European Journal of Molecular & Clinical Medicine

ISSN 2515-8260 Volume 09, Issue 03, 2022

How big should my sample be? A practical approach to sample size calculation in health studies.

Madhusudhana MV[1*], Chandrashekar TN[2], Prashanth HL[1]

[1]Department of community medicine, Shimoga institute of medical sciences, Shimoga-, India.
[2]Department of pathology, Shimoga institute of medical sciences, Shimoga-577201, India.

***Corresponding author**: Dr Madhusudhana MV, Assistant Professor, Department of community medicine, Shimoga institute of medical sciences, Shimoga-577201, India. Phone: +919480329954; Email: madhusudhanarmy@yahoo.com

Abstract:

In any health study, Sample size estimation is important in detecting scientific and statistically significant differences. Therefore, sample size estimation is a critical step in the design of a planned research protocol. "Samples which are too small prove nothing, and samples which are too large can prove anything." For example, using too many participants in a study is expensive and exposes more subjects to the procedure. Similarly, if the study is underpowered, it will be statistically inconclusive and may make the entire exercise a failure. In this article, sample size estimation for survey type of studies, case-control studies, cohort studies, clinical trials, and qualitative research are discussed in detail in a single platform with suitable examples.

Keywords: sample size, health research, study designs, clinical trials.

Introduction:

The very common question that comes to the minds of researchers is How big should my sample be? If the sample size is too small, it may be impossible to make valid and precise generalizations; on the other hand, it is wasteful to study more subjects than required. Also, if the numbers are too large, even a small difference will be drastically significant, and if the sample is too small, they will miss significant difference[1]. Thus, as scientific merit and ethical issues go hand-in-hand, the awareness of the minimum required sample size and appropriate sampling methods are extremely important in achieving scientifically and statistically sound results[2]. Clinical research studies can be classified into surveys, experiments, observational studies, etc. They need to be carefully planned to achieve the objective of the study. It is very important to understand that different study designs need a different method of sample size calculation, and one formula cannot be used in all designs[3].

ISSN 2515-8260 Volume 09, Issue 03, 2022

As the advances in medicine are coming at an accelerating speed, there is a paradigm shift in medical research. Sometime back, the very common tools used were purely cross-sectional studies and other observational studies. Now there are plenty of clinical research tools available that are effective in coming to a conclusion on introducing a new mode of treatment and control of disease or medical condition[4]. In this article, we cover the important issues in calculating appropriate sample sizes for various applied research.

Many articles have been published explaining the methods of calculating sample size, but a lot of confusion still exists. It is very important to understand that method of sample size calculation is different for different study designs. Although various paid and free (online calculators) software provides an easy method of sample size calculation, it is very important to understand the basics of prerequisites needed to calculate the sample size. Let us discuss the factors which affect sample one before proceeding to the actual sample Calculation part.

Power of the study (β):

The power of a statistical test is the probability that it will correctly lead to the rejection of a false null hypothesis. The statistical power is the ability of a test to detect an effect if the effect actually exists. It is the probability that it will result in the conclusion that the phenomenon exists. Power of at least 80% is desirable given the available resources and ethical considerations. Power proportionately increases as the sample size for the study increases[5].

Significance (α) level:

The definition of alpha is the probability of detecting a significant difference when the treatments are equally effective or at risk of false-positive findings. The alpha level used to determine the sample size in most clinical research studies is either 0.05 or 0.01[6].

Precision:

Precision refers to how close estimates from different samples are to each other. For example, the standard error (SE) is a measure of precision. Precision is inversely related to standard error.

ISSN 2515-8260 Volume 09, Issue 03, 2022

When the standard error is small, sample estimates are more precise; sample estimates are less precise when the standard error is large.

Confidence interval (CI):

Confidence intervals consist of a range of values (interval) that act as good estimates of the unknown population parameter. However, in infrequent cases, none of these values may cover the value of the parameter. In general, the formula gives a 95% confidence interval: CI = Estimate ± 2×SE. Thus, a 95% confidence interval reflects a significance level of 0.05.

