the recent research in this area is merging the multi-start idea with one of the schemes that assures a fixed sample size. Compared with simple random sampling, it is less demanding to draw a cluster sample uniquely when the choice of test units is done in the field. The present paper offers an audit of the current work around there and gives a few proposals to study professionals utilizing the cluster sampling design for various testing circumstances.

Keywords: Cluster sampling; Super-population; Sampling variance

Abbreviations: SRS: Simple Random Samples; DE: Design Effect; EPI: Expanded Program for Immunization; NHB: National Horticulture Board

Introduction

Sample surveys are widely used as a means of collecting information to meet a definite need in agriculture, trade, social, educational and economical problems. It has been observed that the sample survey can give very precise information. Since in a sample survey, a part of population is surveyed and inference is drawn about the whole population, the results are likely to be different from the true population values, but the advantage with the sample survey is that this type of error can be measured and controlled. Similarly, the errors which arise due to human factor at the stage of ascertainment and processing of data can be eliminated considerably by employing suitable sampling techniques and properly trained persons in surveys. Sample survey is less time consuming, involves less cost, has greater scope in special coverage's and also has greater operational facilities as compared to complete enumeration. It is for these reasons that sample surveys are being preferred and adopted frequently by the government, scientific organizations, industries, institutes and others since the beginning of 20th century. The purpose of sampling theory is to make sampling more efficient. It attempts to develop methods of sample selection and of estimation that provide, at the lowest cost, estimates that are precise enough for our purpose. This principle of specified precision at minimum cost recurs repeatedly in the presentation of theory. In order to apply this principle, we must be able to predict, for any sampling procedure that is under consideration, the precision and the cost to be expected. So far as precision is concerned, we cannot foretell exactly how large an error will be present in an estimate in any specific situation, for this would require the knowledge of the true value for the population.

The precision of a sampling procedure is judged by examining the frequency distribution generated for the estimate if procedure is applied again and again to the same population. With samples of the sizes that are common in practice, there is often a good reason to suppose that the sample estimates are approximately normally distributed. In any sampling design an important element is the representativeness of the sample drawn from a population. The sample must include the characteristics as close as possible to the value that a researcher might have obtained had he been able to observe the universe or population. The difference between an estimate and the true value of the parameter being estimated constitute the sampling error, as such a good sampling technique is one which gives a smaller sampling error. The sample survey techniques are also commonly used for obtaining information on various social and economic activities of the society. Some specific situations in which sampling techniques can successfully be employed are;

i. When results are needed with maximum accuracy, with a fixed budget or with the minimum number of units having specified degree of reliability.

ii. when the units under investigation show considerable variation for the characteristic under study

iii. when a total count of the population is not possible or is very costly or destructive

iv. when scope of the investigation is very wide and the population is not completely known, and

v. When time, money and other resources are limited. There are various steps involved in the planning and execution of a sample survey. One of the principal steps in a sample survey relate to methods of data collection.

The different methods of collecting data include physical observation or measurement, personal interview, mail enquiry, telephonic enquiry, web-based enquiry, method of registration, transcription from records. The methods relate to collection of primary data from the units/respondents directly, while the last one relates to the extraction of secondary data, collected earlier generally by one or more of the methods. These methods have their respective merits and therefore sufficient thought should be given in selection of appropriate methods of data collection in any survey. The choice of the method of data collection should be arrived at after careful consideration of accuracy, practicability and cost from among the alternative methods.

New trends in cluster sampling

The enumeration of population by various sampling methods where first given by Laplace. These procedures came into widespread use only by the mid nineties of the 19th century. The first account of a strong plea for the use of samples in data collection was made by Kiaer at I.S.I. meeting in Berne and presented a report on his experience with sample surveys conducted in the Norwegian Bureau of Statistics. Simple random samples (SRS), where every member of population has an equal or predetermined chance of selection, and the sampling is single stage, are the simplest to visualize and the level of precision of results derived from them is easy to calculate. Neyman [1] discussed the role of random sampling and demonstrated that random sampling is not only a viable alternative, but also a much superior tool than purposive method of selection. In sample surveys experimenter introduces the probability element by adopting the technique of randomization. The idea of probability structure in planned experiments has been very well discussed by Fisher [2], where he has discussed the role of randomization in the selection of a part from the whole population, which provides a valid method of obtaining an estimate of amount of error committed. SRS are almost impossible to achieve in reality (due to imperfect sampling frames, non-response and so on) and , in any case they are relatively inefficient ways of obtaining a particular number of responses in terms of fieldwork and travelling costs, but such irregularities are absent in cluster sampling.

