Section: Community Medicine

Original Research Article

STATISTICAL APPROACHES FOR SAMPLE SIZE ESTIMATION IN CLINICAL RESEARCH

Shivam Dixit[1], Shivani Jain [2], Sana Afrin[3], Abhay Singh[4], Sanjay Silawat[5]

[1]*Assistant Professor (Statistician), Department of Community Medicine, MGM Medical College, Indore, Madhya Pradesh, India.*
[2]*Senior Resident, Department of Community Medicine, SAMC & PGI, Indore, Madhya Pradesh, India.*
[3]*Assistant Professor, Department of Community Medicine, SAMC & PGI, Indore, Madhya Pradesh, India.*
[4]*Associate Professor, Department of Community Medicine, AIIMS, Raebareli, Uttar Pradesh, India.*
[5]*Tutor, Department of Community Medicine, MGM Medical College, Indore, Madhya Pradesh, India.*

Received : 05/06/2025
Received in revised form : 16/07/2025
Accepted : 04/08/2025

Corresponding Author:
Dr. Shivani Jain,
Senior Resident, Department of Community Medicine, SAMC & PGI, Indore, Madhya Pradesh, India..
Email: dr.jainshivani80@gmail.com

DOI: 10.70034/ijmedph.2025.3.431

Source of Support: Nil,
Conflict of Interest: None declared

Int J Med Pub Health
2025; 15 (3); 2336-2339

ABSTRACT

A sample is a part of a population, which can be selected in different manners, from which the required data is collected. It is not possible to examine or enquire each and every one among the millions of people. It is the sample that is considered to be representative of whole population under study. Estimation of an adequate sample size is therefore pre-requisite before carrying out any study, and is one of the basic and critical steps while planning a research. Each and every study design requires a different method for the calculation of sample size. The aim here is to describe the various methods in a comprehensive manner and under a single umbrella for the ease of understanding.

Key-words: Sample size, Power, Margin of Error.

INTRODUCTION

"What should be the sample size for my study?" During the course of a research study, investigators frequently ask this question.

A sample is a part of a population, which can be selected in different manners, from which the required data is collected. Since it is not possible that each and every one among the millions of people be examined or enquired, it is the sample that is considered to be representative of the whole population under study.[1]

Estimation of an adequate sample size is therefore pre-requisite before carrying out any study, and is one of the basic and critical steps while planning a research.[2,3]

If the sample is too large or too small, various details may be overlooked and this may lead to false approval of results. Conclusions obtained from the sample are generalized to the whole population and with wrong results in hand, a population may be put to the risk of unnecessary interventions or may not get the necessary intervention at all. This may collectively lead to wastage of the investigator's time, as well as, wastage of resources.[2,4]

It has often been noticed that various published studies and articles do not report the calculation of sample size appropriately.[5,6] Some of them have been found to have deficient number of participants. As stated earlier, this variation leads to difficulty in generalizing the results to the whole of the population.[4]

Hence, it is crucial to acknowledge that a single formula can not calculate the sample size for all the study designs. Each and every study design requires a different method for the calculation of sample size.[7] The aim here is to describe the various methods in a comprehensive manner and under a single umbrella for the ease of understanding.

Components of sample size[4,8]: Following are some basic components that one must be familiar with before calculating the sample size for any study-

Alpha- It is a number denoting the level of significance represented by the Greek letter Alpha. It is exactly equal to the probability of a Type I error. Lesser the value of Alpha, found during computation, lesser are the chances of rejecting a true null hypothesis. Out of all levels of significance, the ones most commonly used for alpha are the values of 0.10, 0.05 and 0.01.

Power- It is the probability of not making a Type II error. It can be calculated as (1 – beta). The power of a hypothesis test lies between 0 and 1. If the power is calculated and found close to 1, we can conclude that the test for hypothesis is very good at detecting a false null hypothesis. The value of Beta is usually set at 0.2, but it can be set by the researchers to be lesser. Accordingly, the power may be computed as low as 0.8, but it can be higher. However, the value of powers lower than 0.8 is generally considered to be too low for most researches.

