Review Article

Journal of College of Medical Sciences-Nepal, Vol-19, No 2, Apr-Jun 2023
ISSN: 2091-0657 (Print); 2091-0673 (Online) Open Access
DOI: 10.3126/jcmsn.v19i2.23978

Sample Size Estimation in Medical Research Based upon Study Design

Hari Prasad Upadhyay,[1] Bijay Lal Pradhan[2]

[1]Department of Statistics, Birendra Multiple Campus, Bharatpur-10, Chitwan, Nepal, [2]Department of Statistics, Amrit Campus, Kathmandu, Nepal.

ABSTRACT

Introduction

In medical research the target of researcher is to generalize the finding to the population based upon the information of sample data. This sample has to be representative of the target population, and the number of participants must be appropriate. The determination of minimum optimum sample size is extremely important not only for ethical and economic purposes but also to achieve scientifically and statistically sound results and valid conclusion.

Choosing the best study design and calculation of sample size are the most important tool in any research. Before you plan any type of medical research either you should have good knowledge on research methodology or you should meet to professional statistician and ask for help at the time of planning a research project to avoid methodological errors.

Keywords: Case control; cross-sectional; sample size calculation; power and sample Size; study design.

INTRODUCTION

Population refers to the complete set of observations in a specific area. Its size can be finite or infinite, and its composition can be homogeneous or heterogeneous. A representative subset of this population is referred to as a sample, and the number of observations chosen from the population is known as the sample size. The process of selecting a sample from the population is called sampling. The sample size plays a crucial role in two statistical aspects: the precision of estimations and the study's power to draw conclusions. Therefore, determining the appropriate sample size is a vital step in planning any biomedical research. It is important to establish the minimum or optimum required sample size to ensure ethical and cost-effective practices and to obtain scientifically and statistically valid outcomes and conclusions. The primary objective of calculating the sample size in research is to minimize time, manpower, and costs while obtaining reliable statistics. By employing suitable formulas based on the study design and employing appropriate sampling techniques (such as probability sampling), we can obtain a reliable and effective sample size that allows us to generalize our findings to the

Correspondence: Hari Prasad Upadhyay, Department of Statistics, Birendra Multiple Campus, Bharatpur-10, Chitwan Nepal. Email: hpchalise@gmail.com. Phone: +977-9855058464.

population through estimation (confidence interval).[1]

The purpose of this article is to calculate the sample size for different study designs. This article emphasizes various methods of sample size estimation commonly used in medical research especially formulae method used for simple random sampling. Different formulae are described for comparing different outcome measure (means or proportions) by taking medical data. Calculation of exact sample size is an important part of research design. It is very important to understand that different study design need different method of sample size calculation and one formula cannot be used in all designs.[2]

In clinical research, our objective is to draw conclusions about a population by examining a representative sample. The sample must accurately reflect the target population, with an appropriate and suitable number of participants. The sample size should be large enough to minimize the likelihood of chance differences between groups and increase the chances of detecting clinically significant differences. However, it should not be excessively large, leading to resource waste or unnecessary participant risk. Thus, during the study design phase, it is crucial to calculate the sample size carefully. This calculation ensures that the study is properly designed, as an inadequate sample size can yield unreliable results and expose participants to unnecessary risks. Therefore, determining the optimal or minimum required sample size is of utmost importance, not only for ethical and economic considerations but also to ensure scientifically and statistically sound outcomes.[3]

The study can be designed in either a descriptive or analytical manner. Descriptive studies aim to describe the occurrence pattern of a disease in the entire population. For example, they may examine the prevalence of malnutrition in a school, taking into account factors such as age, gender, socioeconomic status, and so on. On the other hand, analytical studies seek to establish a statistical association between two variables through significance testing. Observational studies involve either retrospective observation (case-control studies) or prospective observation (most cohort studies). Experimental studies, on the other hand, assess the effectiveness of a new drug or treatment compared to a conventional method. It is crucial for researchers to understand that the method of calculating sample size varies across different study designs, and a single formula cannot be universally applied to all study designs.[4]