Variance or standard deviation:

The variance or standard deviation for sample size calculation is obtained either from previous studies or a pilot study. The larger the standard deviation, the larger is the sample size required in a study.

Effect size:

Effect size measures the magnitude of a treatment effect. To estimate the necessary sample size, we need to know the effect size in advance. There are two strategies available to know the effect size before we conduct the study. One approach is to use another data set to predict the likely effect size. For example, one may conduct a small pilot study to obtain a rough estimate. Alternatively, we can use the results from a related study by conducting literature research on a similar topic. A second approach is to use clinical judgment to specify the smallest effect size that the researcher considers relevant [7].

Sample size estimation for cross-sectional studies (surveys):

Cross-sectional studies are conducted to estimate a population parameter like the prevalence of some disease in a community or finding the average value of some quantitative variable in a

ISSN 2515-8260 Volume 09, Issue 03, 2022

population. The formula for sample size estimation for qualitative variable and quantitative variable is different [8].

For qualitative data, we use the following formula,

$$\text{Sample size } (N) = \frac{(Z_{1-\frac{\alpha}{2}})^2 P(1-P)}{d^2}$$

Here,

$Z_{1-\frac{\alpha}{2}}$= standard normal variate (at 5% it is 1.96)

P= anticipated population proportion

d = absolute error (decided by the researcher)

For example, if the anticipated prevalence of Hypertension in a population is 10%, with a precision / absolute error at 5% and type 1 error at 5 % (95% confidence, Z=1.96)
N=1.9620.1(0.9)0.052 = 138

For quantitative data,

If the study's objective is to find out the average blood pressure of the same study participants, at 5% of the type of 1 error and precision of 5 mmHg of either side (more or less than mean systolic BP). Standard deviation, based on previously done studies, is 25 mmHg, then the formula for sample size calculation will be,

$$\text{Sample size}(N) = \frac{(Z_{1-\frac{\alpha}{2}})^2 SD^2}{d^2}$$

$Z_{1-\frac{\alpha}{2}}$= standard normal variate (at 5%it is 1.96)

SD = standard deviation

d = absolute error (decided by the researcher)

$$N = \frac{(1.96)^2 25^2}{25^2} = 96$$

ISSN 2515-8260 Volume 09, Issue 03, 2022

Sample size estimation for case-control studies:

In case-control studies, the group with the disease in question (cases) is compared with those with no disease (controls) regarding exposure to the risk factors. The formula to calculate sample size depends upon the type of variable, whether qualitative or quantitative.

For qualitative variable:

$$\text{Sample size} = \left(\frac{r+1}{r}\right) \frac{(\overline{P})(1-\overline{P})\,(Z_{\beta}+Z_{\frac{\alpha}{2}})^2}{(P_1-P_2)^2}$$

Here,

r = ratio of control to cases.

$$\overline{P} = \frac{P_1 + P_2}{2}$$

Z_{β}=standard normal variate (for 80%power Z_{β} will be 0.84)

$Z_{\frac{\alpha}{2}}$= standard normal variate (for 5%type Z_{α} will be 1.96)

P_1=proportion of smoker in cases.

P_2=proportion of smokers in controls.

P_1-P_2= Effect size.

ISSN 2515-8260 Volume 09, Issue 03, 2022

Example: If a researcher wants to study the association between lung cancer and smoking, the patients with lung cancer form the cases, and healthy people will be the controls. Then the researcher will go back in time to find out the exposure to tobacco smoking in both groups. Finally, the odds ratio will be calculated for both groups. Here we take a hypothetical example where expected proportions (smokers) in cases and control are 0.40(40%) and 0.20(20%) respectively, with an equal number in both the groups at 80% power, the sample size per group will be,

$$N = \left(\frac{1+1}{1}\right)\frac{(0.3)(0.84+1.96)^2}{(0.4-0.2)^2} = 82$$

So the study involves 82 cases and 82 controls.