***Margin of error (E)* -**The margin of error effect size, denoted by 'd' is the minimal difference that the investigator intends to determine between study groups. It is also known as the minimal clinical relevant difference i.e., difference of two means or difference of two proportions. Previously reported data, pilot studies, and educated guesses based on clinical experiences are three methods for estimating the ES. On the basis of clinical experience, the margin of error is usually set at 5%.

***The Variability (SD)* –** A researcher requires to predict the population variance of a specifically given outcome variable which can be calculated using the standard deviation (SD). The investigators frequently use an estimate based on information obtained from previous studies. If no previous study is found, a pilot survey to estimate standard deviation must be conducted.

Prevalence Rate (P) - When estimating sample size, the prevalence rate, or underlying event rate, of the condition being studied in the population is crucial. It is generally calculated from previous studies. If no prior literature is available, a pilot survey must be conducted to assess the prevalence rate.

1. For estimating values by descriptive cross-sectional study: [8,9]

(i) For Quantitative data (mean):

If the standard deviation of the population or 'σ' is known through previous experience or records, the sample size can be calculated using the formula below:

$$n = \frac{Z^2_{1-\frac{\alpha}{2}} * (\sigma)^2}{(E)^2}$$

n = Sample Size

$Z^2_{1-\frac{\alpha}{2}}$ = Confidence level at 95%

σ = Standard deviation

E = Margin of Error

The following are some fundamental rules for estimating sample size:

1. **Power** – It should be above or equal to 80%. As the power increases, so does the sample size. The higher the power, the less likely you are to miss a real effect.

2. **Level of significance** – This is usually set at 5%. As the level of significance decreases, the sample size increases.

Following are the constant values for convention values of α and β:

Z values for conventional values of alpha (a) are:

Alpha	Zα
0.05	1.96
0.01	2.58

Z values for conventional values of beta (β) are:

Beta	Z (1-β)
0.20	0.842
0.10	1.28
0.05	1.64
0.01	2.33

(ii) For Qualitative data (Proportion):

$$n = \frac{Z^2_{1-\frac{\alpha}{2}} * p * q}{(E)^2} * D$$

p = Expected prevalence

q = 100 - p

D = design effect reflecting the sampling design used in the study. It is '1' for Simple Random Sampling. It is taken as '2' for other designs such as Cluster Random Sampling.

(iii) For Finite Population

$$n = \frac{n_0}{1 + \frac{(n_0 - 1)}{N}}$$

Where n is the sample size

n_0 is the calculated sample size

N is the population size

2. For establishing association by case control study or independent sample cohort study: [4,8]

(a) For Quantitative data:

$$n = \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2 \sigma_1 * \sigma_2}{(d)^2 \; or \; (m_1 - m_2)^2}$$

Or

$$n = \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2 2(\sigma)^2}{(d)^2 \; or \; (m_1 - m_2)^2}$$

d = $(m_1 - m_2)$ = Mean difference was observed in the outcome variable

(b) For Qualitative data:

$$n = \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2 (p_1 q_1 + p_2 q_2)}{d^2 \; or \; (p_1 - p_2)^2}$$

Or

$$n = \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2 2pq}{d^2 or\ (p_1 - p_2)^2}$$

p_1 = Event rate in control group

p_2 = Event rate in cases group

d = $(p_1 - p_2)$ = Expected proportion difference between case and control (it may be based on previously published studies), p_1 is the proportion in case and p_2 is the proportion in control.

3. For identifying significant difference between two groups by intervention study (in each group):[4,9]

(a) For Quantitative data: Formula for the Paired t-test

Consider Alpha only:

$$n = \frac{Z^2_{1-\frac{\alpha}{2}} * (\sigma)^2}{(d)^2\ or\ (m_1 - m_2)^2}$$

Consider Alpha and Beta:

$$n = \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2 (\sigma)^2}{(d)^2\ or\ (m_1 - m_2)^2}$$

σ = Standard deviation of the before and after differences

Formula for the unpaired t-test:

Consider Alpha

$$n = \frac{Z^2_{1-\frac{\alpha}{2}} * 2(\sigma)^2}{(d)^2\ or\ (m_1 - m_2)^2}$$

Consider Alpha and Beta:

$$n = \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2 2(\sigma)^2}{(d)^2\ or\ (m_1 - m_2)^2}$$

σ = Standard deviation in response variable

4. Randomized Controlled Trial (Testing Non-inferiority/Non-superiority/Equivalence of Two Treatments): [8,9]

To calculate sample size requirement for testing non-inferiority of means or proportion of an event between two groups (e.g. standard treatment and new treatment).

