Sample Size Estimation in Medical Research

Hari Prasad Upadhaya[1]

[1]Department of Community Medicine- Lecturer (Bio-statistics), College of Medical Sciences, Chitwan

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. **Conclusion:** Choosing the best study design and calculation of sample size are the most important tool in any research. Before any type of medical research is planned, knowledge on research methodology is necessary or help form professional statistician at the time of planning a research project can be taken to avoid methodological errors.

Key words:Case control, Cross-sectional, Sample size calculation, Power and Sample Size, Study design,

INTRODUCTION

The totality of all observation in specific area is called population. Depending upon the number it may be either finite or infinite and based upon the nature of element or observation it may be homogeneous and heterogeneous. The represented part of that population is called sample and number of element (observation) chosen from that population is called sample size; also the process (technique) of choosing sample is called sampling. The size of a sample influences two statistical properties: the precision of our estimates and power of the study to draw conclusions. So, sample size determination is an important step while planning any medical research. The determination of minimum required sample size is extremely important not only for ethical and economic purposes but also to achieve scientifically and statistically sound results and valid conclusion. The objective of calculation of sample size in any research is to minimize time, manpower and cost and to get good statistics. If we take the reliable or effective sample size using suitable formula based upon our study design and choosing by using suitable sampling technique (probability sampling) technique, then we can generalized our finding to the population, through estimation (confidence interval).[1]

In clinical research, our goal is to make an inference regarding something about a population by studying a sample of that population. This sample has to be representative of the target population, and the number of participants must be appropriate. It should be large enough that the probability of finding differences between groups by mere chance is low and that of detecting true, clinically significant differences is high. However, the number of participants should not be so large that resources are wasted or participants are exposed to unnecessary risk. Therefore, in the study design phase, it is essential to perform sample size calculation. To perform this calculation, we should focus of our study design.[2] Inappropriate sample size cannot produce a sound or valid result and expose the participants to unnecessary risk. So, determination of optimum sample size or minimum required sample size is extremely important not only for ethical and economic purposes but also to achieve scientifically and statistically sound results.[3] The study design can be either descriptive type and/or analytical type. The descriptive type study focuses on the entire population describing the pattern of occurrence of disease, for e.g., prevalence of Malnutrition in the school district relative to age, gender, socioeconomic status, etc. Analytical studies, ascertain statistical association between two things by the test of significance. Analytical studies can be either observational type and/or experimental type. In observational type, we simply observe either retrospectively (case-control study) or prospectively (most of cohort study). In experimental studies, we test the efficacy of a new drug or a new treatment with the conventional method. It is very important to understand for the entire researcher that method of sample size calculation is different for different study designs and one blanket formula for sample size calculation cannot be used for all study designs.[4] 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]

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 in which both exposure and outcomes are measured at the same time. Information is collected only once, there is no follow up and provides the snapshot (like a camera) of health problem at a particular point of time.[5] It provides frequency and characteristics of disease in a population at a point of time. These types of study are conducted to access the prevalence (also called prevalence study) of acute or chronic conditions not for the causes of disease or the results of intervention also, not suitable for the study of rare diseases.[6] So, cross sectional studies or cross sectional survey are done to estimate a population parameter like prevalence of some disease in a community or finding the average value of some quantitative variable in a population.[4] Since exposure and disease are measured at the same time, it may not always possible to distinguish whether the exposure precedes or follows the disease. [5]

Descriptive cross sectional study: Descriptive cross-sectional studies simply characterize the prevalence of a health outcome in a specified population. Prevalence can be assessed at either one point in time (point prevalence) or over a defined period of time (period prevalence). Period prevalence is required when it takes time to accumulate sufficient information on a disease in a population, e.g. what proportion of persons served by a public health clinic over a year have hypertension. These prevalence measures are commonly used in public health, often the point or period aspect is not specified.[6]

Analytical cross sectional study: In analytical cross-sectional studies, data on the prevalence of both exposure and a health outcome are obtained for the purpose of comparing health outcome differences between exposed and unexposed. Analytical studies attempt to describe the prevalence of, for example, disease or non-disease by first beginning with a population base. These studies differ from solely descriptive cross-sectional studies in that they compare the proportion of exposed persons who are diseased with the proportion of non-exposed persons who are diseased.[5] These studies are also useful for examining the association between exposure and disease onset for chronic diseases where researchers lack information on time of onset. Examples might include diet and arthritis, smoking and chronic bronchitis, and asthma and exposure to air pollution. Interpretation requires caution regarding potential association of duration of disease with exposure status.[7]

b. **Cohort study design:**

Prospective cohort (concurrent; longitudinal study) - An investigator identifies the study population at the beginning of the study and accompanies the subjects through time. In a prospective study, the investigator begins the study at the same time as the first determination of exposure status of the cohort. [8] When proposing a prospective cohort study, the investigator first identifies the characteristics of the group of people he/she wishes to study. The investigator then determines the present case status of individuals, selecting only non-cases to follow forward in time. Exposure status is determined at the beginning 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 study that compares patients who have a disease or outcome of interest (cases) with patients who do not have the disease or outcome (controls), and looks back retrospectively to compare how frequently the exposure to a risk factor is present in each group to determine the relationship between the risk factor and the disease.[8]

Case control studies are observational because no intervention is attempted and no attempt is made to alter the course of the disease.[10] The goal is to retrospectively determine the exposure to the risk factor of interest from each of the two groups of individuals: cases and controls. These studies are designed to estimate odds.[6] For example this type of study can be used to study the serum vitamin D status in a group of Migraine patients (case) compared with healthy person (control). Controls need not be in good health, inclusion of sick people is sometimes appropriate, as the control group should represent those at risk of becoming a case. Controls should come from the same population as the cases, and their selection should be independent of the exposures of interest.[10] In the case control study the case and control should be match by age, sex, area, socioeconomic status etc. [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).