Different type of Study design

A research design is a plan, structure and strategy used in the research. So, it is called the blue print of research. The most commonly used study designs in medical research are:

a. **Cross-sectional:** This is an observational study where both exposure and outcomes are assessed simultaneously. Data is collected only once, without any follow-up, providing a snapshot of the health issues at a specific moment in time, much like a camera.[5] It offers information about the frequency and characteristics of diseases within a population at a particular point in time. These studies are primarily conducted to determine the prevalence (also known as a prevalence study) of acute or chronic conditions, rather than investigating the causes of diseases or the outcomes of interventions.[4] They are not suitable for studying rare diseases. Therefore, cross-sectional studies or surveys are performed to estimate population parameters such as disease prevalence in a community or to

determine the average value of a quantitative variable across a population. Due to the simultaneous measurement of exposure and disease, it may not always be possible to discern whether the exposure precedes or follows the disease.[5]

Descriptive cross sectional study: Descriptive cross-sectional studies involve the characterization of the frequency of a particular health condition in a specific population. This frequency can be determined at a single moment in time (referred to as point prevalence) or over a specific period (known as period prevalence). Period prevalence is necessary when it takes time to gather enough data on a disease within a population. For example, it may be used to determine the proportion of individuals with hypertension among those who have received care at a public health clinic over the course of a year. These prevalence measures are frequently employed in public health research, often without specifying whether the assessment is based on a specific point or period of time.[6]

Analytical cross sectional study: Analytical cross-sectional studies involve collecting data on the prevalence of both exposure and a health outcome in order to compare differences in health outcomes between individuals who are exposed and those who are not. These studies aim to describe the prevalence of a disease or non-disease by starting with a specific population. They are distinct from purely descriptive cross-sectional studies because they compare the proportion of diseased individuals among those who are exposed to the proportion of diseased individuals among those who are not exposed. These studies are particularly valuable when investigating the connection between exposure and the onset of chronic diseases where information on the timing of disease onset is lacking. For instance, they can be used to explore the relationship between diet and arthritis, smoking and chronic bronchitis, or asthma and exposure to air pollution. However, it is important to interpret the results cautiously, considering the potential influence of disease duration on the exposure status.[7]

B. Cohort study design: Prospective cohort (longitudinal study): The investigator initiates the study by identifying the study population and remains with the subjects throughout the duration of the research. In a prospective study, the investigator commences the study simultaneously with the initial assessment of the cohort's exposure status. When proposing a prospective cohort study, the investigator first identifies the specific traits of the target group for examination. Subsequently, the investigator determines the current case status of individuals, exclusively selecting non-cases to be monitored over time. The determination of exposure status takes place at the outset of the study. [8]

Retrospective cohort study (historical cohort; non-concurrent prospective cohort): An investigator accesses a historical roster of all exposed and non exposed persons and then determines their current case/non-case status.[9] The investigator initiates the study when the disease is already established in the cohort of individuals, long after the original measurement of exposure. Doing a retrospective cohort study requires good data on exposure status for both cases and non cases at a designated earlier time point.[8]

c. Case control design: A case-control study is conducted to compare patients with a specific disease or outcome of interest

(cases) to patients without the disease or outcome (controls). This study looks back retrospectively to determine how frequently a particular risk factor is present in each group, aiming to establish a relationship between the risk factor and the disease.[8] It is an observational study that does not involve any intervention or attempt to alter the course of the disease. The main objective is to retrospectively assess the exposure to the risk factor in both the cases and controls. These studies are designed to estimate odds. For instance, this type of study can be employed to examine the serum vitamin D levels in a group of migraine patients (cases) compared to healthy individuals (controls).[6] Controls do not necessarily have to be in good health; including sick individuals may be appropriate if they represent those at risk of developing the disease.[8] The selection of controls should be independent of the exposures of interest and they should come from the same population as the cases.[10] It is important to match the case and control groups by factors such as age, sex, area, socioeconomic status, etc., in a case-control study.[8]

Statistical terminology

Type of error. [11]

The error committed in rejecting, null hypothesis when null hypothesis (Ho) is true is called type first error (α).