For quantitative variable:

When quantitative variables like blood pressure and body weight are measured, we use the following formula,

$$N = \left(\frac{r+1}{r}\right)\frac{SD^2\,(Z_\beta + Z_{\frac{\alpha}{2}})^2}{\delta^2}$$

r = ratio of control to cases

SD = standard deviation based upon previous data.

δ = expected mean difference between cases and controls.

Zβ, Zα/2 = already discussed for the qualitative variable.

Sample size calculation for cohort studies:

In Cohort studies, healthy individuals with or without exposure to a particular risk factor will be observed over time to see the development of the event in both the groups[10]. Let us take a hypothetical example of smoking and lung cancer; if the researcher wants to know whether smoking causes lung cancer, he selects a group of smokers and others who don't smoke. Both groups will be followed up for a specified period, and at the end of the study, the groups will be compared for the development of lung cancer. For such studies, the formula for sample size will be,

ISSN 2515-8260 Volume 09, Issue 03, 2022

$$N = \frac{\left[Z_{\propto}\sqrt{\left(1+\frac{1}{\eta}\right)\overline{P}(1-\overline{P})} + Z_{\beta}\sqrt{P_1} \right]^2}{(P_1 - P_2)^2}$$

Here,

Z_β, $Z_{\alpha/2}$ = already discussed

η= no of controls per experiment group

P_1=probability of the event in the control group

P_2= probability of an event in the experiment group.

$$\overline{P} = \frac{P_2 + \eta P_1}{\eta + 1}$$

Suppose if the researcher wants to see the impact of smoking on lung cancer, according to previous studies, the proportion of lung cancer in smokers is 40% and 20% among nonsmokers, sample size at 5%significance and 80% power with an equal number in both groups. On substitution of the values, the sample size works out to be 169 in each group.

Sample size calculation in clinical trials:

Planning the sample size for a clinical trial requires prior information. The type of prior information depends on the statistical methods which are to be used. For example, if the desired parameters cannot be estimated, it may be desirable to perform a pilot study in advance to estimate the appropriate population parameters[11].

The following formula can be used to calculate the sample size.

For qualitative variable:

ISSN 2515-8260 Volume 09, Issue 03, 2022

Sample size per group$= \frac{2SD^2\left(Z_{\frac{\alpha}{2}} + Z_\beta\right)^2}{\delta^2}$

Here,

SD = estimated standard deviation based upon previous studies or pilot study

δ = expected difference between 2 groups considered as clinically significant (as decided by the researcher)

Zα = standard normal variate for level of significance (1.96 for 5% significance)

Zβ = standard variate for power (0.84 for 80% power)

For example, if the researcher wishes to estimate the number of subjects in each group to detect a difference of 0.25 between mean levels of hemoglobin with 2 sided significance level of 5% and power of 80%. The estimated standard deviation of hemoglobin between the two groups was 0.51 units based on previous studies. In this case, the formula will be,

$N = \frac{2(0.51)(0.84+1.96)^2}{0.25^2}$= 66. The study involves 66 subjects in each group.

For quantitative variable:

In this scenario number of subjects needed in each group will be estimated by the formula

$$N = \frac{2\left(Z_{\frac{\alpha}{2}} + Z_\beta\right)^2 P(1-P)}{(P_1 - P_2)^2}$$

For example, if the researcher wants to estimate the number of subjects in each group to detect that the new treatment is more effective than the previous treatment. The existing drug has a success rate of 30%, and the new drug is expected to reach 60% using a two-sided significance level of 5% and a power of 80%.

Here,

ISSN 2515-8260 Volume 09, Issue 03, 2022

The P1= success rate of present drug=30/100=30%

P2 = success rate of new drug = 60/100=60%

Zα and Zβ will be the same as already discussed.

