A Brief Review of Conditions, Circumstances and Applicability of Sampling Techniques in Computer Science Domain

Nabila Amir [1], Fouzia Jabeen[1], Sidra Niaz[1]

Department of Computer Science of Shaheed Benazir Bhutto Women University, Peshawar

nabilaameer1@gmail.com

fouzia.jabeen@sbbwu.edu.pk

sidraniaz@sbbwu.edu.pk

2020 IEEE 23rd International Multitopic Conference (INMIC) | 978-1-7281-9893-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/INMIC50486.2020.9318209

Abstract—**It is tricky to examine the whole population. To save time and money researchers select a small group called sample for best representation of the whole population. Sampling technique is used for choosing sample in the population. Sampling techniques has two types': probability and non-probability. In research the sampling techniques play an important role, for decision making and analysis of data in different fields. In this paper, we have presented a review of sampling techniques from different perspectives, to enable researchers to select a best sampling technique while doing research. After study the review they will come to know which sampling technique is best suited according to the resources (time, money and effort etc.) available in the domain.**

Keywords—Population, Sample, Sampling, Random, Systematic

I Introduction

It is uncertain that the researchers should be capable to collect data from whole population. Therefore, it is significant to choose sample. The whole set of data from which the researcher select samples are called Population. Researchers applying different sampling techniques to drawn samples from the whole population, because of limitation of time and resources to examine the whole population[1]. There are various sampling techniques used in statistics, survey methodology and quality assurance [2]. Sampling techniques are divided into two categories such as: Probability sampling and non-probability sampling.

Probability sampling is the way of choosing the sample components that ensures each element of population occur the same chance of selection[3]. There are different types of probability random sampling. The types of probability random sampling are: 1) *Simple random sampling* is a technique in which each and every member of the whole population has an equal chance of begins chosen. The drawback of simple random sampling is[1]:A complete list of all elements of the entire population must be available. In some cases, the gaining of sample can be expensive if the factors were scattered biologically, e.g. while surveys are conducting from personal interviews. The average errors of estimators can be high.2) A sampling technique in which element are chose at regular interval (in terms of time, space or order) is called systematic random sampling[4]. (3)In stratified random sampling, the entire population is classified into small categories known as strata. These strata are chosen in such a way that they show each element of the population with same characteristics [5]. It is a technique which tries to limit the likely samples to those which are "less extreme" by following that whole parts of the population are represented in the sample, because to increase the efficiency, that is to increase the error in judgment[6].Some draw backs of stratified random sampling are: (4*) Cluster sampling* is another technique, in which the population is classified into samples (units). In which each group is called clusters, which must be representative for the population. In which the heterogeneity of the population is clear. But each unit in the cluster is same characteristics[8]. (5) *Multistage sampling* can be a complex form of cluster sampling as it also divides the population into groups on the bases of their geographical location. Then one or more clusters are chosen for sampling. Multistage sampling and cluster sampling are sometime creates confusion, as multistage sampling follow a hierarchical structure of existing clusters within the population[9].

Non-probability sampling is technique in which some of the elements have zero chance of selection or in which the probability of selection is not correctly determined[10].The types of non-probability sampling are:(1)A *volunteer sampling* technique may be used when the other techniques are not possible. Generally samples of volunteers should be selected with care. Though, since all survey research include some degree of volunteering[11].(2)*Convenience sampling* is best

978-1-7281-9893-4/20/$31.00 ©2020 IEEE

technique as they are often willingly and easily available. Normally, this technique is best suited sampling technique for students because it is an inexpensive and an easy option as compared to other sampling techniques[12]. Convenience sampling frequently uses to solve a number of issues related with research. For example, instead of targeting unknown individuals, it is easier to use friends or family as part of sample [13]. (4)*Purposive sampling* technique also called judgmental, selective, or subjective sampling technique. A purposive sampling technique uses for the selection of units based on personal judgment rather than randomization. This critical sampling technique is the "representative" of the population of interest without sampling at random[11].(5)A non-probability sampling technique called *Quota sampling*, in which samples are choose on the basis of predetermined properties so that the total sample will have the similar distribution of properties as the larger population[13].(6) In *snowball sampling* technique one member of the population is approached at a time and then is asked to refer the investigator to the other elements of the population. This technique is most appropriate for small populations that are hard to access due to their closed nature (secret societies and inaccessible professions) [13].

