Sample size and power analysis in medical research

Sanjay P. Zodpey
Clinical Epidemiology Unit, Department of Preventive and Social Medicine, Government Medical College, Nagpur, Maharashtra, India.

Address for Correspondence: Dr. Sanjay P. Zodpey, A/303, Amar Enclave, Prashant Nagar, Ajni, Nagpur - 440015, India.
E-mail: spzodpey@hotmail.com

ABSTRACT

Among the questions that a researcher should ask when planning a study is "How large a sample do I need?" If the sample size is too small, even a well conducted study may fail to answer its research question, may fail to detect important effects or associations, or may estimate those effects or associations too imprecisely. Similarly, if the sample size is too large, the study will be more difficult and costly, and may even lead to a loss in accuracy. Hence, optimum sample size is an essential component of any research. When the estimated sample size can not be included in a study, post-hoc power analysis should be carried out. Approaches for estimating sample size and performing power analysis depend primarily on the study design and the main outcome measure of the study. There are distinct approaches for calculating sample size for different study designs and different outcome measures. Additionally, there are also different procedures for calculating sample size for two approaches of drawing statistical inference from the study results, i.e. confidence interval approach and test of significance approach. This article describes some commonly used terms, which need to be specified for a formal sample size calculation. Examples for four procedures (use of formulae, readymade tables, nomograms, and computer software), which are conventionally used for calculating sample size, are also given

KEY WORDS: Sample size, Power analysis, Medical research

INTRODUCTION

Medical researchers primarily consult bio-statisticians for two reasons. Firstly, they want to know how many subjects should be included in their study (sample size) and how these subjects should be selected (sampling methods). Secondly, they desire to attribute a p value to their results to claim significance of results. Both these bio-statistical issues are interrelated. If a study does not have an optimum sample size, the significance of the results in reality (true differences) may not be detected. This implies that the study would lack power to detect the significance of differences because of inadequate sample size.[1] Whatever outstanding results the study produces, if the sample size is inadequate their validity would be questioned.

If the sample size is too small (less than the optimum sample size), even the most rigorously executed study may fail to answer its research question, may fail to detect important effects or associations, or may estimate those effects or associations too imprecisely. Similarly, if the sample size is too large (more than the optimum size), the study will be more difficult and costly, and may even lead to a loss in accuracy, as it is often difficult to maintain high data quality. Hence, it

How to cite this article: Zodpey SP. Sample size and power analysis in medical research. Indian J Dermatol Venereol Leprol 2004;70:123-8.
Received: December, 2003. **Accepted:** January, 2004. **Source of Support:** Nil.

is necessary to estimate the optimum sample size for each individual study.[1] For these reasons, in recent years, medical literature has focused increasing attention on sample size requirements in medical research[2] and peer reviewed journals seriously look for the appropriateness of sample size in their manuscript review process.

Basically, the issue of sample size can be addressed at two stages of the actual conduct of the study. Firstly, one can calculate the optimum sample size required during the planning stage, while designing the study, using appropriate approaches and information on some parameters. Secondly, the issue of sample size can be addressed through post-hoc power analysis at the stage of interpretation of the results. In practice, the size of a study is often restricted because of limited financial resources, availability of cases (rare diseases) and time limitation. In these situations the researcher completes the study using the available samples and performs post-hoc power analysis.[1]

It is also important to note that the requirement for estimating the sample size depends primarily on the study design and the main outcome measure of the study. There are various study design options available for conducting medical research. A medical researcher needs to select an appropriate study design to answer the research question. There are many different approaches for calculating the sample size for different study designs. For example, the procedure of calculating the sample size is different for a case-control design than for a cohort design. Similarly, there are different approaches for calculating the sample size for cross-sectional studies, clinical trials, diagnostic test studies, etc. Moreover, within each study design there could be more sub-designs and the sample size calculation approach would vary accordingly. For case-control studies, the approach for calculating the sample size is distinct for matched and un-matched designs. Hence, one must use the correct approach for computing the sample size appropriate to the study design and its subtype.[1]

The second important issue that should be considered while computing the sample size is the primary outcome measure. The primary outcome measure is usually reflected in the primary research question of the study and also depends on the study design. For estimating the risk in a case-control study the primary outcome measure would be the odds ratio, but while estimating the risk in a cohort study it would be the relative risk. In a case-control study, the primary outcome measure could be the difference in means/ proportions of exposure in cases and controls, crude odds ratio, adjusted odds ratio, attributable risk, population attributable risk, prevented fraction, etc. While calculating the sample size, one of these primary outcome measures has to be specified since there are distinct approaches for calculating the sample size for each of these outcomes.[3] Similarly, for each study design there could be many outcomes and a researcher needs to specify the main outcome measure of the study.

For drawing a statistical inference from the study results two approaches are used: estimation (confidence interval approach) and hypothesis testing (test of significance approach). The procedures for calculating the sample size for these two approaches differ and are available in the literature.[1,2,4,5] A researcher needs to select the appropriate procedure for computing the sample size and accordingly use the approach of drawing a statistical inference subsequently.

Moreover, one also needs to specify some additional parameters depending upon the approach chosen for calculating the sample size. They are hypothesis (one or two tailed), precision, type I error, type II error, power, effect size, design effect, etc. For understanding the principles of sample size calculation and power analysis, one should have an understanding of these commonly used terms.