δ = P1-P2 = 30%

P = Joint success rate = Total number of success/total number of subjects =

90/200= 45%

1-P = Joint failure rate= 1- joint success rate =100-45=55%

On substitution of the values sample size works out to be 43 in each group.

Sample size in qualitative research:

For, e.g., Knowledge, Attitude, and Practice (KAP) studies, the sample size in qualitative research will depend upon the expected response, e.g., the proportion of couples using family planning methods. Based on the expected response, the usual method for estimating sample size can be employed. However, the general assessment of KAP cannot be performed based on a single parameter. If we use an approach based on proportions, we need to calculate the sample size for each parameter separately. In such situations, usually, a score is assigned to the correct response to an item. Thus, a total score for all the correct responses of each individual member is obtained. The total score can then be treated as a continuous or a dichotomous response for analysis[11].

Discussion:

Sample size determination is an important step in the design of a research study. Appropriately sized samples are essential to infer with confidence that samples estimated are reflective of underlying population parameters. The two important factors to be noted in this article are. First, we assume that the selection of individuals is random and unbiased, and second, the studies in which mean is calculated, the measurements are assumed to have normal distributions.

ISSN 2515-8260 Volume 09, Issue 03, 2022

The concept of statistical power is more associated with sample size; the power of the study increases with an increase in sample size. Ideally, the minimum power of a study required is 80%. The ideal study for the researcher is one in which the power is high. As a general rule, higher power is achieved by increasing the sample size[6]. As a result of this, the readers are requested to refer to the standard table to find a normal variate for the appropriate level of significance and power.

As a standard, any scientific research considers a significance level of 5% and power of 80% necessary to draw sound scientific conclusions. The researcher can opt for a higher level of significance and power for a better validity of the results. In studies involving control groups like in case-control studies, usually case to control ratio will be kept the same(one control for a case), but if the researcher wishes to provide a more valid basis for generalization, he can use more than two control groups[12].

This article was designed to make researchers understand the basics of prerequisites needed for calculating sample size. However, readers are also requested to refer to other articles and from other sources also and once they have understood the basics, they can use different online software available for sample size calculations.

References:

1. Sackett DL. Bias in analytical research. Journal of chronic diseases 1979.32:51.
2. Nyi Nyi Naing. Determination of Sample Size. Malays J Med Sci. Jul 2003; 10(2): 84–86.
3. Abramson JH. AbramsonZH. Survey Methods in Community Medicine. Epidemiological research programs clinical trials. In Churchill Livingstone; 1999; pp 15-37.
4. Wittes J. Sample size calculations for randomized controlled trials. Epidemiol Rev. 2002; 24:39–53.
5. Lwanga SK, Lemeshow S. Geneva: World Health Organization; 1991. Sample size determination in health studies. A Practical manual; pp. 1–3.
6. Mohar D, Dulbarg CS. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA.1994; 272:122–4.

7. Effect size. Al therapy statistics Beta (Internet). Updated 2014 (cited on 2014 Dec 25). Available from https://www.ai-therapy.com/psychology-statistics/effect-size-calculator#sizes.
8. Jaykaran C, Tamoghna B.How to calculate sample size for different study designs in medical research? Indian journal of psychological medicine. 2013; 35 (2): 121-126.
9. Chadha VK Sample size determination in health studies NTI Bulletin 2006; 42(3&4): 55–62.
10. Ivo F, Donna S.Power and Sample Size Calculations for Case-Control Studies of Gene-Environment Interactions with a Polytomous Exposure Variable. American journal of epidemiology. 1997; 146 (7): 596-604.
11. Gupta SK. Principles of clinical research and methodology. In: Jaypee publishers. India 2009; 246-50.
12. Hennessy S, Bilker WB, Berlin JA, Strom BL. Factors influencing the optimal control-to-case ratio in matched case-control studies. Am J Epidemiol. 1999 Jan 15; 149(2):195-7.