To best of our knowledge in this paper have covered the three aspects of sampling techniques: 1. Investigation of the conditions and circumstances to determine best utilization of sampling techniques, 2. To identify which sampling techniques are used in combination to achieve best results, 3. To identifying the application areas of sampling techniques in computer science domain.

The purpose to review the literature is to able the researchers about the importance of knowledge in probability and statistical needs in their research studies in different fields such as in modern computer science, software engineering, and other fields. These needs arise to make decisions under uncertainty. Probability and statistics pay an important role to solve the problems and make decisions under uncertain conditions in computer science, calculate probabilities and forecasts, evaluate performance of computer systems and networks.

The rest of paper is ordered in following sections: In section 2, we discuss the condition to determine the suitability of sampling techniques. In section 3, the discussion is about to find which of the sampling techniques can be used in combination. In section 4, we dig out Computer Sciences' domain application areas in which sampling techniques can be best applicable. In section 5, a comparative study of sampling techniques is done. In the last section 6 we end up with conclusion.

I. ANALYSIS OF CONDITION AND CIRCOMSTANCES TO DETERMINE BEST UTILIZATION OF SAMPLING TECHNIQUES

For making a research meaningful, beneficial and successful, it is important to select an appropriate sampling technique. In some cases, probability sampling techniques are better to use but in various cases non-probability sampling techniques are best suited. Probability sampling techniques are not good when population is infinite and not precisely defined then in such conditions non-probability sampling techniques are better to use. Moreover, it also depends on the presence of resources (i.e. time, money, efforts etc.)[6]. Probability sampling techniques is used when enough resources are available; otherwise the non-probability sampling technique is preferred.

A. The Conditions for Appropriate Applicability Of Probability Sampling Techniques

Researchers select probability sampling technique according to the requirement of research and the presence of resources needed during research. Now researchers have to choose the type of probability sampling technique which is precise and representative for researcher requirement. So, for this purpose we presented some guidelines to choose an appropriate probability sampling technique, which are following:

- If the population which chooses for study is homogenous and list of elements are complete and exist, data collection can be done efficiently on random distributed items, and the cost of sample is small enough to make efficiency less important than simplicity, then simple random sampling is better choice to use.

- If you select population which is homogenous, and list of elements cannot be created for any reason so systematic sampling technique is used [6].

- If the population you choose is heterogeneous and is not distributed in a very large geographical area, then stratified random sampling technique is best [6]. Stratified random sampling is also used when variability within strata is large, variability between strata is less and the variables upon which the population is stratified are highly correlated with the require dependents variables.

- If the population is homogenous but distributed in a large area, cluster sampling is used. Cluster sampling also suitable in conditions if the available list of elements is not reliable and it is costly to prepare, or if the element list is available but the location or detection of the elements may be tricky.

- If the population is homogenous and also distributed over large area then multistage sampling is used. When all elements in the selected clusters may be costly or not necessary, then also multistage cluster sampling is useful.

B. *The Conditions for Appropriate Applicability of Non-Probability Sampling Techniques*

If researchers want to selection-probability sampling technique, then for research i.e. non-probability sampling techniques suited with the resources and condition for your research then find out what non-probability sampling technique is good choice for your research. The following is the list of guidelines to choose best technique.

- The *volunteer sampling* technique is best for saving time and/or efforts in search of units.

- If your chosen population is very general type i.e. found everywhere and not completely define the *convenient sampling* technique is used [15].

- If the population is heterogeneous also it is essential to take representation from each subgroup, and then choose *quota sampling*. Researchers also prefer quota sampling when to study characteristics of a particular subgroup or determine the relationships between various subgroups [16].

- If the elements of that population which you selected are not easily available (hidden), lesser in number and also difficult to find, then the *snowball sampling* technique is to use [7].

In the next section we did analyses to find out which sampling techniques can be used in combination for obtaining fruitful results.