The efficiency of cluster sampling has been studied by Smith [3], Hansen & Hurtwiz [4], where it has been discussed that the relative efficiency of cluster sampling increases with the increase in mean square within clusters. Mahalonobis [5] studied the problem of determination of optimum cluster size from the points of view of both variance and cost, where he mentioned that for a given sample size, the sampling variance increases with cluster size and decreases with increasing number of clusters and on the other hand, the cost decreases with the cluster size and increases with the number of clusters. Hence, it is necessary to determine a balancing point by finding out the optimum cluster size and the number of clusters in the samples which can minimize the cost for a fixed variance. Hansen & Hurtwiz [6] discussed the role of sampling variance of an estimator in case of cluster sampling and proved that in case of cluster sampling, variance depends upon the number of clusters in the sample, the size of cluster, the intra-class correlation coefficient.

On the basis of many agricultural surveys Jessen [7,8], and Madow & Madow [9] developed a general law to predict how mean square within clusters changes with the size of cluster. Hansen & Hurtwiz [6] discussed that in many practical situations, cluster size is positively correlated with the variable under study and in these cases, it is advisable to select the clusters with probability proportional to the number of elements in the cluster. A good discussion of numerical values of intra-class correlation coefficient for different elements within cluster in cluster sampling have been given by Hurtwiz & Madow [10], they have shown the intra-class correlation coefficient as a "measure of homogeneity" of the clusters in cluster sampling. Several good introductory books have been written on Sampling, including Hurtwiz [10], Yates F [11]. Deming [12], Kish [13], Des Raj [14], and Cochran [15]. All these books present relatively leisurely introduction to the sampling methodology. Singh et al. conducted a study on fresh fruits in Tamil Nadu and discussed the role of cluster sampling for studying the cultivation practices and yield of guava and showed that cluster sampling has been found to be very suitable for studying the cultivation practices of horticultural crops in absence of sampling frames.

Paul Harris has discussed the role of DE (design effect) in cluster sampling, and has shown how clustering can make the true variance different from that of the theoretical one and showed that in practice, samples are never totality of simple random type and clustering can make the true variance different from that of the theoretical one. The extent to which it differs that is, the ratio of true to theoretical variance is known as design effect. Paul Harris has very well used the concept of intra-class correlation to access the design effect of clustered samples. Using Cochran's formula, he has shown one can calculate and summate the two components of variance in the clustered sample, that is within and between components and intra-class

How to cite this article: Iqbal J M, Faizan D, Mansha G. A Review on the Recent Development on the Cluster Sampling. Biostat Biometrics Open Acc J. 2018; 5(5): 555673. DOI: 10.19080/BBOAJ.2018.05.555673.

correlation measure , which is a measure of the total respondent to respondent variability in a survey measure, is accounted by differences between clusters. Collins & Goodhardt [16] have applied the intra-class correlation approach and have verified the assertion of the limited benefits of large cluster sizes in contributing to the levels of statistical precision.

A more advanced textbook on Sampling is written by Sukhatme & Sukhatme [17] in which an extensive treatment of inference has been given to agricultural surveys. A unique feature of this book is that a large number of exercises with real sets of agricultural data from various fields are included. Singh & Chaudary [18] have provided a thorough discussion on philosophical as well as theoretical issues in statistical analysis of survey data and main focus of this book is on agriculture. Machado [19] presented a comparison between the results obtained from a complete enumeration forest inventory and cluster sampling methods systematically distributed over the inventoried area located in the tropical rain forests of Brazilian Amazon and found that total volume and the number of trees for all species obtained from cluster sampling were very close to their true value. Shackman [20], has discussed the role of design effect in case of cluster sampling and has shown how design effect is used to determine how much larger the sample size or confidence interval needs to be and discussed how design effect in case of cluster sampling increases as the cluster size increases, and as the intra-class correlation increases. Gilbert [21] has discussed the role of cluster sampling for estimation of mortality in Iraq due to the invasion of American troops and showed that cluster sampling can be used to estimate high mortalities in cases such as wars, famines and natural disasters.

