Review Article

How to Calculate Sample Size for Different Study Designs in Medical Research?

Jaykaran Charan, Tamoghna Biswas[1]

ABSTRACT

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. In this short review we tried to educate researcher regarding various method of sample size calculation available for different study designs. In this review sample size calculation for most frequently used study designs are mentioned. For genetic and microbiological studies readers are requested to read other sources.

Key words: *Medical research, sample size, study designs*

INTRODUCTION

In the recent era of evidence-based medicine, biomedical statistics has come under increased scrutiny. Evidence is as good as the research it is based on, which in turn depends on the statistical soundness of the claims it make. One of the important issues faced by a biomedical researcher during the design phase of the study is sample size calculation. Various studies published in Indian and International journals revealed that sample size calculations are not reported properly in the published articles. Many of the studies published in these journals have less than required sample size and hence less power.[1-3] Many articles have been published in existing literature explaining the methods of calculation of sample size but still a lot of confusion exists.[4-6] It is very important to understand 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. In this article different formulae of sample size calculations are explained based on study designs. Readers are advised to understand the basics of prerequisites needed for calculation of sample size calculation through this article and from other sources also and once they have understood the basics they can use different paid/freely available software available for sample size calculations. For simple study designs formulae given in this article can be used for sample size calculation.

Sample size calculation for cross sectional studies/surveys

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.

For qualitative variable

Suppose an epidemiologist want to know proportion of children who are hypertensive in a population then this formula should be used as proportion is a qualitative variable.

Department of Pharmacology, Govt. Medical College, Surat, Gujarat, [1]Independent Researcher, Kolkata, West Bengal, India

Address for correspondence: Dr. Jaykaran Charan
Department of Pharmacology, Govt. Medical College, Surat, Gujarat, India. E-mail: drjaykaran@yahoo.co.in

$$\text{Sample size} = \frac{Z_{1-\alpha/2}^{2}\, p(1-p)}{d^{2}}$$

Here

$Z_{1-\alpha/2}$ = Is standard normal variate (at 5% type 1 error ($P<0.05$) it is 1.96 and at 1% type 1 error ($P<0.01$) it is 2.58). As in majority of studies P values are considered significant below 0.05 hence 1.96 is used in formula.
p = Expected proportion in population based on previous studies or pilot studies.
d = Absolute error or precision – Has to be decided by researcher.

For example, let us assume that a researcher wants to estimate proportion of patients having hypertension in paediatric age group in a city. According to previously published studies actual number of hypertensives may not be more than 15%. The researcher wants to calculate this sample size with the precision/absolute error of 5% and at type 1 error of 5%. So if we use the above formula

$$\text{Sample size} = \frac{1.96^{2} \times 0.15(1-0.15)}{0.05^{2}} = 196$$

So for this cross sectional study researcher has to take at least 196 subjects. If the researcher want to increase the error (decrease the precision) then denominator will increase and hence sample size will decrease.

For quantitative variable
Suppose the same researcher is interested in knowing average systolic blood pressure of children of the same city then below mentioned formula should be used as blood pressure is a quantitative variable.

$$\text{Sample size} = \frac{Z_{1-\alpha/2}^{2}\, SD^{2}}{d^{2}}$$

$Z_{1-\alpha/2}$ = Is standard normal variate as mentioned in previous section.
SD = Standard deviation of variable. Value of standard deviation can be taken from previously done study or through pilot study.
d = Absolute error or precision as mentioned in previous section

So if the researcher is interested in knowing the average systolic blood pressure in pediatric age group of that city at 5% of type of 1 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

$$\text{Sample size} = \frac{1.96^{2}(25)^{2}}{5^{2}} = 96$$

So researcher will have to take the blood pressure of at least 96 children to know average systolic blood pressure in paediatric age group.

Sample size calculation for case control studies
In case control studies cases (the group with disease/condition under consideration) are compared with controls (the group without disease/condition under consideration) regarding exposure to the risk factor under question.