III THE COMPATIBILITY OF SAMPLING TECHNIQUES WITH EACH OTHER

In this section, we check the compatibility of sampling techniques with each other in different conditions. The analysis in this section helps in making decision to opt for best combination of sampling techniques in a problem domain.

A. *Stratified Cluster Sampling*

Cluster and stratified sampling are used with combination because the whole population can be classified into n strata and a cluster sample can be chosen from each stratum[17].

B. *Stratified Systematic Sampling*

Stratified sampling can also merge with systematic sampling. This is suitable to use in case when we select to stratify the sample on the bases of some criteria and from each stratum select an individual systematic sample with starting points. The result of this technique is more precise than the simple stratified random sampling if systematic sampling within strata is more precise than simple random sampling[18].

C. *Simple random, systematic and stratified sampling*

The above stated techniques are used to analyze loss of information and imbalances during sampling of network traffic. These techniques are used in combination for different purposes. For example to handle imbalances and loss of information[19].

In section 4 we talk about the application area of sampling techniques in computer science domain, to show the importance of sampling techniques in computer science domain.

IV. APPLICATION OF SAMPLING TECHNIQUES IN COMPUTER SCIENCE DOMAIN

Sampling techniques have many applications in computer science area. In this section we have categorized the applications areas in which the sampling techniques are successful. The brief analysis and discussion is done in this section.

A. *Role of Simple Random, Systematic and Stratified random Sampling in Networking*

Networked computers communicate with each other to share their data with each other, and they are growing day by day. Because of which the threats in this environment takes place. Against these network anomalies preventive action can be taken through analyzing the data. But because of very huge growth of network traffic, to process the entire data is not feasible. Different sampling techniques have analyzed loss of information and imbalances during sampling of network traffic. Such as simple, systematic and stratified random sampling techniques and also a new technique re-sampling (Under-sampling & over-sampling) is used. In these sampling techniques stratified sampling play important role in overcoming the loss of information while in simple sampling and systematic sampling loss of information take place which cause for wrong decision. But stratified sampling does not able to handle the imbalances of network traffic dataset. For this problem of imbalances of network traffic re-sampling technique is used to solve this issue[19].

Networks are usual representation of data in different region like information networks, biological and social networks. Though, the networks are massive, because of which conventional analytical techniques are impossible. Data decreasing methods are crucial. Various methods have been used to complete the challenge of data decreasing, which range from principle component analysis to clustering analysis. Sampling techniques play an important role in the massive data analysis. In the work [20]the author propose research graph sampling techniques to speed up the huge network data analysis. For this aim, the authors proposed a model of estimate degree distribution and using stratified strategy in multifaceted networks. Stratified sampling technique is one of the probability sampling techniques which is used for solving various problems prominently in networking domain like to speed up the huge network data analysis and complex networks which is efficient and give unbiasedness [20].

C. *Role of Cluster Sampling in Data Mining, Image Handling, collaborative filtering and big data*

- Cluster sampling technique is used for the organization of a huge number of unlabeled face images for different areas ranging from social media to law enforcement. In these areas there are huge number of faces up to hundreds of millions, and unfortunately this data is not clean and clear because often the face image labels are either missing or noisy[21].

- Nowadays the size of data is growing rapidly, from which a valuable data extraction and analyzing is very tedious task**.** For this task several data mining techniques are used to extract this important information from big data. The accessible data mining algorithms and conventional sampling algorithms can resolve the data size issue but has limitation when work with big data. Therefore, decreasing the size of data and performing sampling investigation of data. The authors [22]proposed an efficient technique called cluster sampling arithmetic. The algorithm work by dividing the data set into small multiple data blocks, which execute in parallel on multiple distributed nodes. In this way the size of data is reduced and show strong scalability for big data and ensuring the distribution property of data information. It is commonly used in marketing research[14].

- Recommendations of movies, books and CDs dependent on overlap of concentration are often called collaborative filtering. U. Lyle H et al[23] planned an official statistical representation of collaborative filtering, and evaluate different algorithms for testing the model parameters involving variations of K-means clustering and Gibbs Sampling. The proposed representation is simply extended to handle clustering of objects with various attributes[23].