The error committed in accepting null hypothesis (Ho) when null hypothesis (Ho) is false is called type II error.

Hypothesis testing		
Test result	Accept Ho	Reject Ho
Ho is true	True result (1-α)	Type I error (α)
Ho is false	Type II error (β)	True result (1-β)

Level of Significance (α)[12]: The determination of the sample size in a study depends on the desired level of error, known as 'α' or Type I error. Type I error refers to the probability of incorrectly claiming a difference in readings when there is actually no difference (a false positive), resulting in the erroneous rejection of the null hypothesis. The Type I error is predetermined, and its maximum acceptable value is referred to as the level of significance. In most academic research studies, the alpha level used to determine the sample size is typically either 0.05 or 0.01. For crucial findings, the study requires higher precision, so the alpha error is set at a lower level. A lower alpha level corresponds to a larger sample size and a more precise study.[11]

α-error	10%	5%	1%
2-sided value	1.645	1.96	2.5758

Power of test (1- β):[12]

Power refers to the likelihood of rejecting the null hypothesis when the alternative hypothesis is true. It gauges a test's capacity to reject the null hypothesis when it should be rejected. When considering a specific significance level, the test's power can be enhanced by having a larger sample size. The minimum acceptable power level is typically set at 80%, meaning there is an eight in ten chance of detecting a difference of the specified effect size.[3] Determining the power of a study is done prior to data collection as it assists in determining the required sample size. Power represents the probability that the test will correctly identify a difference if it exists. In most cases, a study aims for a power of 80%, which translates to a 20% chance of missing a genuine difference. Occasionally, a higher power of 90% is set, resulting in a 10% possibility of false negative results due to β error. Type II error occurs when two variables are wrongly considered equivalent when they are actually different. Power is directly proportional to the sample size in a study, meaning that as the

sample size increases, so does the power.[13]

Power	80%	85%	90%	95%
Value	0.8416	1.0364	1.2816	1.644

Sample size calculation for cross sectional studies/surveys[4]

Cross-sectional studies or surveys are conducted to determine a population parameter, such as the prevalence of a specific disease in a community or the average value of a quantitative variable in a population. The formula for determining sample size varies depending on whether the variable is qualitative or quantitative.[4]

A. For qualitative variable

Suppose an epidemiologist want to know proportion of under five (6 to 59 month) children suffered from chronic malnutrition in a population then this formula can be used to determine the sample size.

$$Samplesize\ (n) = \frac{z^2_{\alpha/2}P(1-P)}{e^2} \text{(For infinite population)}$$

$$Samplesize = \frac{z^2_{\alpha/2}P(1-P)}{e^2 + \frac{z^2_{\alpha/2}P(1-P)}{N}} \text{(For finite population)}$$

(For infinite population)

(For finite population)

Zα = Standard normal variate value, its value is 1.96 at 95%CI, 1.64 at 90% CI and 2.58 at 99% CI

P = Expected value of proportion in population or prevalence of disease, this value needs to be taken from the previous study

e = Allowable error or margin of error which is decided by the researcher for study, generally its value is taken as 5%

Illustration with an example

A Doctor in of Department of Community Medicine wishes to estimate the prevalence of tuberculosis among children under five years of age in its locality. How many children should be included in the sample so that the prevalence may be estimated to within 5 percentage points of the true value with 95% confidence, if it is known that the true rate is unlikely to exceed 15%?

$$Samplesize = \frac{z^2_{\alpha/2}P(1-P)}{e^2} = \frac{1.96 * 1.96 * 0.2 * 0.8}{0.05 * 0.005} = 246$$

So, in the case of cross section study researcher has to choose 246 respondents for the study.

Sample size calculation for two population proportion case.

