Journal of Advanced Research in Medical Science & Technology
Volume 7, Issue 1 - 2020, Pg. No. 17-21
Peer Reviewed & Open Access Journal

Review Article

Sampling in Research Series 1: Basic Concepts in Estimating Sample Size

SC Mohapatra[1], Sunil Kumar Chamola[2]

[1]Former HOD, Community Medicine, BHU, Varanasi and Former Dean FMHS and Dean Academic affairs SGT University.
[2]Assistant Professor, Community Medicine, FMHS, SGT University, Gurgaon.

DOI: https://doi.org/10.24321/2394.6539.202003

INFO

Corresponding Author:
SC Mohapatra, Community Medicine, BHU, Varanasi and Former Dean FMHS and Dean Academic affairs SGT University or Presently Advisor Cum Consultant, SGT University, Gurgaon
E-mail Id:
vishwamegh@gmail.com
Orcid Id:
https://orcid.org/0000-0002-9605-0867

How to cite this article:
Mohapatra SC, Chamola SK. Sampling in Research Series 1: Basic Concepts in Estimating Sample Size. *J Adv Res Med Sci Tech* 2020; 7(1): 17-21.

Date of Submission: 2020-04-22
Date of Acceptance: 2020-05-14

ABSTRACT

Sampling is a process/ method of drawing a representative group of units or cases from a particular population. Sampling is based on probability theory.It is very important for each researcher in all fields to understand and select a statistically and epidemiologically valid sample; which has become extinct from the cognitive arena of researcher today.Some people take a thumb rule of 100 samples. If asked why 100, why not 99 or 101; usually there is no answer.If there is no valid response to “why this number of sample”, then the study is to be discarded.Even many journals today publish papers without proper sample size being considered in the study. In fact such journals should also be made to stop publishing. Thus it was imperative to highlight the value of sample size calculation as CME. Sampling based on probability theory determines result of chance.The methodology used to select a sample from a larger population includes special procedure and calculations that needs to be followed.

Keywords: Sampling, Sample Size, Sampling Methods, Sampling Error, Probability & Non-Probability Sample

Introduction

Quality assurance in epidemiological and managerial studies or studies related to Health System Research (HSR)or any kind of study, survey or research methodology, sampling selection of a subset of individuals from within a statistical population to estimate characteristics of the whole population is Vital and unfortunately not followed in most of the theses or researches today. One can observe irrational sampling studies being published also in many journals some also call them as International journal erroneously. Unfortunately, the researchers do not understand the advantages of sampling such as lower cost, faster data collection as well as quality assurance of a study without which it loses the title as a “Study”. Therefore, in most of the cases in daily life, business and industry the information are gathered by means of *proper Sampling method with adequate size* for accepting the study. The results of a properly taken sample enable the investigator to arrive at generalisations that are valid for the entire population. The process of generalising sample results to the population is called *Statistical Inference*.

Basic Terminology

Before discussing the sampling theory and distribution of a statistic, we shall be discussing basic definitions of the important terms which are very helpful to understand the fundamentals of statistical inference.

Population

In general sense “population means a large group of people who live in a particular geographical area. For example, the group of people who live in Uttarakhand, the group of people working in a college, students enrolled in the MBBS programme in any Medical college etc. In statistics, population is a collection or group of individuals/ items/ units/ observations under study. For example, the collection of books in a library, the particles in a salt bag, the rivers in India, the students in a classroom etc. are considered as populations in Statistics. The population is classified into two types as given below.

Finite and Infinite Population

A countable number of units or observations seen in a population, is called a “finite population”, say 1.3 crore Indians,30 patients of Gall bladder cancer…and so on. Even non-living thing studied can also be named as “finite population” such as books in the library, number of trees in Tarai or Trains of India etc.

On the other hand, if a population contains uncountable units or observations, it is entitled as“infinite population”. The grains of Bengal grams in a bag and the stars on the sky (as compared to finite Moon in the sky) etc. are “infinite population”**.**

Often times people take a sample of 100. When asked why 100 why not 99 or 101. There is usually no answer to this question. Sample size or type cannot be taken from the air, it must be derived from statistics.