- H. Bangui et al studies on big data technologies connected clustering algorithms and probable usage of IoT (internet of things). This work identified a group of research problems that can be utilized as research program for big data clustering . The goal of this study is to identifying and merging the research gaps between Big data clustering algorithms and IoT[24].

D. *Role of Stratified sampling in Pattern Recognition*

Y. Ye et al[25] presented a stratified sampling technique in order to choose the feature subspaces for random forest with high dimensional data. The benefit of using stratified sampling is that it ensures that each subspace consist of sufficient informative features for grouped in high dimensions. Testing on both synthetic data and a range of real datasets in gene classification, image categorization, and face recognition data sets consistently shows the effectiveness of this proposed method. The work is demonstrated to better that of state-of-the-art algorithms such as SVM (support vector machine), the random forest's four variant such as RF (relative frequency), enrich-RF, oblique-RF and nearest neighbor (NN) algorithms.

E. *Role of Simple random sampling in DBMS*

This work[26] reviewed the recent literature based on sampling from databases. The key techniques for at least simple random sampling from databases are now enough well understood to allow the development of model implementation of DBMSs which involves sampling facilities. The work of statistician, auditors, and query optimizers will greatly be supported by inclusion of sampling operators in DBMS.

F. *Role of Simple and Stratified random Sampling in Evaluation of Alignment for Ontology Matching*

This work presenting two statistically-founded evaluation methods called simple and stratified random sampling, in order to assess ontology-matching performance that are based on alignment application. Some researchers investigate the performance of an alignment and the other investigates the alignment itself. For the evaluation of alignment stratified random sampling has two main advantages as compare to simple random sampling. (1) The individual estimation of subpopulation makes it simpler to study the situation for the performance of corresponding techniques. In case, if the strata are selected in way that differentiates among various usage of the correspondences, one can conclude about the behavior of the correspondences in a domain. (2) Evaluation outcomes for the whole population obtain by combing the results from stratified random sampling are more precise as compare to simple random sampling[5].

This work[27] investigates snowball sampling, arecruitment technique that uses research into participant's social networks to access precise population. Initiate with the hypothesis that research is 'formed', the paper present one account of snowball sampling and using social networks to 'make' research. Due to low number of potential participants or the sensitivity of the topic, snowball sampling is often used. This work considers how the employ technique of snowball sampling that uses interpersonal relations and links between people, both includes and excludes individuals. Considering this, the paper challenge that due to use of social networks and interpersonal relations, snowball sampling forms how individuals take action and cooperate with groups, couple interviews and interviews.

V. COMPARATIVE STUDY OF SAMLING TECHIQUES

In this section we compare the sampling techniques in different perspectives such as in terms of efficiency, way of selecting samples, types of population etc. We categorize this comparison in following two groups.

1. Probability sampling with each other
2. Probability sampling with non-probability sampling techniques

A. *Comparison of Probability Sampling with each other*

Here we compare probability sampling techniques with each other to know which probability sampling technique is efficient to use and which sampling technique give more precision, and for which type of data it is used.

Simple and stratified random sampling techniques are used for the selection of samples in a population. But both techniques behave different in the selection of samples. Simple random sampling technique randomly select the samples in the population thus each sample is equally likely to happen. While stratified random sampling classified the whole population into smaller subgroups called strata on the bases of shared properties. Then a random sample is selected from each stratum which directly proportion to the size of stratum. At the end samples subgroups are merging to create random sample.
Stratified random sampling technique generally give more correct representation of population then simple random sampling as a complete list of population from each stratum is created [15].

Systematic random sampling is a probability sampling technique which is popular because of its practicality. Systematic sampling with contrast to simple random sampling is easier to select systematic samples when the selection is performing in the domain. When explicit or implicit stratification is available in the sampling context systematic random sampling also give large precise estimators as compared to *simple random sampling* [4].

