What are the Major Sampling Methods to Obtain a Representative Sample for a Clinical Study?

Sada N. Dwivedi

Considering the need for genuine generalization of observed findings under a clinical study among the target study population, it is essential to conduct the study on a representative sample from the study population. Along with focusing on the honest following of each of the major steps required under research methodology for a clinical study, it becomes essential to use an appropriate sampling method to ensure the representativeness of the study sample. In other words, the study sample should ideally consist of major characteristics of the study population, especially those expected to influence study findings. For this, the focus of the present write-up is to briefly describe major sampling methods so that researchers can easily be aware of them and choose one of the feasible and appropriate sampling methods for their clinical study.

Access this article online

Website:
www.cijmr.com

DOI:
10.58999/cijmr.v3i01.159

Keywords:
Study Population, Representative Sample, Sampling Methods, Research Methodology, Clinical Study

Introduction

To guide public health and/or clinical practice based on derived authentic and reliable evidence in clinical research, generalizing observed findings among the target population is required as a first step in this direction. In this regard, the researcher of a particular study is supposed to randomly draw a required sample from the study population to appropriately answer a planned research question.[1] A researcher may or may not be in a position to use the most appropriate sampling method to ensure better representativeness of drawn sample for the study population, but he needs to be familiar with each of the various sampling approaches so that in case of any deviation from appropriate sampling method due to feasibility, he understands the related limitation of the study sample and its implications while analyzing the data, interpreting the analytical results and deriving conclusions including generalizations of the results. Keeping in view of this, all the major sampling methods for a random/ non-random selection of samples from the study population have been briefly discussed in this write-up. Keeping in view feasibility and other issues, this will provide a long way for any researcher to get well acquainted with them along with their relative merits and demerits and facilitate decision-making regarding the involvement of sampling method under a clinical study.

As a matter of fact, a complete enumeration of a study population is rarely possible in a clinical study. In other words, to carry out a research study, one often needs to select a representative sample from the study population[2, 3] using appropriate and feasible random sampling method (s).[4] They are briefly described in the successive sections.

Sampling Methods

The sampling methods are broadly classified as random sampling (with known knowledge about probability/chance of selection of a sampling unit in the sample) and non-random sampling (without any knowledge about probability/chance of selection of a sampling unit in the sample) methods.[4] For a better understanding of sampling methods (random and non-random), one needs to be well acquainted with basic terminologies, which are briefly stated as follows:

Study Population

This is basically an aggregate of all the eligible individuals/patients in a target setting among whom research findings need to be applicable/generalized. For example, to study the prevalence of respiratory problems in the Ujjain district of Madhya Pradesh (MP), the study population will be total general population of Ujjain

All India Institute of Medical Science, New Delhi, Delhi, India.

Correspondence to: Sada N. Dwivedi, All India Institute of Medical Science, New Delhi, Delhi, India. E-mail: dwivedi7@hotmail.com

Submitted: 06/02/2024 Revision: 28/02/2024
Accepted: 04/03/2024 Published: 20/04/2024

How to cite this article: Dwivedi SN. What are the Major Sampling Methods to Obtain a Representative Sample for a Clinical Study?. Central India Journal of Medical Research. 2024;3(1):6-11.

district; to study prevalence of respiratory problems among patients attending RD Gardi Medical College, Ujjain, study population will be total patients attending RD Gardi Medical College; and to study prevalence of chronic obstructive pulmonary disease (COPD) among symptomatic respiratory disease people attending medical OPD at RD Gardi Medical College, study population will be total symptomatic respiratory disease people attending medical OPD at RD Gardi Medical College. If sampling units differ from individuals, aggregate of considered sampling units in a target setting will constitute the study population.

Sampling Unit

Basically, the considered unit for sampling is known as a sampling unit. For example, sampling at the individual level may not always be feasible. Sampling is often considered at the household level, village/block, district, state, region, and country. Sometimes, sampling is carried out at the school level. Further, sampling may also be carried out at the hospitals, on weekdays, with physicians, and so on.

Sampling Frame

A listing of all the sampling units in a study population is referred as a sampling frame. For example, for individual-level sampling, the sampling frame will consist of listing each eligible individual in a study population. A listing of each eligible household will be a related sampling frame for household-level sampling. Likewise, a list of schools will be known as the sampling frame to consider school-level sampling.