Normal Distribution

As Per the Normal Distribution curve, the distribution of observations follows a rule approximately 68% of all observations fall within±1 SD (standard deviation) of the mean. Approximately 95% of all observations fall within ± 2 SD of the mean. Approximately 99.7% of all observations fall within±3 SD of the mean. Figure number one depicts this concept.

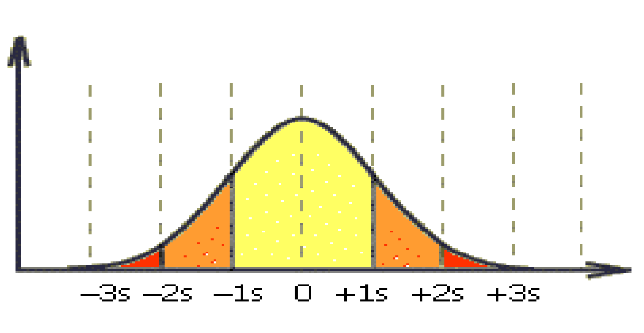

Figure 1.Role of standard deviation in selecting observations in normal circumstances

Sample

To extract the information from all the elements or units or items of a large population may be, in general, time-consuming, expensive and difficult. In this situation, a small part of the population is selected from the population which is called a sample. Thus, the sample can be defined as: “A part/ fraction/ subset of the total population.”

Statistical Errors

The inaccuracies or errors in any statistical investigation, i.e., in the collection, processing, analysis and interpretation of the data may be broadly classified as follows:

- Sampling errors
- Non-sampling errors

Sampling Errors

In a sample survey, since only a small portion of the population is studied, its results are bound to differ from the census results and thus have a certain amount of error. This error would always be there, no matter that the sample is drawn at random and is highly representative. This error is attributed to fluctuations of sampling and is called sampling error. *Sampling error due to the fact that only a subset of the population has been used to estimate the population parameters and draw inferences about the population.* Thus sampling error is present only in a samples survey and is completely absent in census method.

Sampling errors are primarily due to the following reasons:

- Faulty selection of the sample
- Substitution
- Faulty demarcation of the sample units
- Error due to bias in the estimation method
- Variability of the population

Non-Sampling Errors

Non-sampling errors are not mistakes attributed to chance and are a consequence of certain factors which are which were otherwise within human control. In other words, they are due to some human errors during the study at some stage or other and are possible to be traced but created due to ignorance or circumvention. This could be undertake at any stages such asplanning, execution, management or evaluation of the study, data collection or analysis or even report writing. Someimportant factors/ situationsof non-sampling errors are given below:

- Faulty planning of the study
- Faulty definitions of disease or situation. A good example is the definition of diarrhoea in under-five. It’s now as per UNICF, “ If the mother says diarrhoea in under-five, it is diarrhoea”, while most people take diarrhoea as per frequency or constituents of the stool of the child. If the definition is not used properly entire data collected will be lost. Same as the definition of ORS since the homemade sugar-salt solution is not ORS as per the belief of many!
- Incomplete sampling frame

ISSN: 2394-6539
DOI: https://doi.org/10.24321/2394.6539.202003

- A vague and imperfect questionnaire which might result in the incomplete or wrong information
- Defective methods of interviewing and asking questions.
- The personal bias of the investigator
- Failure of respondents, memory to recall the events or happenings in the past
- Lack of trained and qualified investigators and lack of supervisory staff
- Improper coverage
- Publication errors

Types of Sampling

There are two major types of sampling i.e. probability and non-probability sampling,

Probability Sampling

Probability sampling is a scientific technique of drawing samples from the population following the role of chance for each unit pre-assigned to be included in the sample with a Chance otherwise known as "probability". In such a sample, each sample has an equal probability of being chosen in the sample lot.

Non-Probability Sampling

There are five types of non-probability sampling technique that can be taken as adequate. However one has to consult an Epidemiologist who can advise the type of non-probability sample that can be chosen for the study. They could be convenience sampling, purposive sampling and snowball sampling, etc. Type of snowball sampling in-hospital study and community settings differ in methodology.

Sample Size Calculation

For quantitative analysis we have a given formula to calculate the appropriate sample size:

N=$\frac{Z^2\sigma^2}{D^2}$ Where Z=confidence interval, σ=standard deviation; D= marginal error

For Qualitative Analysis

N=$\frac{Z^2pq}{D^2}$ Where z=confidence interval, p=prevalence rate, q=1-p and D=precision error.