A good discussion of regression analysis and cluster sampling has been given by Wretman [22], where he has proved that to estimate the variance of the regression coefficients correctly, one should include the information of clustering in regression analysis. Paul Milligan [23] discussed the role of cluster sampling in expanded Program for immunization (EPI), where he has showed how cluster sampling can be used to estimate the vaccination coverage, when an up-to-date household sampling frame is not available. Tipping & Pickering [24] carried out some work to look how precision of survey estimate, design effect and intra-class correlation values for health measure changes with size of the cluster and showed that it is possible that values can potentially double as the cluster size is halved and collaboration of design effect, intra-class correlation and precision of survey estimates is very important in cluster sampling. A good discussion of cluster sampling has been demonstrated by Andrew [25], where the impact of various aspects of cluster sampling on the level of statistical reliability of a survey has been discussed. Saifuddin [26] has discussed the role of cluster sampling in the estimation of health insurance coverage in Baltimore city UK and has discussed the balance between variance efficiency and cost efficiency in case of cluster sampling.

A good discussion on the effect of auxiliary information on the variance of cluster sampling has been proposed by Nina & Zhang [27], where they have discussed that the use of auxiliary information removes the extra variance that is due to the variation in the cluster sizes. Moreover, it reduces the loss of efficiency to the extent it reduces the conditional intra-class correlation given the covariates and also in case of cluster sampling one generally needs to balance between the likely loss of efficiency as compared to sampling of elements and potential administrative and operational advantages that are important in practice. In recent years horticulture has emerged as an important component of the Indian economy, and yearly information of Horticulture crops is given by National Horticulture Board (NHB, Ministry of Agriculture, Govt of India). Tauqeer [28] has provided a number of methodological issues related to apple data and has carried a series of surveys to evolve a sampling methodology for estimating area and yield of fruits and vegetables and has given extensive discussion on choice of different sampling procedures and the relations between the statistical methods and their application in agricultural research. Chandra [29] has discussed the role of R software in survey data analysis, and main emphasis is given on stratified and cluster sampling.

Rama Rao [30] has discussed that cluster sampling becomes statistically more efficient if each cluster can be made to represent most of the possible observations that can be obtained from the universe and in contrast, if each cluster represents only a few different universe observations then cluster sampling will be less statistically efficient than a simple random sample of the same size. Leo [31] has demonstrated that cluster sampling should be used only when it is economically justified or when reduced costs can be used to overcome losses in precision and showed that given a fixed budget, the researcher may be able to use a bigger sample with cluster sampling than with the other methods. When the increased sample size is sufficient to offset the loss in precision, cluster sampling may be the best choice. Venables & Ripley [32] is the excellent book published on R and S plus which deals with introductory as well as advanced concepts of S programming with suitable examples. Lumley [33], is the most modern contribution of sample surveys using R-software, where he has explained the use of survey packages in sample surveys.

Jeelani et al. [34] discussed a common motivation for cluster sampling is to reduce the average cost per interview given a fixed budget this can allow an increased sample size. Assuming a fixed sample size the technique gives more accurate results when most of the variation in the population is within the groups, not between them. The present work is an attempt to show that cluster sampling is more efficient than simple random sampling provided the mean square within the clusters is maximum and there is a negative intra-class correlation coefficient between elements within clusters as relative efficiency of

How to cite this article: Iqbal J M, Faizan D, Mansha G. A Review on the Recent Development on the Cluster Sampling. Biostat Biometrics Open Acc J. 2018; 5(5): 555673. DOI: 10.19080/BBOAJ.2018.05.555673.

cluster sampling increases with increase in mean square within clusters. Different estimators of cluster sampling are applied and their results are compared with simple random sampling using the same sample size. Different computer programmes are developed using R-software. All these functions are run on real data set generated on Apple crop in year 2010-11 from district Ganderbal of Kashmir valley. The programme they developed in R-Software are

cluster1(x,N)

#This is function developed in R-software. It takes the arguments x (name of the data), N (total number of clusters) and returns cluster mean (ynbar), cluster variance (vynbar) and standard error (se).