Representative Sample

A representative sample of a study population has its all potential characteristics, especially those expected to influence study findings. For example, age and gender are potential confounders in almost all clinical studies. Hence, the selected sample should also have an almost similar geriatric population if the study population is 10% (above 60 years of age). Further, the selected sample ought to have a comparable level of gender distribution to that of the study population. Likewise, if a study population is cancer patients, the proportion of patients in each grade in the sample needs to be similar to that in the study population.

Random Sampling Methods

Under these methods, as stated earlier, the probability or chance (equal or unequal) of getting selected is known for each sampling unit. Therefore, sometimes they are also addressed as probability sampling methods. These methods remain to the most valid choice to derive a representative sample from a target study population. The major random sampling methods are briefly described below:

Simple Random Sampling Method

To use this method, a complete sampling frame for the target study population is required. For example, to randomly select 500 families from the Rural Field Practice Area, RD Gardi Medical College, Ujjain, a list of all the families of the Rural Field Practice Area with specific serial numbers will be required. Under this method, the probability/chance of being selected for every family remains equal. However, use of this method may be appropriate only if the study population is homogeneously distributed, especially concerning the health status of interest. In case of the non-availability of sampling frame or heterogeneity among study population, one of the other appropriate sampling methods must be used. To draw a representative sample using simple random sampling, using a random number table often available in almost every book involving statistics/biostatistics/research methodology. It is recommended that another CME write-up be made on a random number table, and its detailed use will be made available later. Traditional methods like tossing a coin, especially to draw large sample size, become non-random.

As an inbuilt feature of the random number table, it can be used from any side/angle with a random start. If the population size consists of four digits (e.g., 5000), one has to consider four columns or rows. Often, the column-wise choice becomes friendlier. After fixing four columns, one can close the eyes and put the pencil/ pen on the table that provides random start (e.g., 0357). From that point one can move down and keep recording four digits' numbers falling in the range of 0001 to 5000 in sequence until a target sample of 500 families is selected. No further change in the derived list of 500 families is required for a simple random sampling with replacement. Under this approach, a family may get selected more than once. On the contrary, under simple random sampling without replacement, a family will not be considered again once selected. It is a preferred valid method to be used.

The use of simple random sampling is very popular to select a representative sample from the study population. It is evident from the fact that all the methods to explore minimum sample size for a particular study presume that simple random sampling will be used. Otherwise, an appropriate design effect is required to adjust the explored sample size. For example, while using cluster sampling (described later) in place of simple random sampling method, explored sample size under a study to estimate burden of any health problem had to be doubled. As stated earlier, design effect varies from one study design to another.

Systematic Random Sampling

There are situations when the sampling frame of the study population/selected higher-order sampling units are unavailable. To overcome this problem, one can simply make use of systematic random sampling. For example, suppose in a selected village, there are 150 families (as told by the village leader) and only 12 families need to be selected, especially in the absence of a sampling frame of the village. In that case, a division of 150 by 12 will give a sampling fraction of approximately 13. One can randomly plan the selection process from any direction of the village. Out of first 13 families, for a random start, one has to choose first family randomly To find a random number between 01 to 13, instead of using a random number table in this regard, one can simply pull out a currency note from the pocket and consider last two digits (e.g., 11). From family 11 onwards, every 13th family in the village may be selected, (i.e., 11, 24, 37, 50, 63 and so on) to obtain the required random sample of 12 families. If there is no random start in selecting the first sampling unit, this method will be systematic sampling but not systematic random sampling.

To use systematic random sampling, one has to make sure that families in the village do not involve any unknown systematic pattern that can make the drawn sample skewed. For instance, if families in almost each of the considered groups of 13 families are in order of increasing socioeconomic status, selected families might be predominantly from high socioeconomic status. Also, families of low socioeconomic status might be present inappropriately in the sample. In summary, the sample may not represent the study population. It may obviously result into distortion in research findings, leading to inappropriate implications.