How large a sample would be required to estimate the proportion of pregnant women in a population who seek prenatal care within the first trimester of pregnancy, to within 5% of the true value with 95% confidence? It is estimated that the proportion of women seeking such care will be between 25% and 40%.

$$Samplesize = \frac{Z^2_{\frac{\alpha}{2}}(p_1 * (1-p_1) + p_2(1-p_2))}{d^2}$$

$$= \frac{1.96 * 1.96(0.25 * 0.75 + 0.40 * 0.60)}{0.05 * 0.05} = 657$$

B. For quantitative variable

If the researcher is interested in knowing the average systolic blood pressure in pediatric age group of Chitwan district at 5% of type of I error and precision of 5 mmHg of either side (more or less than mean systolic BP) and standard deviation, based on previously done studies, is 25 mmHg then formula for sample size calculation will be

$$Samplesize = \frac{z^2_{\alpha/2}\sigma^2}{e^2} \text{ (For infinite population)}$$

$$Samplesize = \frac{z^2_{\alpha/2}\sigma^2}{e^2 + \frac{z^2_{\alpha/2}\sigma^2}{N}} \text{ (For finite population)}$$

Sample size for comparative study: If the objective of research is to compare different independent group then sample size will be calculate as

$$Samplesize(n) = \frac{p_1(1-p_1) + p_2(1-p_2)(z_\alpha + z_\beta)^2}{(p_2 - p_1)^2}$$

Where, b is power of test, the value of at 80% power of test its value is 0.84, is z-score value at level of significance at 95%CI its value is 1.96, proportion exposed of among first group, proportion exposed of among second group.

Illustration with an example

Suppose a researcher wants to calculate a sample size to compare the level of knowledge regarding organ donation among Medical and nursing students. Reviewed showed that the level of knowledge among nursing and medical students are 90% and 71.85% respectively. Then for his/her study sample size will be:

$$Sample\ size\ (n) = \frac{0.718(1-0.718) + 0.9(1-0.91)(1.96+0.84)^2}{(0.91-0.78)^2}$$
$$= 69.99\ (In\ each\ group)$$

The total sample size for each group is 70 and for over all optimum sample size will be 140.

For case control study [4]

Formula for difference in proportions

To find the sample size in case control study for difference in proportion following formula is used:

$$n = \frac{r+1}{r}\frac{\bar{p}(1-\bar{p})\left(Z\beta + Z_{\frac{\alpha}{2}}\right)^2}{(p_1 - p_2)^2}$$

Where r is the ratio between cases to control, is average proportion of exposed=(p1+p2)/2, Z_b is power of test, is level of significance, p_1 proportion exposed of among control group, proportion exposed of among case group.

For, illustration let's take following example

For 80% power, Z_b=0.84, for 95% CI or 5% level of significance level, Z_a=1.96. For equal number of cases and controls, r =1. The proportion exposed in the control group is 20%. To get proportion of cases exposed

$$P_{case\ expose} = \frac{OR * P_{prop\ expose\ in\ control}}{P_{prop\ expose\ in\ control}(OR-1)+1}$$

$$= \frac{2*0.2}{0.2(2-1)+1} = \frac{0.4}{1.2} = 0.33$$

Hence, Average proportion exposed

= (0.33+0.20)/2=.265

Now using formula, the required sample size is

$$n = \frac{1+1}{1}\frac{0.265\ (1-0.265)(0.84+1.96)^2}{(0.33-0.20)^2} = 181$$

Therefore, n=362 (181 cases, 181 controls)

Formula for difference in means

To find the sample size in case control study for difference in mean value following formula is used:

$$n = \frac{r+1}{r}\frac{\sigma^2\left(Z\beta + Z_{\frac{\alpha}{2}}\right)^2}{(d)^2}$$

Where r is the ratio between cases to control, is the Standard Deviation (SD) of the outcome if interest, Z_b is power of test, is level of significance, clinically important effect size, d, researcher wish to detect in the test i.e. the smallest difference in means that it would be clinically meaningful to detect.