Example 1: A study is to be performed to determine a certain parameter in a community. From a study a standard deviation of 46 was obtained. If a sample error of up to 5 is to be accepted. How many subjects should be included in this study at 95% level of confidence?

Solution: Here we have given sd=34

Error was 5% Z=1.96 at 95% ,N=$\frac{Z^2\sigma^2}{D^2}$, N=$\frac{1.96^2\times34^2}{5^2}$ = 177.63 ~ 178 samples are to be considered.

Example 2: It was desired to estimate the proportion of anaemic children in a certain preparatory school. In a similar study at another school, a proportion of 28 % was detected. Compute the minimal sample size required at a confidence limit of 95% and accepting a difference of up to 4% of the true population. For the matter of Error, it cannot be more tha 10% either. *Then*

Solution"

N=$\frac{Z^2pq}{D^2}$ or $\frac{4pq}{D^2}$ = N=$\frac{1.96^2\times.28\times.72}{(.04)^2}$=484.04 ~ 484 or its usually taken as $4pq\,/\,L^2$where p=Prevalence; q=!00 or 1 - Prevalence; and L = permissible limit of error. However permissible limit of error should maximum be within 10%, a higher value of which always damages the quality of the study . If Error is taken as 5%, it means 5% of the prevalence ie 0.04. Taking the absolute value of 5 is incorrect and cannot be taken since that introduces Sampling Error. Today many wise investigators make mistake at this point by taking Absolute value of the error. Let us examine in the same example if the absolute value of 4 instead of 0.04 the sample size will become $\frac{1.96^2\times.28\times.72}{(4)^2}$=0.032 as the sample size in place previous sample as *484.* So one has to be careful in calculating the Sample Size.

Example 3: The sample size of immunization coverage of under-fives was needed to be studied where the previous coverage showed 80% with absolute precision of 4% of the value at 95% confidence interval.

Solution: The sample size in this situation is calculated by the formula: n=Z^2.p.(1-p)/d^2

Where d+ Absolute precision,=0.04;P= Expected proportion in the population=0.80 and Z=1.96

Thus n=$(1.96)^2$.(0.80).(0.02)/$(0.04)^2$ = 384

Example 4: What should be the sample size of pregnant women who seek care at health centre within 1st trimester?

Solution: Percentage of women seeking care as per previous study=P=25%

Relative Precisione= 5% of 25%=1.25, Z=1.96 & confidence level=95%

Thus n=z^2.(1-P) /e^2P)=4610 women

Example 5: The prevalence of Alcoholism in a village was seen to be 25%. The villagers were given health education in different modalities and style for one year. How many people should be studied to see the reduction in alcoholism, is it is desired to be 90% sure of detecting a rate of 20% at 5% level of significance?

Solution: Earlier study rate= (P_0)=25%, Anticipated rate=(P_1)=20%, Level of significance needed (α)=5%, Power of test (1-β)= 90%

Thus n={(1-α)[P_0(1-P_0) +Z (1-β).$P_{\alpha(1-P\alpha}$)}$^{2\,(P}$$_0$-$P_\alpha$)^2= 601

The usefulness of the Pilot Study

This sampling technique helps to gain initial primary data about a topic. These findings can serve as pointers and help the investigator to decide the further line of action. This is more important for evaluating a probable prevalence of a diseases/condition if it is not known. But pilot study data and the sample size calculated thereof are not valuable for universal use and remains specific for the particular area of study. But a number of pilot studies taken at different situations can be further authenticated by a Meta-analysis of the multiple observations even if taken by pilot studies.

Census Method

In the census method, we resort to 100% inspection and enumeration of each and every unit of the population. The census method seems to provide the more accurate and exact information as compared to sample enumeration as the information. Moreover, it affords a more extensive and detailed study. For instance, the population census conducted by the Government of India every ten years collects information not only about the population but also obtains data relating to age, marital status, occupation, religion, education etc.

When to Use Census Method?

Census method is recommended in the following situations:

1. If the information is required about each and every unit of the population, there is no way but to resort to 100% enumeration.
2. In any manufacturing process in industry, 100% enumeration should be taken recourse to under the following condition:

a. The occurrence of a defect may cause loss of life or serious casualty to personnel.
b. A defect may cause serious malfunction of the equipment.

