International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2016): 79.57 | Impact Factor (2015): 6.391

Statistical Sampling – A Primer for Beginners

Deepti Srivastava

Indian Statistical Service (2000 batch), Director, Ministry of Skill Development & Entrepreneurship, India

Abstract: ***This article provides beginners an insight on selection of sampling methodology for any study particularly in social sector studies. It also highlights some common errors which need to be avoided for having a robust and reliable study.***

Keywords: Simple Random Sampling, Precision, Stratification, Snowball Sampling, Estimates, Haphazard

1. Introduction

An appropriate sampling methodology is an important ingredient for right results/interpretations. It has often been observed that newspapers/media articles quote studies conducted on some topic of interest with some exciting findings. The authenticity of any study largely depends on its sampling methodology including sample size. It is a well known fact that a wrong sampling methodology may lead to extremely contrary results and errors. In social sector studies, it becomes more pertinent to adopt an appropriate sampling methodology for understanding impact of certain programmes or schemes or its evaluation.

2. Requisite Steps for Adopting Sampling Strategy

As such there is no foolproof method to decide the best sampling strategies but some cautions can be taken to minimize errors.

2.1 Defining the objectives

Before jumping into the selection of sampling methodology the objectives/purpose of the study should be well defined. The final expected outcome from the study should be clear in researcher's mind before deep diving. This objective should also make clear that whether estimates are required or only statement about the sample selected will be sufficient. For example if in a study of soil quality of certain area whether one wants the quality of soil details like a presence of nitrogen, phosphorous, nutrients, moistness etc. OR one really wants to estimate yield from the soil of that particular area. Simply identifying lucid objectives will throw light on the answering following two questions;

1) Whether one wants to conduct complete enumeration or wants to go ahead with sample survey?
2) Whether study is being conducted for calculating estimates or simple mean/median/mode is needed.

a. Complete enumeration vs sample survey

The complete enumeration or census requires a lot of time and resources if sampling frame is large while sample survey saves not only time but leads to more precise information in lesser resources. To illustrate, take the example of estimating number of mentally ill patient in a village. One way is to do house listing of the whole village and visit door to door for collecting information. The other way is to use snowball sampling where a knowledgeable person or health worker of the village may be contacted and he/she will provide some information about mental patients in particular households. Thereafter, these households will provide information about other similar cases in the vicinity. It has been found that purposive sampling is more effective than the census in the cases where information is not easily accessible or it is related to some negative indicator. However, the researcher who is applying purposive should be capable enough to identify right resources and possible pitfalls otherwise there may be some bias in the findings. Further, if population size is very large it is practically difficult to do census or complete enumeration. Sample surveys in such cases are convenient and provide more precise estimates.

b. Estimates vs averages: For taking a decision on sampling design, one needs to know that whether estimates are needed or not. If one needs to get an estimate after completing the study, a probability sampling[3] will be needed otherwise one may go ahead with purposive sampling[4].

2.2 Determination of Population to be studied

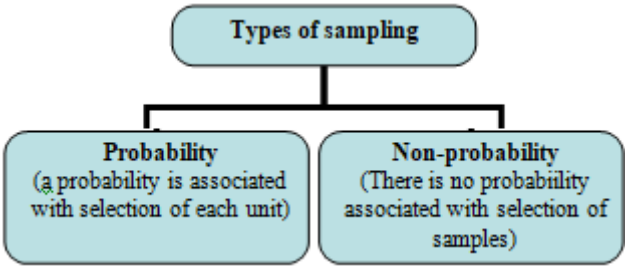
The population from where the sample is to be drawn or on which study is to be conducted should be defined in clear and unambiguous terms. The geographical, administrative and other boundaries should be clearly specified before initiating the study.

2.3 Freezing Sampling Frame

Sampling frame is the list of units in the population for which data is to be collected. For example to estimate drop outs girls of 10 to 16 year age group in a village, the list of all the girls of the same age group is needed. A sampling frame is a key feature around which the selection and estimation revolve. Now it is to be noted that in case of purposive sampling, we may go ahead without sampling frame also or in other words if one does not have the complete listing of units (like in case of mental health patients/dropouts), one may go ahead with purposive sampling.

2.4 Selection of proper sampling design

The reliability of the estimates depends upon the selection of sampling design. There are mainly two types of sampling;



3. Probability Sampling

Most frequently used probability sampling are described as following;

3.1 Simple Random Sampling (SRS): In SRS method an equal probability of selection is assigned to each available unit of population. Other methods of sampling are often preferred to Simple random sampling on the grounds of convenience or increased precision. Following are the methods of selecting a random sample;

a. Lottery method
b. Use of random number tables
c. Remainder approach
d. Quotient approach
e. Independent choice of digits

Random vs Haphazard – A Random selection is used loosely by many researchers who get confused between random and haphazard. Random as explained above is a very organized way of sample selection where each unit needs to be numbered/listed and there should be a complete frame while haphazard is a manner where a user can pick any unit without having the complete idea of the frame or population.