The formula for sample size calculation for this design also depends on the type of variable (qualitative or quantitative). Here formula for independent case control study is mentioned. To read these formulae in more detail other texts should be referred.[7,8]

For qualitative variable
Suppose a researcher want to see the link between childhood sexual abuses with psychiatric disorder in adulthood. He will take a sample of adult persons with psychiatric disorder and will take another sample of normal adults having no psychiatric disorders. He will then go retrospectively to see history of childhood sexual abuse in both groups. Exposure to both groups will be compared and odds ratio will be calculated. Here number of people exposed to childhood sexual abuse is qualitative variable hence this formula will be used for such type of design

$$\text{Sample size} = \frac{r+1}{r}\, \frac{(p^{*})(1-p^{*})(Z_{\beta}+Z_{\alpha/2})^{2}}{(p_{1}-p_{2})^{2}}$$

r = Ratio of control to cases, 1 for equal number of case and control
p* = Average proportion exposed = proportion of exposed cases + proportion of control exposed/2
Z_{β} = Standard normal variate for power = for 80% power it is 0.84 and for 90% value is 1.28. Researcher has to select power for the study.
$Z_{\alpha/2}$ = Standard normal variate for level of significance as mentioned in previous section.
$p_{1}-p_{2}$ = Effect size or different in proportion expected based on previous studies. p_{1} is proportion in cases and p_{2} is proportion in control.

So if the researcher wants to calculate sample size for the above-mentioned case control study to know link between childhood sexual abuse with psychiatric

disorder in adulthood and he wants to fix power of study at 80% and assuming expected proportions in case group and control group are 0.35 and 0.20 respectively, and he wants to have equal number cases and control; then the sample size per group will be

$$\text{Sample size} = \frac{2}{1} \frac{(p^*)(1-p^*)(0.84+1.96)^2}{(0.35-0.20)^2}$$

p* = Average proportion exposed = (proportion of exposed cases + proportion of control exposed)/2
= (0.35 + 0.20)/2 = 0.275

$$\text{So sample size} = \frac{2}{1} \frac{(0.275)(1-0.275)(0.84+1.96)^2}{(0.35-0.20)^2}$$
$$= 138.9$$

So the researcher has to recruit at least 139 subjects in cases and equal number in control as he wants to have equal number in both.

For quantitative variable
Suppose a researcher wants to see the association between birth weight and diabetes in adulthood. The birth weight being a quantitative data, the researcher will select one group i.e. cases that will be diabetic adults and other group i.e. control will be non-diabetic adults. Both groups will be traced back for data regarding childhood weight. The formula for sample size calculation is

$$\text{Sample size} = \frac{r+1}{r} \frac{SD^2(Z_\beta + Z_{\alpha/2})^2}{d^2}$$

SD = Standard deviation = researcher can take value from previously published studies
d = Expected mean difference between case and control (may be based on previously published studies.)
r, Z_β, $Z_{\alpha/2}$ are already explained in previous sections.

So if researcher think that difference in mean weight between case and control may be around 250 gm and SD is 1 Kg then considering equal number of cases and control and 80% power the sample size will be

$$\text{Sample size} = \frac{2}{1} \frac{1^2(0.84+1.96)^2}{0.25^2} = 250.88$$

Hence researcher has to take 251 subjects in each group (case and control).

Sample size calculation of cohort studies
In cohort studies healthy subjects with or without exposure to some risk factor are observed over a time period to see the event rate in both groups. If a researcher wants to see the impact of weight training exercise on cardiovascular mortality then he will select two groups, one consisting of subjects who do exercise and another consisting of those who don't do. These groups will be followed up for a specific time period to see cardiovascular mortality in both groups. At the end of the study period both groups will be compared for cardiovascular mortality. The formula for sample size is

$$\text{Sample size} = \frac{\left[Z_\alpha \sqrt{\left(1+\frac{1}{m}\right)p^*(1-p^*)} + Z_\beta \sqrt{p1(1-p1)/m + p2(1-p2)} \right]^2}{(p_1 - p_2)^2}$$

Z_α = Standard normal variate for level of significance
m = Number of control subject per experimental subject
Z_β = Standard normal variate for power or type 2 error as explained in earlier section
$p1$ = Probability of events in control group
$p2$ = Probability of events in experimental group p

$$P^* = \frac{p2 + mp1}{m+1}$$

So suppose the researcher wants to see the impact of weight training exercise on cardiovascular mortality and according to previous studies proportion of cardiovascular death in case may be around 20% and in control it can be around 40% hence sample size calculation for 5% of significant level and 80% power with equal number of case and control will be

$$\frac{\left[1.96\sqrt{\left(1+\frac{1}{1}\right)0.30(1-0.30)} + 0.84\sqrt{0.40(1-0.40)/1 + 0.20(1-0.20)} \right]^2}{(0.40-0.20)^2} = 59.41$$

So, the researcher has to take 59 samples in each group.