The codes of the function follow

cluster1(x,N)

#This is function developed in R-software. It takes the arguments x (name of the data), N (total number of clusters) and returns cluster mean (ynbar), cluster variance (vynbar) and standard error (se).

The codes of the function follow

cluster1<-function(x,N)

{ x=data.frame(x) #N=total no. of clusters in the population n=nrow(x)

m=ncol(x)

yibar=apply(x,1,mean)

myibar2=sum(m*yibar^2)

si2=apply(x,1,var)

ynbar=sum(yibar)/nyibar2=sum(yibar^2)nybar2=ynbar^2 vynbar=((1/n)-(1/N))*(1/(n-1))*(yibar2-n*nybar2)

s e = s q r t (v y n b a r) list(clusterMean=ynbar,ClusterVariance=vynbar,se=se)

Recently, Lundberg & Strand [35] studied several estimators, not including the striplet estimator of Fewster [36], for the sampling variance of the two-dimensional systematic sampling design when applied in land use surveys. They concluded that variance estimation by stratification gives good overall results but may underestimate the variance when spatial autocorrelation is absent while treating the systematic sample as a SRS is safe and conservative when spatial autocorrelation is absent or unknown. It seems like there is a growing literature in this specific area and in the area of spatial sampling in general [37-39].

Conclusion

Sampling method used when assorted groupings are naturally exhibited in a population, making random sampling from those groups possible. The use of the technique requires the division or classification of the population into groups, defined by their assorted characteristics or qualities. In this paper we have made the attempt to cover all the developments towards the cluster sampling from the several authors. From the manuscript it can be concluded that the cluster sampling is one of the designs of sampling which has taken the keen interest not only of statisticians but also non-statisticians have contributed a lot in the field. Thus the cluster sampling has better usage rather than other types of sampling design in real life.

References

1. Neyman J (1934) On the two different aspects of representative method; The method of stratified sampling and the method of purposive selection. Jour Roy Stat Soc 97: 558-606.
2. Fisher RA (1939) The comparison of samples with possibly unequal variances. Annuals of Eugenetics 9(2): 174-180.
3. Smith TMF (1938) The foundations of survey sampling: a review. Jour Roy Stat Soc A 139: 183-204.
4. Hansen MH, Hurwitz WN (1943) On the theory of sampling from finite populations. Ann Math Stat 14: 333-362.
5. Mahalanobis PC (1940) Report on the sample census of jute in Bengal. Indian Central Jute Committee.
6. Hansen MH, Hurwitz WN (1946) The problem of non- response in sampling surveys. Jour Amer Stat Assoc 41: 517-529.
7. Jessen A (1926) Report on the representative method in statistics. Bull Int Stat Inst 22: 359-380.
8. Jessen A (1926) The representative method in practice. Bull Int Stat Inst 22: 381-439.
9. Madow WG, Madow LH (1944) On the theory of systematic sampling. Ann Math Stat 15: 1-24.
10. Hurtwiz WN, Madow WG (1953) Sample Survey Methods and Theory. New York: John Willey and sons.
11. Yates F (1946) A review of recent statistical developments in sampling and sampling surveys. Jour Roy Stat Sco A 109: 12-42.
12. Deming WE (1960) Sample Design in Business Research. John Willey and Sons, NewYork, USA.
13. Kish L (1965) Survey sampling. John Willey and sons, NewYork, USA.
14. Raj D (1972) The Design of Sample Surveys. McGraw-Hill, NewYork, USA.
15. Cochran WG (1977) Sampling Techniques. John Wiley and Sons, New York, USA.
16. Collins M, Goodhardt G (1978) Value for money in research design, paper presented at the Proceedings of the MRS Annual Conference, Brighton.
17. Sukhatme PV, Sukhatme BV, Sukhatme S, Ashok C (1984) Sampling theory of surveys with applications. Iowa State university Press, U.S.A.
18. Singh D, Chaudary F (1985) Theory and Analysis of Sample survey Designs. New Age Publications, New Delhi, India Pp 196-221.
19. Machado SA (1985) Complete enumeration forest inventory versus cluster sampling method applied in the Amazonic rain forests. Revista Florista 52: 122-130.
20. Shackman G (2001) Sample size and Design effect. Paper presented at American Statistical Association. New York.