3.2 Stratified Random Sampling (StRS)

In StRS, the population of N units is divided into n subpopulations called strata. The sub populations are non-overlapping and comprise the whole population i.e. $N_1+N_2+\ldots+N_n = N$. Stratification provides an opportunity to divide the heterogeneous population into homogenous subpopulations and estimates with greater precision. In StRS, the allocation of the sample to different strata is done on the basis of following;

a. Stratum size i.e. no. of units in the stratum
b. the variability within the stratum
c. the cost of taking in taking observations per sampling unit in the stratum

A good allocation is where maximum precision is obtained with minimum resources. There are following four methods of allocation of samples to different strata;

a) **Equal allocation** – for administrative convenience, the equal number of samples are allocated to different strata
b) **Proportional Allocation** – In this sampling fraction is same in all strata. Numerous estimates can be made with greater speed and a higher degree of precision.
c) **Neyman allocation** – Also called "minimum variance allocation". Sample allocation in this depends on stratum size and stratum variance. This method is difficult to use as it is not easy to get stratum variance in all the cases.
d) **Optimum allocation** – In this allocation of samples are done with an objective of minimizing variance for a specified cost of conducting a survey.

3.3 Systematic Random Sampling (SyRS)

In SyRS, the first unit is selected with the help of random numbers and the rest get selected automatically according to some predefined pattern. It is popular for its simplicity and Operational convenience. It is a commonly used technique if a complete and up-to-date sampling frame is available. It is useful in forest surveys in estimating the volume of timber, in fisheries for estimating total catch of fish, estimation of lactation yield etc. It is of mainly two types;

a) **Linear Systematic Sampling**: In this, the population is linearly ordered. N=nk (where N is the total number of units in the population, n is the no. of units to be selected as sample and k is an integer). If the first unit selected is i, i+k, i+2k …..i+(n-1)k gets selected in the sample.
b) **Circular Systematic Sampling**: If N≠nk, first unit i.e. i is selected randomly and then every kth unit is selected till n samples are obtained.

It can not be used when a population has an unforeseen periodic bias as this may contribute to bias in the estimates. Another drawback is in SyRS, sampling variance is estimated with single sample

4. Non-Probability Sampling

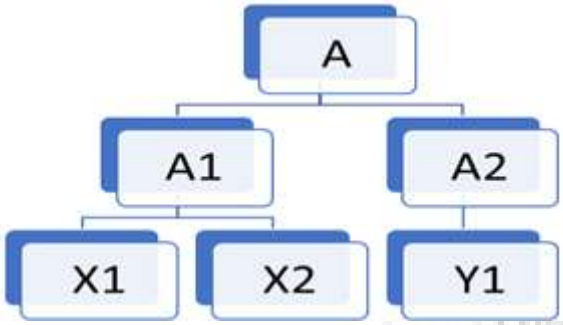
In this, the choice of selection of sample entirely depends on the judgment of the sampler. This method is also called **purposive or judgment** sampling. In this, the sampler inspects the whole population and selects a sampling which he considers best or most feasible. If the degree of precision of estimates is not expected to be made and the sampler is experienced and well versed with the characteristics of the population, purposive sampling is most effective. In some cases of measuring a negative indicator from the sample like maternal mortality, the number of mental patient in the population or number of households having AIDS patient, no. of dropouts in a village etc., the purposive sampling like the snowball is most effective sampling methodology.

Snowball Sampling: It is a kind of purposive (nonprobability) sampling. The name snowball is there because as the ball rolls down, it gets thicker and thicker like a snowball. It is an approach for locating key informants that are needed for any particular study/research. For illustration, in case of no. of dropouts in the children of 6-14 age group, the children can be identified easily through snowball sampling, the process for selecting 10 households with such dropouts will be as follows:

Step 1: 2/3 knowledgeable persons/members of VEC/ teachers/ SHGs will be contacted to know about the locations of the households having dropout children of 6-14 age group. Address of at least one such household will also be obtained from them.

Step 2: Identified households will be contacted and it will be confirmed that they belong to the required category (i.e. having dropped out children of 6-14 age group) or not. If the household belongs to the required category, the household schedule will be canvassed. Further, it will be tried to locate some other households of the required category from the members of this household and the process will go on.

Step 3: As and when the 10 distinct households will be identified, the process will be stopped whenever requisite number of sampling.



5. Advantages of Statistical Sampling

Sampling is a very useful statistical tool which can be used for gap studies, determination of baselines, concurrent and impact evaluation of projects particularly in social sector studies. It provides a platform for assessing the feature of a particular geography or population with lesser resources and time. The accuracy of the sampling procedures, if chosen judiciously, is generally high and results are statistically acceptable. The only caution is to apply due diligence to reduce bias.

References

[1] William G. Chochran, "Sampling Techniques", third Edition, John Wiley & Sons, 1999
[2] P.V. Sukhatme & B.V. Sukhatme, "Sampling theory of surveys with applications", second edition, IOWA State University Press and Indian Society of Agricultural Stattistics, New Delhi
[3] Parimal Mukhopadhyay, "Theory and methods of survey sampling", Prenctice Hall of India Private Limited, 1998.

Author Profile

Deepti Srivastava is M.Sc.(Statistics) from University of Allahabad, India. She belongs to the 2000 batch of the prestigious Indian Statistical Service and has experience of working in various organizations of the Government of India (GOI), viz. Ministry of Water Resources, Planning Commission of India and Ministry of Health & Family Welfare. Her fields of interest are Monitoring & Evaluation and Project Management. Currently, she is posted as Director in Ministry of Skill Development & Entrepreneurship, GOI.