Cluster and *stratified sampling* are probability sampling techniques and selection of samples are done randomly in both the techniques. But the ways of sampling are different based on population i.e. when the population is homogenous the cluster sampling technique is applied. In the case when there is heterogeneity exist in the population the stratified sampling technique is used[13]. The addition of cluster or strata makes the main difference between cluster sampling and stratified random sampling. In cluster sampling the researchers only choose several clusters from groups of clusters randomly. While in stratified sampling all the strata from whole population are sampled[29]. Table 1 summarizes the comparison.

B. *Comparison of Probability Sampling with Non-Probability Sampling Techniques*

Stratified is probability sampling technique while *Quota* is non-probability sampling technique. In stratified sampling the samples are randomly chooses while in quota sampling the samples are chooses non-randomly. But there are some similar points in these two techniques such as: both the techniques are used for heterogeneous population. In both techniques population is classified into smaller subgroups[13].

Convenience sampling is non-probability sampling technique. It is inexpensive and requires less effort. It is casual and easy as compared to simple random sampling (where the availability of a well-defined population is mandatory, if the list of elements not present then create, sample randomly from the list. Therefore, random sampling require a lot of effort[15].

Purposive sampling is non-probability sampling technique as compared to simple and stratified sampling techniques. In simple random sampling and stratified random sampling every sample has a known probability of choosing, while in a purposive or judgmental sampling the units that are selected as samples by researchers are exclusive[30]. Table 2 summarizes the compression.

Table 1:Comparison among probability sampling technique

	Simple random sampling	**Stratified random sampling**	**Cluster sampling**	**Systematic sampling**	**Multistage sampling**
Type of population	Homogenous	Heterogeneous	Homogenous	Homogenous	Homogenous
Availability of list of elements	Complete element list	Elements with similar properties	Available element is not reliable	Element list is not available	Element list may be costly
Way of selection of elements	Random selection	Classification of population in strata	Form clusters of population	Selection base on regular interval	Step-by- step clusters take place

Table 2. Comparison among probability sampling and non-probability sampling techniques

	Simple random sampling	Stratified random sampling	Quota sampling	Purposive sampling	Convenience sampling
Sampling type	Probability sampling	Probability sampling	Non-probability sampling	Non-probability sampling	Non-probability sampling
Population type	Homogenous	Heterogeneous but strata are randomly selected	Heterogeneous also it is essential to take representation from each subgroup	The criteria are predefined	Very general type (found everywhere)
Way of selection of sampling	Random selection	Random selection of strata	Non-random selection	Exclusively selection	Those units are selected which are easily available
Performance	Costly and time Consuming	Costly and require a lot of effort	Inexpensive	Inexpensive	Inexpensive

VI. CONCLUSION

This review provides a brief discussion about the selection of sampling techniques for two types of problems (heterogeneous or homogeneous also for the availability of resources). In the present study by reviewing the literature we conclude that probability sampling techniques are commonly used as compare to non-probability sampling techniques. Stratified sampling and cluster sampling play very important role in computer science domain such as in networking, data mining for extraction of big data and for other problems. Simple random sampling also has role in computer science. Furthermore, most sampling techniques are suitable to use with each other in different circumstances, which provide a way to researchers to select best couples for their research study.

REFERENCES

[1] H. Taherdoost and H. Group, “Sampling Methods in Research Methodology; How to Choose a Sampling Technique for Research,” no. September, 2017.

[2] “Sampling.” [Online]. Available: https://en.wikipedia.org/wiki/Sampling_(statistics. [Accessed: 06-Jul-2019].

[3] N. Hospital, “Probability Sampling - A Guideline for Quantitative Health Care Research,” vol. 12, no. 2, pp. 95–99, 2015.

[4] S. A. Mostafa and I. A. Ahmad, “Recent developments in systematic sampling: A review,” *J. Stat. Theory Pract.*, vol. 12, no. 2, pp. 290–310, 2018.

[5] W. R. Van Hage, A. Isaac, and Z. Aleksovski, “Sample Evaluation of Ontology-Matching Systems,” 2007.

[6] S. I. Classification, “Stratified random sampling.”

[7] P. L. Barreiro and J. P. Albandoz, “Population and sample . Sampling techniques,” 2001.