Cluster Random Sampling:

There are situations when in the first instance, one has to randomly select higher-level sampling units (e.g., villages in rural area, wards in urban area; and schools), often addressed as clusters, instead of individuals. Further, in case of no demarcation of such sampling units (e.g., wards in a city), they may be considered jointly and divided into various segments (i.e., generated clusters) using landmarks like lanes/roads. For instance, for a long term epidemiological study to assess the effect of toxic gas exposure in the night of 2nd/3rd December 1984 on human health in Bhopal, use of cluster sampling was unavoidable.[5] Likewise, for many of the school-level studies, first of all schools are randomly selected instead of individual students. To use this sampling method, clusters need to be of similar characteristics. Otherwise, the drawn sample may not be representative of the study population, and accordingly, the obtained results may neither be valid nor generalizable to the considered study population.

Probability Proportional to Size Sampling

In case of sampling using higher order sampling unit (e.g., village) involving its varying sizes, if size is known for its each individual unit (i.e., village), the probability proportional to size (PPS) sampling method is recommended.[6] To be more specific, there are villages (i.e., clusters) involving different number of households in each village. In this case, each village may be selected with a probability that is proportional to the number of households in that village. As an example, if there are four villages having 20, 40, 60 and 80 households respectively, the probability of their selection in the sample will be 1/10 (20/200), 2/10 (40/200), 3/10 (60/200) and 4/10 (80/200), respectively. In other words, a bigger cluster will have a higher chance of selection. The PPS sampling is known to provide unbiased estimates of the parameters under study. However, villages need to be selected with replacement sampling to avoid complexity under this method. This sampling method is often integrated within a large-scale community survey, such as various rounds of national family health surveys in India[2] and other national/ international studies.

Stratified Random Sampling

Sometimes public health indices (e.g., specific morbidity indices; infant mortality rate) are likely to vary from one area to another area (e.g., villages, wards, blocks, districts, states) at a higher extent, mainly due to varying extents of potential confounders. In such situations, a study on such indices through a sample using simple random sampling method from study population may not provide valid results. Under such circumstances, to make a drawn sample from a study population its representative, study populations need to be divided into different strata considering potential confounders and a sample needs to be drawn proportionately from each stratum. For example, villages may be divided into various strata like those with public health centers, within one kilometer of public health centers; and those beyond 1 kilometer. As another example, sample of a toxic gas-exposed population in Bhopal, MP, (due to toxic gag exposure from leakage of gases from Union Carbide Plant) had to be drawn from mildly exposed area, moderately exposed area and severely exposed area[5]. This sampling approach is known as stratified random sampling if the considered sampling unit is selected randomly. As obvious, if used properly, this sampling method helps in obtaining better representative sample leading to more precise and valid results.

Multistage Random Sampling

The large-scale surveys often involve multistage sampling, that is, sampling at two or higher stages. This is mainly to ensure a representative sample that also reduces the cost and time involved in the study. Under this method, samples are taken from higher sampling units to smallest sampling units. Further, to ensure random sampling, use of an appropriate and feasible random selection method at every stage is necessary. For instance, under various rounds of National Family Health Survey (NFHS) including its recently completed fifth round (NFHS-5)2, a two-stage sampling was used in rural and urban areas. Specifically, in rural areas, villages were randomly selected as primary sampling units (PSU) using PPS random sampling at the first stage; households within selected PSUs were randomly selected as secondary sampling units using systematic random sampling at second stage. Likewise, in urban area, Census Enumeration Blocks (CEB) was randomly selected at first stage; and households within selected CEB were randomly selected at second stage. Likewise, similar sampling design was also used in the national household drug abuse survey in India[3]. Further, collected data using this sampling design has a hierarchical structure once variables at various levels are recorded.[7]