Non-inferiority of two means

$$n = \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2 [v_1 + v_2]}{[d - (m_1 - m_2)]^2}$$

Non-inferiority of a proportion

$$n = \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2 [p_1)(1 - p_1) + p_2(1 - p_2)]}{[d - (p_1 - p_2)]^2}$$

m_1= Anticipated mean of the variable in standard treatment

m_2
= Anticipated mean of the variable in new treatment

v_1=Anticipated variance of the variable in standard treatment group

v_2=Anticipated variance of the variable in new treatment group

p_1=Anticipated event rate during a fixed period in standard treatment group

p_2=Anticipated event rate during a fixed period in new treatment group

d = Maximum difference between means or proportions within which one group is regarded non-inferior to the other

5. Sample Size in Survival Studies: [4,8,9]

Any two of the following should be known in order to conduct a study with the objective of testing a hazard ratio of the risk of vital events such as mortality in the two groups at some specified points in time:

Expected survival rate in people who were exposed to the factor of interest (or treatment group) = p_1

Expected survival rate in people who were not exposed to the factor of interest (or control group) = p_2

If the expected hazard ratio, that is, the ratio of the risks of a critical event (e.g., death, disease recurrence) in the two groups, does not change over time, it can be calculated as:

Hazard ratio (h) $= \frac{log p_1}{\log p_2}$

Then, the required minimum sample size can be calculated as:

$$n = \frac{N_E}{(2 - p_1 - p_2)}$$

Where, N_E is the minimum number of desired events and may be estimated as:

$$N_E = \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2 (h + 1)}{2(h - 1)}$$

Accordingly, the formula for minimum number of sample in each group can be calculated as:

$$n = \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2 [log p_1 + log p_2]^2}{[log p_1 - log p_2]^2 (2 - p_1 - p_2)}$$

CONCLUSION

As we all know, studying the entire population in any study is not realistic nor feasible, and therefore the sample. Determining the sample size is one of the first practical procedure while designing a research study and is of vital importance. We ought to be familiar with the essential concepts of medical statistics in order to develop a research project and comprehend the contents of a research paper. The calculations can be complicated, and thus, in order to avoid any kind of methodological error, it is best to seek advice from an expert statistician for the same.

REFERENCES

1. Mahajan's Methods in Biostatistics for Medical Students and Research Workers. 9th ed. Jaypee Brothers Medical Publishers (P) Ltd; 2018.

2. Das S, Mitra K, Mandal M. Sample size calculation: Basic principles. Indian J Anaesth. 2016;60(9):652-656. doi:10.4103/0019-5049.190621
3. Gogtay NJ. Principles of sample size calculation. Indian J Ophthalmol. 2010;58(6):517-518. doi:10.4103/0301-4738.71692
4. Gupta KK, Attri JP, Singh A, Kaur H, Kaur G. Basic concepts for sample size calculation: Critical step for any clinical trials!. Saudi J Anaesth 0;0:0.
5. Ranganathan P, Deo V, Pramesh CS. Sample size calculation in clinical research. Perspect Clin Res. 2024;15(3):155-159. doi:10.4103/picr.picr_100_24
6. Althubaiti A. Sample size determination: A practical guide for health researchers. J Gen Fam Med. 2022;24(2):72-78. doi:10.1002/jgf2.600
7. Chander NG. Sample size estimation. J Indian Prosthodont Soc. 2017;17(3):217-218. doi:10.4103/jips.jips_169_17
8. KP S, Srikantiah C. Sample Size estimation and Power analysis for Clinical research studies. Journal of human reproductive sciences. 2012;5(1):7-13.
9. Sakpal TV. Sample size estimation in clinical trial. Perspect Clin Res. 2010 Apr-Jun; 1(2): 67–69