How to cite this article: Iqbal J M, Faizan D, Mansha G. A Review on the Recent Development on the Cluster Sampling. Biostat Biometrics Open Acc J. 2018; 5(5): 555673. DOI: 10.19080/BBOAJ.2018.05.555673.

21. Gilbert B (2006) Mortality after the 2003 invasion of Iraq: A cross sectional cluster sample survey. Lancet 368(9545): 121-128

22. Wretman (2003) Use of regression analysis in random cluster samples. Paper presented at American Statistical Association. New York, USA.

23. Milligan P (2003) Comparison of two stage cluster sampling methods for health surveys in developing countries. Int J Epidemiol 45: 221-234.

24. Tipping S, Pickering P (2004) Impact of geographical size of clusters on the precision of survey estimates. Survey Methods Newsletter 23 Winter.

25. Andrew H, Roger S (2006) Cluster sampling: a false economy. International Journal of Market Research 47: 231-239.

26. Saifuddin M (2006) Methods in Sample Surveys: Cluster sampling. Journal of American Statistical Association 80: 111-123.

27. Nina H, Zhang L (2009) A note on the effect of Auxiliary information on variance of cluster sampling. Journal of official Statistics 25: 397-404.

28. Tauqueer A (2007) Methodological issues related to Horticultural Statistics. Indian Agricultural Statistics Research Institute (IASRI). New Delhi, India.

29. Chandra H (2009) Analysis of survey data using R-software. Indian Agricultural Statistics Research Institute (IASRI). New Delhi, India.

30. Rao R (2009) History and Development of survey based estimation and analysis. Survey Methodology 30: 3-29.

31. Leo L (2009) Role of cluster sampling in market research. International Journal of Market Research 49: 306-321.

32. Venables WN, Ripley BD (2009) Modern Applied Statistics with S-PLUS, 4th edition, Springer Verlag, New York, USA.

33. Lumely T (2010) Complex survey samples in R R Development Core Team. University of Washigton, Department of Biostaistics, USA.

34. Jeelani MI, Mir AH, Maqbool S, Nazir N, Shah AA (2012) Cluster sampling of survey data and their application agriculture using R-software. International Research Journal of Agricultural Economics and Statistics 3(1): 35-39.

35. Lundberg L & Strand GH (2014) Comparison of variance estimation methods for use with two-dimensional systematic sampling of land use/land cover data. Environmental Modelling & Software 61: 87-97.

36. Fewster RM (2011) Variance estimation for systematic designs in spatial surveys. Biometrics 67(4): 1518-1531.

37. Bowley AL (1926) Measurement of the precision attained in sampling. Bull Int Stat Inst 22(1): 6-62.

38. Hansen MH, Hurwitz WN, Madow WG (1953) Sample survey Methods and Theory. I&II, Wiley, New York, USA.

39. Tyagi KK (2001) Sample survey techniques in Agricultural research: Lecture notes. Indian Agricultural Statistics Research Institute (IASRI). New Delhi, India.

DOI: 10.19080/BBOJ.2018.05.555673.

How to cite this article: Iqbal J M, Faizan D, Mansha G. A Review on the Recent Development on the Cluster Sampling. Biostat Biometrics Open Acc J. 2018; 5(5): 555673. DOI: 10.19080/BBOAJ.2018.05.555673.