It is worthy of mention here that these formulas for case control and cohort study are for independent design studies. They are not for matched case control and cohort studies. These formulae can be modified or corrected depending on population size or ratio between sample size and population size. Detailed text should be read to know more about technical aspects of sample size calculation.[7,8] Readers are advised to use various freely available epidemiological calculators like openEpi given in appendix to calculate sample size formula.

Sample size calculation for testing a hypothesis (Clinical trials or clinical interventional studies)

In this kind of research design researcher wants to see the effect of an intervention. Suppose a researcher want to see the effect of an antihypertensive drug so he will select two groups, one group will be given antihypertensive drug and another group will be give placebo. After giving these drug s for a fixed time period blood pressure of both groups will be measured and mean blood pressure of both groups will be compared to see if difference is significant or not. Complex formulae are used for this type of studies and we want to advise readers to use statistical software for calculation of exact sample size. The procedure for calculation of samle size in clinical trials/intervention studies involving two groups is mentioned here.In the case of only two groups method of calculation is mentioned here but if design involves more than two groups then statistical software like G Power should be used for sample size calculation. But understanding of various prerequisites which are needed for sample size calculation is very important.

Formula for sample size calculation for comparison between two groups when endpoint is quantitative data

When the variable is quantitative data like blood pressure, weight, height, etc., then the followingformula can be used for calculation of sample size for comparison between two groups.

$$\text{Sample size} = \frac{2SD^2(Z_{\alpha/2}+Z_\beta)^2}{d^2}$$

SD – Standard deviation = From previous studies or pilot study
$Z_{\alpha/2} = Z_{0.05/2} = Z_{0.025} = 1.96$ (From Z table) at type 1 error of 5%
$Z_\beta = Z_{0.20} = 0.842$ (From Z table) at 80% power
d = effect size = difference between mean values

So now formula will be

$$\text{Sample size} = \frac{2SD^2(1.96 + 0.84)^2}{d^2}$$

For example, suppose a researcher wants to see the effect of a potential antihypertensive drug and He wants to compare the new drug with placebo. Researcher thinks that if this new drug reduces this blood pressure by 10 mmHg as compared to placebo then it should be considered as clinically significant. Let us assume standard deviation found in previously done studies was 25 mmHg. Suppose the researcher selects the level of significance at 5% and the power of study at 80%., and he thinks suitable statistical test in this condition

will be two tailed unpaired *t* test. The effect size in this condition is 10 mmHg. Hence sample size will be

$$\text{Sample size} = \frac{2(25)^2(1.96 + 0.84)^2}{10^2} = 98$$

So in this case the researcher needs 98 subjects per group.

Formula for sample size calculation for comparison between two groups when endpoint is qualitative

When the endpoint of a clinical intervention study is qualitative like alive/dead, diseased/non diseased, male/female etc., then the following formula can be used for sample size calculation for comparison between two groups. Suppose the researcher is interested in knowing protective effect of a drug on mortality in patients of myocardial infarction. He selected two groups of patients of myocardial infarction one group was given that drug and another group was given placebo. The both groups were kept under observation and at the end of study death in both groups were compared. For sample size of this type of study below mentioned formula can be used.

$$\text{Sample size} = \frac{2(Z_{\alpha/2} + Z_\beta)^2 \, P(1-P)}{(p_1 - p_2)^2}$$

$Z_{\alpha/2} = Z_{0.05/2} = Z_{0.025} = 1.96$ (From Z table) at type 1 error of 5%
$Z_\beta = Z_{0.20} = 0.842$ (From Z table) at 80% power
$p_1 - p_2$ = Difference in proportion of events in two groups
P = Pooled prevalence = [prevalence in case group (p_1) + prevalence in control group (p_2)]/2

In above example, let us assume that previous study says that 20% of patient of myocardial infarction die within a specified time. The researcher feels that if the drug being tested increases survival to 30% then the finding can be considered as clinically significant. Effect size will be difference between proportions. 0.2 – 0.3 = –0.1. At 5% of significance level and 80% power sample size will be

Pooled prevalence = (0.20 + 0.30)/2 = 0.25

$$\text{Sample size} = \frac{2(1.96 + 0.84)^2 \, 0.25(1-0.25)}{(-0.1)^2}$$
$$= 294$$

So researcher needs 294 subjects per group.