For, illustration let's take following example

For 80% power, Z_b=.84, For 95% CI or 5% level of significance level, Z_a=1.96

For equal number of cases and controls, r=1. The standard deviation of the characteristic you are comparing is 10.0

You want to detect a difference in your characteristic of 5.0 (one half standard deviation)

An equal number of cases and controls (r=1)

$$n = \frac{1+1}{1}\frac{10^2(0.84+1.96)^2}{(5)^2} = 63$$

Therefore, n = 126 (63 cases, 63 controls)

For Cohort study [4]

$$n' = \frac{\left[Z\alpha\sqrt{(r+1)\bar{P}\,\bar{Q}} - Z\beta\sqrt{rp_1q_1 + p_2q_2}\right]^2}{r(p_2-p_1)^2}$$

$$Required sample size(n) = \frac{n'}{4}\left[1+\sqrt{1+\frac{2(r+1)}{n'r(p_2-p_1)}}\right]^2$$

Where, for 80% power, Z_b=.84, For 95% CI or 5% level of significance level, Z_a=1.96, For equal number of expose and unexposed group, r=1, p_1 prevalence among unexposed group, p_2 prevalence among exposed group

For, illustration let's take following example

A medical researcher wants to prove that overweight adult have higher risk of diabetes mellitus as compared to normal weight adult. From the review of literature identify the rate of disease among those with or without the risk factors. Review showed that the chance of having diabetes mellitus among overweight was 32% and among normal adult was 7%.

p_1=0.07, q_1=1-0.07=0.93

p_2=0.32, q_2 =1-0.32=0.68

r = ratio of exposed to unexposed generally we chose equal so =1/1=1

$$\bar{P} = \frac{(p_1+p_2)}{r+1} = \frac{(0.07+0.32)}{1+1} :$$

$$= 0.195 \; and \; \bar{Q} = 1 - P = 1 - 0.195 = 0.805$$

Then using formula,

$$n' = \frac{\left[1.96\sqrt{(1+1)0.195*0.805} - 0.84\sqrt{1*0.07*0.93+0.32*0.68}\right]^2}{1*(0.32-0.07)^2} = 38.23$$

$$Required sample size(n) = \frac{38.236}{4}\left[1+\sqrt{1+\frac{2(1+1)}{38.236*1*(0.32-0.07)}}\right]^2 = 46$$

So, he should select 46 overweight adult and 46 normal adult. The optimum sample size is 46+46=92

Randomized Control Trials[15]

A. For quantitative outcome variable

When the outcome variable is continuous then following formula can be used for calculation of sample size

$$Sample size = \frac{\left(Z\beta + Z_{\frac{\alpha}{2}}\right)^2 2\sigma^2}{(\mu 1 - \mu 2)^2}$$

Here, μ1 is the population mean in group I and μ2 is the population mean in II group, σ is the population variance. Where, for 80% power, Z_b=0.84, for 95% CI or 5% level of significance level, Z_a=1.96.

For, illustration let's take following example

A medical researcher is planned to do a RCT to determine the efficacy of new treatment for diabetes mellitus is better than existing treatment. Mean glucose level is taken from review of literature for both existing (Group I is 140 mg/dl) and novel treatment group (125 mg/dl) also population variance expressed as SD is 20 mg/dl. Then sample size is as:

$$Sample size = \frac{(0.842+1.96)^2 2*20^2}{(140-125)^2}$$

$$= 28\ (in\ each\ group)$$

B. For binary outcome variable

When the outcome variable is in binary outcome (two outcome like "yes" or "No" then) following formula can be used for calculation of sample size

$$Samplesize = \frac{\left(Z\beta + Z_{\frac{\alpha}{2}}\right)^2 (p1q1 + p2q2)}{(p1 - p2)^2}$$

Here, p1 is the proportion of the subjects with disease in group I and p2 is the proportion of the subjects with disease in group II. Where, for 80% power, Z_b=0.84, for 95% CI or 5% level of significance level, Z_a=1.96.