[8] P. Sedgwick, “Multistage sampling,” no. August, pp. 9–11, 2015.

[9] “research methods.” [Online]. Available: https://courses.lumenlearning.com/atd-herkimer-researchmethodsforsocialscience/chapter/chapter-8-sampling/. [Accessed: 06-Jul-2019

[10] S. Elder, *Sampling methodology. International Labour Office. (2009).* .

[11] R. L. Ackoff, “Book Reviews,” p. 1954, 1954.

[12] H. Taherdoost, H. Business, S. Sdn, C. Group, and K. Lumpur, “Sampling Methods in Research Methodology ; How to Choose a Sampling Technique for,” vol. 5, no. 2, pp. 18–27, 2016.

[14] “Clustering sampling.” [Online]. Available: http://home.iitk.ac.in/~shalab/sampling/chapter9-sampling-cluster-sampling.pdf. [Accessed: 02-Jul-2019].

[15] W. J. Lammers and E. Babbie, “Sampling Techniques,” *Fundam. Behav. Res.*, pp. 1–23, 2005.

[16] “Quota-sampling.” [Online]. Available: https://blog.socialcops.com/academy/resources/quota-sampling-when-to-use-how-to-do-correctly/. [Accessed: 10-Jul-2019].

[17] “Multistage cluster sampling.” [Online]. Available: http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/MetQ/Talk4.pdf. [Accessed: 08-Jul-2019].

[18] “Multistage sampling.” [Online]. Available: http://www2.stat-athens.aueb.gr/~jpan/diatrives/Tzavidis/chapter4.pdf. [Accessed: 06-Aug-2019].

[19] R. Singh, H. Kumar, and R. K. Singla, “Issues Related to Sampling Techniques For Network Traffic Dataset,” vol. 3, no. 4, pp. 75–85, 2013.

[20] J. Zhu, H. Li, M. Chen, Z. Dai, and M. Zhu, “Enhancing stratified graph sampling algorithms based on approximate degree distribution,” *Adv. Intell. Syst. Comput.*, vol. 764, pp. 197–207, 2019.

[21] C. Otto, S. Member, D. Wang, and A. K. Jain, “Clustering Millions of Faces by Identity,” vol. 8828, no. c, pp. 1–14, 2017.

[22] J. Zhao, J. Sun, Y. Zhai, Y. Ding, C. Wu, and M. Hu, “A Novel Clustering-Based Sampling Approach for Minimum Sample Set in Big Data Environment,” *Int. J. Pattern Recognit. Artif. Intell.*, vol. 32, no. 2, p. 1850003, 2018.

[23] L. H. Ungar and D. P. Foster, “Clustering Methods for Collaborative Filtering What is Collaborative Filtering ?,” pp. 114–129.

[24] H. Bangui, M. Ge, and B. Buhnova, “Exploring Big Data Clustering Algorithms for Internet of Things Applications,” no. IoTBDS, pp. 269–276, 2018.

[25] Y. Ye, Q. Wu, J. Zhexue, M. K. Ng, and X. Li, “Stratified sampling for feature subspace selection in random forests for high dimensional data,” *Pattern Recognit.*, vol. 46, no. 3, pp. 769–787, 2013.

[26] C. S. Division and S. Jose, “Random sampling from databases ." a survey,” 1995.

[27] K. Browne and K. Browne, “Snowball sampling : using social networks to research non - heterosexual women Snowball Sampling : Using Social Networks to Research Non- heterosexual Women,” vol. 5579, no. November, 2017.

[28] L. A. Palinkas, S. M. Horwitz, C. A. Green, J. P. Wisdom, N. Duan, and K. Hoagwood, “Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research,” *Adm. Policy Ment. Heal. Ment. Heal. Serv. Res.*, vol. 42, no. 5, pp. 533–544, 2015.

[29] “quota-sampling.” [Online]. Available: https://explorable.com/quota-sampling. [Accessed: 06-Aug-2019].

[30] “Systematic random sampling.” [Online]. Available: https://www.investopedia.com/ask/answers/071615/when-it-better-use-systematic-over-simple-random-sampling.asp.