So simple calculation for sample size when comparison is for two independent groups can be done manually

by given formulae but for more than two groups or for matched data and for other complex calculations software should be used [appendix 1].

Sample size formula for animal studies

For animal studies there are two method of calculation of sample size. The most preferred method is the same method which has been mentioned in sample size calculation for testing the hypothesis. While all efforts should be done to calculate the sample size by that method, sometimes it is not possible to get information related to the prerequisites needed for sample size calculation by power analysis like standard deviation, effect size etc.In that condition a second method can be used this is called as "resource equation method".[9] In this method a value E is calculated based on decided sample size. The value if E should lies within 10 to 20 for optimum sample size. If a value of E is less than 10 then more animal should be included and if it is more than 20 then sample size should be decreased.

E = Total number of animals – Total number of groups

Suppose in an animal study a researcher formed 4 groups of animal having 8 animals each for different interventions then total animals will be 32 (4 × 8). Hence E will be

E = 32 – 4 = 28

This is more than 20 hence animals should be decreased in each group. So if researcher takes 5 rats in each group then E will be

E = 20 – 4 = 16

E is 16 which lies within 10-20 hence five rats per group for four groups can be considered as appropriate sample size. This is a crude method and should be used only if sample size calculation cannot be done by power analysis method explained in above section for testing the hypothesis.

REFERENCES

1. Jaykaran, Saxena D, Yadav P, Kantharia ND. Negative studies published in medical journals of India do not give sufficient information regarding power/sample size calculation and confidence interval. J Postgrad Med 2011;57:176-7.
2. Jaykaran, Yadav P, Kantharia ND. Reporting of sample size and power in negative clinical trials published in Indian medical journals. J Pharm Negative Results 2011;2:87-90.
3. Jaykaran, Saxena D, Yadav P. Negative studies published in Indian Medical journals are underpowered. Indian Pediatr 2011;48:490-1.
4. Naduvilath TJ, John RK, Dandona L. Sample size for ophthalmology studies. Indian J Ophthalmol [serial online] 2000;48:245. Available from: http://www.ijo.in/text.asp?2000/48/3/245/14864. [Last cited in 2012 Sep 2].
5. Patra P. Sample size in clinical research, the number we need. Int J Med Sci Public Health 2012;1:5-9.
6. Shah H. How to calculate sample size in animal studies? Natl J Physiol Pharm Pharmacol 2011;1:35-9.
7. Kasiulevicius V, Sapoka V, Filipaviciute R. Sample size calculation in epidemiological studies. Gerontology 2006;7:225-31.
8. Cai J, Zeng D. Sample size/power calculation for case-cohort studies. Biometrics 2004;60:1015-24.
9. Festing MF, Altman DG. Guidelines for the design and statistical analysis of experiments using laboratory animals. Institute for Laboratory Animal Research. ILAR J 2002;43:244-58.

APPENDIX 1: – FREE SOFTWARE AND CALCULATORS AVAILABLE ONLINE FOR SAMPLE SIZE CALCULATION

Open Epi
http://www.openepi.com/OE2.3/Menu/OpenEpiMenu.htm

Biomath
http://www.biomath.info

EpiTools epidemiological calculators.
http://epitools.ausvet.com.au/content.php?page=home

Java applets for power and sample size
http://homepage.stat.uiowa.edu/~rlenth/Power/#Download_to_run_locally

StatPages
http://statpages.org/

Free statistics
http://www.freestatistics.info/index.php

Department of Biostatistics, Vanderbilt University

http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize

G Power
http://www.psycho.uni-duesseldorf.de/aap/projects/gpower

Power analysis for ANOVA
http://www.math.yorku.ca/SCS/Online/power

Statistical considerations for clinical trials and scientific experiments
http://hedwig.mgh.harvard.edu/sample_size/size.html

Statistics calculators
http://danielsoper.com/statcalc3/default.aspx

Laboratory animal science program
http://web.ncifcrf.gov/rtp/lasp/intra/calculation.asp

WINPEPI (PEPI for windows)
http://www.brixtonhealth.com/pepi4windows.html