Quantitative Research-Sample Size:

When conducting probability sampling it is important to use a sample size that is appropriate to the aims and objectives of the research.

A common error is to assume that the sample should be a certain percentage of the population, for example, 10%. In reality, there is no such relationship and it only the size of the sample that is important.

A probability sample size of 100+ is considered a large enough sample to conduct Statistical analysis.

Statistics and Samples

1. When presenting your research you need to be able to demonstrate, how representative of the whole population the sample data you have collected is.
2. There is two statistical tests used to do this:

- Standard Error
- Confidence Levels

Standard Error

- Using the standard deviation of the population and the sample size a statistical calculation can measure the degree of error likely to occur between the results of a sample and the results of a census, this is call the standard error?
- The larger the sample the lower the standard error.
- When a probability sample of 100+ is undertaken the distribution can usually be assumed to be normal
- When the sample has normal distribution, we can use the z score approach to obtain confidence limits for the sample mean.

Standard Error of Mean

The square root of the variance of the sample means:

SE of sample mean = SD/ √n

SE of sample proportion = $\frac{\sqrt{pq}}{n}$

Application of Standard Error of Mean:

- To determine whether a sample is drawn from the same population or not when its mean is known.
- To work out the limits of desired confidence within which the population mean should lie.

Confidence Interval or Fiducial Limits:

Confidence limits are two extremes of measurements within which 95% of observations would lie.

- Lower confidence limit=mean – ($t_{0.05}$ × SEM)
- Upper confidence limit=mean + ($t_{0.05}$ × SEM)

The important difference between ‘p’ value and confidence interval is confidence interval represents clinical significance and ‘p’ value indicates statistical significance.

Errors of Non-Observation

1. Sampling Error
2. Errors of coverage
3. Non-response errors due to:

- The inability of the person responding to come up with the answer
- Refusal to answer
- Inability to contact the sampled elements

These errors can be classified as due to the interviewer, respondent, instrument, or method of data collection.

Interviewers: Interviewers can affect the quality of study directly.

- Non-nutral attitude of the interviewer
- Giving lead answer while interviewing
- Gender, racial, and ethnic bias

Respondents: Respondents need to answer correctly and honestly.Some times sensitive questions lead to inbuilt respondents' bias. Basic errors/ bias are:

- Recall bias: When the respondent cannot remember when did it happen.
- Prestige bias: exaggerates to 'look' better
- Intentional deception: lying
- Incorrect measurement: does not understand the units or definition

Type I Error

The probability of finding a difference with our sample compared to population, and there really isn't one Known as the (or "type 1 error") Usually set at 5% (or 0.05)

Type 2 Error

The probability of not finding a difference that actually exists between our sample compared to the population Known as the β (or "type 2 error"). Power is (1- β) and is usually 80%.

Confidence Levels

- Confidence levels are calculated using the Central Limit Theorem
- Using this and the sampling error we can then use the area below the normal distribution curve to make predictions about our sample
- As well as making predictions we can use the properties of the normal distribution curve to provide us with confidence levels
- The confidence level or the confidence interval is that place where your H_0 is accepted
- The concept does not mean that we are 95% sure that a single sample mean lies within these limits
- There are three confidence levels 68%, 95% and 99%
- 95% confidence is considered acceptable in social research, medical research often requires 99% confidence, particularly in Drug Trials

There will be a continuation of the present CME by the authors titled 'Sampling in Research series 2: estimating sample sizes in an especial situation' in our next issue.

Conflict of Intrest: None

References

1. Bartlett JE, Kotrlik JW, Higgins C. Organizational research: Determining appropriate sample size for survey research" (PDF). *Information Technology, Learning, and Performance Journal* 2001; 19(1): 43-50.
2. Kish L. Survey Sampling. Wiley. ISBN 978-0-471-48900-9. 1965.
3. Scott S. Determining Sample Size: How to Ensure You Get the Correct Sample Size. Qualtrics. Retrieved 19 September 2018. 2013.
4. Israel, Glenn D. Determining Sample Size. University of Florida, PEOD-6. Retrieved 29 June 2019. 1992.
5. Rens van de Schoot, MilicaMiočević (eds.). Small Sample Size Solutions (Open Access): A Guide for Applied Researchers and Practitioners. Routledge. 2020.