Multiphase Random Sampling

This sampling design involves data collection on larger sample size initially (i.e., phase 1) and on its sub-sample later (i.e., phase 2). Likewise, the phases involved in a clinical study may be three or higher. This design helps in conducting a study at a reduced cost by minimizing the cost of data collection through costly clinical procedures. As the best example of multiphase sampling, after toxic gas exposure from the union carbide plant in Bhopal during 2nd /3rd December 1984 night, a long-term epidemiological study[5] was started through the registration of a cohort of people from mildly exposed area, moderately exposed area, severely exposed area and also from unexposed area. Detailed information on exposure along with socioeconomic, demographic, vital statistics and symptomatic morbidities was collected at first instance. Considering various categories using data on symptomatic morbidities in each area, sub-samples were randomly drawn for respiratory studies for one of the various studies. Again, from samples under respiratory study, sub-samples were randomly drawn for radiological study. Likewise, from samples under radiological study, sub-samples were randomly drawn for pathological/ immunological study. As evident under this sampling design, a sub-group of registered cohort under each area has data on each aspect starting from epidemiological, respiratory, radiological, and pathological/ immunological data. This data could be used for linkage analysis among various public health/ clinical aspects along with important clues for generalisation of the analytical results. Further, this sampling design could also help reduce required time and cost of doing all tests on complete registered cohort. It may be worthwhile to mention here that one should be able to clearly differentiate between multiphase and multistage sampling designs, they are not same.

Interpenetrating Random Sampling

To control the quality of data, this method involves drawing two or more random samples of the required minimum sample size independently from the same study population. This design uses all three basic principles of experimental design: randomization, replication and local control. Intuitively, randomization controls the systematic error, whereas replication and local control minimize the random error in the data set. This sampling design may help assess associated factors with different sources of variation, such as enumerators/ field workers, submitted schedules, methods of data collection, and data management/analytical methods. In other words, interpenetrating sub-samples may help quantify the listed non-sampling errors. It may also help quantify the sampling error using first sampling units and correct related bias. However, under this sampling design, sub-samples may or may not overlap one another more so in case of multistage sampling. This method is conventionally used in national-level studies carried out in India by the National Sample Survey Organisation (NSSO), especially under sample registration system.[8] As a limitation of this design, study becomes costly due to replicated samples in comparison to a single sample.

Interpenetrating sampling is sometimes also referred as interpenetrating subsampling and replicated sampling. Basically, this sampling method acts as a major ingredient under the following major resampling methods[9]:

- *Jackknife Resampling Method*

Under any advanced epidemiological modeling in public health/ clinical research, its cross-validation often relies on a Jackknife resampling method. Under this resampling method, for a study involving a size "n" sample, repeated subsamples of size (n-1) are obtained by randomly omitting one record. Further, the bias and variance estimated for each subsample are aggregated to obtain their Jackknife estimates. In addition to a point estimate, it also helps in obtaining an interval estimate (confidence interval).

- *Bootstrap Resampling Method*

This is another resampling method used in cross-validation of developed epidemiological models under public health/ clinical research. Under this method, from a considered sample, repeated random samples of same size are collected through replacement method. Like the case of Jackknife method, the bias and variance estimated for each sample are aggregated to obtain their Bootstrap estimates. It also helps in obtaining interval estimates (confidence intervals) along with their point estimates.

Non-Random Sampling Methods

Under these methods, as stated earlier, the probability or chance (equal or unequal) of getting selected is not known for each sampling unit. Therefore, sometimes they are also addressed as non-probability sampling methods. These methods remain inferior/ invalid in determining a representative sample from a target study population. Hence, compared to random sampling methods, these methods involve a high risk of sampling bias. They have limited scope for inferences/ conclusions. Accordingly, these methods are often used in qualitative research and/ or exploratory studies to understand newer and/ or unexplored areas of research. The major non-random sampling methods[4] are briefly described below:

Purposive Sampling Method

This method is also known as the judgment sampling method. Under this method, selection of a sample totally relies on the investigator who he thinks may be a better choice regarding the planned research. This method generally derives detailed knowledge about the planned topic instead of inferences. As an example of using this sampling method, to know opinions about various aspects of the quality of education in a college, one may purposefully select a number of students from various academic programs of the college. However, to use this method, it is better to have clarity in related inclusion and exclusion criteria. Also, expected observer bias on the part of the investigator using this method needs to be noted.

Convenient Sampling Method

As evident from its name, a study using this method includes only those individual sampling unit (e.g., individuals, families, villages), which can be easily accessible to the investigator. As an example of using this sampling method, one may conveniently ask students of only those classes in which he teaches to know opinions about various aspects of the quality of education in a college. As obvious, the derived sample may not be representative of the study population and is likely to suffer from both sampling and selection bias. Hence, the findings may not be generalizable.