$$\text{Type first error}(\propto) = \text{prob}(\text{Reject Ho when Ho is true})$$

$$\text{Type first error}(\propto) = \text{prob}(\text{Reject Ho when Ho is true})$$

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

$$\text{Type second error}(\beta) = \text{prob.}\,(\text{Accept Ho when Ho is false})$$

$$\text{Type second error}(\beta) = \text{prob.}\,(\text{Accept Ho when Ho is false})$$

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

Level of Significance (α)[12]: Sample size is determined according to 'α' or Type I error - how much error is allowable in a study. Type I error is the probability of falsely claiming the difference in reading but actually there is no difference (false positive), and the null hypothesis is rejected erroneously. Type I error is fixed in advance, and its upper limit of tolerance is known as level of significance. The alpha level used in determining the sample size in most of the academic research studies are either 0.05 or 0.01. For critical results, the study should be more precise hence α-error is set at lower level. Lower the alpha level, larger is the sample size and more precise will be the study.[11]

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

Power of test (1- β):[12]

Power is the probability of rejecting the null hypothesis when the alternative hypothesis is true. It measures the ability of a test to reject the null hypothesis when it should be rejected. At a given significance level, the power of the test is increased by having a larger sample size. The minimum accepted level is considered to be 80%, which means there is an eight in ten chance of detecting a difference of the specified effect size.[3] Also, Power of the study is determined before collecting the data, as it helps in determining the sample size. Power is the probability that the test will correctly identify the difference if it is there. Usually, most study accepts power of 80% i.e., 20% chance of missing the real difference. Sometimes a larger study power is set at 90% i.e., 10% possibility of false negative results due to β error. Type II error is falsely stating that the two variables are equivalent when they are actually different. Power proportionality increases as the sample size for study increases.[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 cross sectional survey are done to estimate a population parameter like prevalence of some disease in a community or finding the average value of some quantitative variable in a population. Sample size formula for qualitative variable and quantities variable are different.

A. For qualitative variable

$$Sample\ size = \frac{z^2_{\alpha/2}P(1-P)}{e^2} Sample\ size = \frac{z^2_{\alpha/2}P(1-P)}{e^2}$$

(For infinite population)

$$Sample\ size = \frac{z^2_{\alpha/2}P(1-P)}{e^2 + \frac{z^2_{\frac{\alpha}{2}}P(1-P)}{N}}$$

$$Sample\ size = \frac{z^2_{\alpha/2}P(1-P)}{e^2 + \frac{z^2_{\frac{\alpha}{2}}P(1-P)}{N}}$$

(For finite population)

Z=Standard normal variate value, it value is 1.96 at 95%CI, 1.64 at 90% CI and 2.57 at 99% CI

P=Expected value of proportion in population, this value we have to take from the previous study

e= Allowable error or margin of error which is decided by the researcher for study, generally its value is than 10%.

This formula can be used to determine the sample size if an epidemiologist wants to know proportion of under five (6 to 59 month) children suffering from chronic malnutrition.

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%?

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

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%.

$$Sample\ size = \frac{Z^2_{\frac{\alpha}{2}}(p_1 * (1-p_1) + p_2(1-p_2))}{d^2} = \frac{1.96 * 1.96(0.25 * 0.75 + 0.40 * 0.60)}{0.05 * 0.05} = 657$$

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

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

$Sample\ size = \frac{z^2_{\alpha/2}\sigma^2}{e^2 + \frac{z^2_{\alpha/2}\sigma^2}{N}}$ $Sample\ size = \frac{z^2_{\alpha/2}\sigma^2}{e^2 + \frac{z^2_{\alpha/2}\sigma^2}{N}}$ (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

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

Where $_{,\beta}$ is power of test, the value of $z_\beta z_\beta$ at 80% power of test is 0.84, $Z_\alpha Z_\alpha$ is z-score value at $\alpha\alpha$ level of significance at 95%CI its value is 1.96, $p_1\ is\ p_1\ is$ proportion exposed of among first group, $p_2\ is\ p_2\ is$ 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:

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

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, $\bar{p}\bar{p}$ is average proportion of exposed, Z_β is power of test, $Z_{\frac{\alpha}{2}}Z_{\frac{\alpha}{2}}$ is level of significance, $p_1\ is\ p_1\ is$ proportion exposed of among control group, $p_2\ is\ p_2\ is$ proportion exposed of among case group.

For, illustration let's take following example

For 80% power, Z_β=0.84, for 95% CI or 5% level of significance level, Z_α=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 $\bar{p}\bar{p}$= (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, $\sigma\sigma$ is the Standard Deviation (SD) of the outcome if interest, Z_β is power of test, $Z_{\frac{\alpha}{2}}Z_{\frac{\alpha}{2}}$ 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_β=.84, For 95% CI or 5% level of significance level, Z_α=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]

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

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

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

Where, for 80% power, Z_β=.84, For 95% CI or 5% level of significance level, Z_α=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,

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

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

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

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

CONCLUSION

Optimum sample size is an essential component of any type of medical research. It is not uncommon for studies to be underpowered and fail to detect treatment effects due to inadequate sample size.[14] When any research is planned, first of all extensive review should be done to collect the necessary information for calculation of sample size and choosing the suitable study design. There are lots of literatures and software available to calculate the sample size. So, professional guidance from a statistician at the time of planning a research will help in avoiding 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.

Corresponding Address
Hari Prasad Upadhya
Department of Community Medicine
Bharatpur, Chitwan, Nepal.
E-mail Id: hpchalise@gmail.com