For, illustration let's take following example

A medical researcher is planned to do a RCT on Novel drug on DM, let's take blood glucose level in to binary outcome variable like below 125 mg/d. Past review shows that proportion of subject with DM in group I (existing treatment group) is 20% and in novel treatment group (group II) is 10%. With 80% power, Z_b=0.84, for 95% CI or 5% level of significance level, Z_a=1.96 then sample size will be

$$Samplesize = \frac{(0.842 + 1.96)^2 (0.2 * 0.8 + 0.1 * 0.9)}{(0.2 - 0.1)^2}$$

$$= 196\ (in\ each\ group)$$

Sample size calculation for diagnostic accuracy study[15]

Based on Sensitivity

$$TP + FN = \frac{\left(Z_{\frac{\alpha}{2}}\right)^2 Sensitivity(1 - Sensitivity)}{e^2}$$

$$Sample\ size\ (n1) = \frac{TP + FN}{Prevalence\ of\ disease}$$

Based on Specificity

$$FP + TN = \frac{\left(Z_{\frac{\alpha}{2}}\right)^2 Specificity(1 - Specificity)}{e^2}$$

$$Sample\ size\ (n2) = \frac{TP + FN}{1 - Prevalence\ of\ disease}$$

Then effective sample size will be larger among n^1 and n^2

ConclusionS

Determining the Optimum sample size is an essential component in any form of biomedical research. It is not uncommon for studies to lack statistical power and consequently fail to identify treatment effects due to insufficient sample size. When preparing for research, it is essential to conduct a thorough review to gather the necessary information for sample size calculation and selecting an appropriate study design. Numerous sources of literature and software tools are available for sample size estimation. Research serves as a means of addressing problems or uncovering new information in a particular field, and since it involves collaborative efforts, consulting a professional statistician during the research planning phase is advisable to avoid methodological errors.

REFERENCES

1. Eng J. Sample size estimation: how many individuals should be studied? Radiology. 2003;227(2):309-13.
2. Dhulkhed VK, Dhorigol M, Mane R, Gogate V, Dhulkhed P. Basic statistical concepts for sample size estimation. Indian Journal of Anaesthesia. 2008;52(6):788.
3. McCrum-Gardner E. Sample size and power calculations made simple. International Journal of Therapy and Rehabilitation. 2010;17(1):10-4.
4. Charan J, Biswas T. How to calculate sample size for different study designs in medical research? Indian journal of psychological medicine. 2013;35(2):121.
5. Levin KA. Study design III: Cross-sectional studies. Evidence-Based Dentistry. 2006;7(1):24-5.
6. Pine C, Pitts N, Nugent Z. British Association for the Study of Community Dentistry (BASCD) guidance on sampling for surveys of child dental health. A BASCD coordinated dental epidemiology programme quality standard. Community dental health. 1997;14:10-7.
7. Woodward M. Epidemiology: study design and data analysis: Chapman and Hall/CRC; 2013.
8. Park K. Preventive and social medicine. Jabalpur. 23rd ed. Jabalpur,482001(M.P) India: M/s Banarasidas Bhanot; 2015.
9. Euser AM, Zoccali C, Jager KJ, Dekker FW. Cohort studies: prospective versus retrospective. Nephron Clinical Practice. 2009;113(3):c214-c7.
10. Selby JV, Friedman GD, Quesenberry Jr CP, Weiss NS. A case–control study of screening sigmoidoscopy and mortality from colorectal cancer. New England Journal of Medicine. 1992;326(10):653-7.
11. Lieberman MD, Cunningham WA. Type I and Type II error concerns in fMRI research: re-balancing the scale. Social cognitive and affective neuroscience. 2009;4(4):423-8.
12. Newman SC. Biostatistical methods in epidemiology: Wiley Online Library; 2001.
13. Zar JH. Biostatistical analysis: Pearson Education India; 1999.
14. Zodpey SP. Sample size and power analysis in medical research. Indian Journal of Dermatology, Venereology, and Leprology. 2004;70(2):123.
15. Mukhia R. Principles of Health Science Research: Key Concepts and Methods for Health Care Professionals: Mohan Raj Sharma. Journal of KIST Medical College. 2022 Aug 8;4(8):96-8.

Citation: Upadhyay HP, Pradhan BL. Sample Size Estimation in Medical Research Based upon Study Design. JCMS Nepal. 2023; 19(2); 237-45.