Quota Sampling

Under quota sampling, as indicated by its name, a pre-decided number of sampling units is also selected without involving random selection. In other words, under this method, the investigator controls the sample size. As such, it controls the structure of the study sample. For example, to know the average health expenditure by people in a community, one may take a pre decided sub-sample from various community strata like children, adults and geriatrics groups. Hence, the sample may not represent the study population due to both sampling and selection bias.

Voluntary Response Sampling

Under this method, contrary to respondents being selected by the investigator for conventional direct contact under all other sampling methods, respondents volunteer themselves. In other words, a voluntary response is obtained from the respondents at their convenience. For instance, online opinion survey about various aspects of the quality of education in a college. Intuitively, such a sample often involves self-selection bias; respondents may be of a peculiar characteristic. For example, most of them might support a particular aspect of quality education, leading to biased responses.

Snow Ball Sampling

This methodology is used in a situation when it is difficult to identify a sampling unit of a particular study population. Also, it can be used if it is difficult to identify persons involved with a sensitive phenomenon. For instance, this method may be useful to conduct a study related to drug addicts instead of using house to house survey. Under this approach, once a drug addict consents to be part of the study, he may help in meeting with other drug addicts whom he knows in the study area.[10] Obviously, this method will remain to be non-random sampling and also drawn sample may be influenced by sampling bias.

Conclusion

Sampling under clinical research is a process to select a sample, a subgroup of the considered study population. The methods of sampling broadly consist of two types: probability sampling and non-probability sampling. They are also known as random sampling methods and non-random sampling methods, respectively. Contrary to non-random sampling methods, every sampling unit has a known chance of being included in the sample for a quantitative study under random sampling methods. It ensures the representativeness of the drawn sample to the study population. Each of the various random sampling methods described above has its own relative merits and demerits. In light of these factors and feasibility, a researcher may choose one of the random sampling methods for a quantitative study. In other words, the random sampling method used may vary from one study to another. On the other hand, non-random sampling methods are often used in qualitative research. Like random sampling methods, they may also vary from study to study.

References

1. Dwivedi SN. How to Formulate a Research Question, Hypothesis and Objective for a Clinical Study?. Central India Journal of Medical Research. 2023;2(3):3-7.
2. International Institute for Population Sciences (IIPS) and ICF. 2021. National Family Health Survey (NFHS-5), 2019-2021: India: Volume 1. Mumbai: IIPS.
3. Dwivedi SN, Pal H, Srivastava A, Pandey A, Nath J. National household survey on drug abuse (NHSDA) in India: Methodological Appraisal. INT J CURR SCI. 2014, 11(1-9).
4. Sundaram KR, Dwivedi SN, Sreenivas V. Medical Statistics: Principles and Methods. Wolters and Kluwer (Health). New Delhi. 2015 (Second Edition).
5. Sriramachari S. Health Effects of the Toxic Gas Leak from the Union Carbide Methyl Isocyanate Plant in Bhopal- Technical report on Population Based Long Term Epidemiological Studies (1985-1994) at Bhopal Gas Disaster Research Centre, Bhopal, MP, under Indian Council of Medical Research, Ansari Nagar, New Delhi.
6. WHO/ UNICEF/ ICCIDD. Assessment of iodine deficiency disorders and monitoring their elimination. A guide for programme managers. Second Edition. Geneva, World Health Organization, 2001. (Document No. WHO/NHD/01.1).
7. Dwivedi SN, Sundaram KR. Epidemiological models and related simulation results for understanding of contraceptive adoption in India. International J Epidemiology. 2000; 29:300-307.
8. Sample Registration System Reports. Office of the Registrar General and Census Commissioner India, Ministry of Home Affairs, Govt. Of India, New Delhi.
9. Wu CFJ. Jackknife, Bootstrap and other Resampling Methods in Regression Analysis. The Annals of Statistics. 1986; 14(4): 1261-1295.
10. Siddiqui NA, Rabidas VN, Sinha SK, Verma RB, Pandey K, Singh VP, Ranjan A , Topno RK , Lal CS , Kumar V , Sahoo GC , Sridhar S , Pandey Arvind. Snowball Vs. House-to-House Technique for Measuring Annual Incidence of Kala-azar in the Higher Endemic Blocks of Bihar, India: A Comparison. PLoS Negl Trop Dis. 2016; 10(9): e0004970. https://doi.org/10.1371/journal.pntd.0004970