DESCRIPTION OF SOME COMMONLY USED TERMS[1]

Random error

It describes the role of chance, particularly when the effects of explanatory or predictive factors have already been taken into account. Sources of random error include sampling variability, subject to subject differences and measurement errors. It can be controlled and reduced to acceptably low levels by averaging, increasing the sample size and by repeating the experiment.

Systematic error (Bias)

It describes deviations that are not a consequence of chance alone. Several factors, including the patient selection criteria, might contribute to it. These factors may not be amenable to measurement, but can usually be removed or reduced by good design and conduct of the experiment. A strong bias can yield an estimate very far from the true value, even in the wrong direction.

Precision (Reliability)

It describes the degree to which a variable has the same value when measured several times. It is a measure of consistency. Sometimes it simply refers to the width of the confidence interval. It is a function of random error (the greater the error, the less precise the measurement), the sample size, the confidence interval required and the variance of the outcome variable. A larger sample size would give precise estimates.

Accuracy (Validity)

It indicates the degree to which the variable actually represents what it is suppose to represent. It is a function of systematic error or bias.

Null hypothesis

This is a hypothesis which states that there is no difference among groups or that there is no association between the predictor and the outcome variables. This hypothesis needs to be tested.

Alternative hypothesis

This is a hypothesis that in some sense contradicts the null hypothesis. It assumes that there is a difference among the groups or there exists an association between the predictor and outcome variable. If an alternative hypothesis cannot be tested directly, it is accepted by exclusion if the test of significance rejects the null hypothesis. There are two types of alternative hypothesis: one-tailed (one-sided) hypothesis and two-tailed (two-sided) hypothesis. One-tailed hypothesis specifies the difference (or effect or association) in one direction only. For example, patients with pancreatic cancer will have a higher rate of coffee drinking as compared to control subjects. Two-tailed hypothesis specifies the difference (or effect or association) in either direction. For example, patients with pancreatic cancer will have a different rate of coffee drinking – either higher or lower – as compared to control subjects. A one-tailed approach leads to a smaller sample size. However, the decision to use the one- or two-tailed approach depends on the clinical or biological importance or relevance of the research question and prior knowledge about effect or association. This decision should not be based on sample size considerations.

Type I (α) error

It occurs if an investigator rejects a null hypothesis that is actually true in the population. It is the error of falsely stating that two drug effects are significantly different when they are actually equivalent. This is the probability of erroneously finding a disease exposure association, when none exists in reality. The probability of making α error is called as level of significance and is conventionally considered as 0.05 (5%). For computing the sample size its specification in terms of Z_α is required. The quantity Z_α is a value from the standard normal distribution corresponding to α. For a one-sided test of the hypothesis, Z_α is taken to be the value of the standard normal distribution corresponding to a. For α two-sided test, Z_α is taken to be the value that is exceeded with probability $\alpha/2$. The sample size is inversely proportional to type I error.

Type II (β) error

It occurs if the investigator fails to reject a null hypothesis that is actually false in the population. It is the error of falsely stating that two drug effects are equivalent when they are actually different. This is the probability of not erroneously finding disease exposure association, when it exists in reality. For computing the sample size its specification in terms of Z_β is required. The quantity Z_β is a value from the standard normal distribution corresponding to β. For either a one-sided or two-sided test, Z_β is taken to be the value that is exceeded with probability β. The values of Z_α and Z_β for the selected values of α and β are presented in Table 1. The sample size is inversely proportional to type II error.

Power (1-β)

This is the probability that the test will correctly identify a significant difference or effect or association in the

Table 1: Unit normal deviates Z_α and Z_β for selected values of α and β

α or β	One-sided Z_α and Z_β	Two-sided Z_α
0.05	1.64	1.96
0.10	1.28	1.64
0.20	0.84	1.28

Z_β is the same for one-sided and two-sided tests

sample should one exist in the population. This is expressed as 1-β. The sample size is directly proportional to the power of the study. The larger the sample size, the study will have greater power to detect significance of difference or effect or association.

Effect size

The effect size refers to the magnitude of the effect under the alternative hypothesis. It should represent the smallest difference that would be of clinical or biological significance. It varies from study to study. For example, a treatment effect that reduces mortality by 1% might be clinically important, while a treatment effect that reduces transient asthma by 20% may be of little clinical interest. It is also variable from one statistical procedure to the other. It could be the difference in cure rates, or a standardized mean difference or a correlation coefficient. If the effect size is increased, the type II error decreases. Power is a function of an effect size and the sample size. For a given power, 'small effects' require larger sample size than 'large effects'. Table 2 shows the sample size for various effect sizes at a fixed power and level of significance.

Design effect

Geographic clustering is generally used to make the study easier and cheaper to perform. The effect on the sample size depends on the number of clusters and the variance between and within clusters. In practice this is determined from previous studies or from studies of a similar type in literature, and is expressed as a constant called 'design effect', often between 1.0 and 2.0. It is the ratio of the variance when cluster sampling is used to the variance when simple random sampling is used. The sample sizes for simple random samples are multiplied by the design effect to obtain the sample size for the clustered sample.

Table 2: Sample size for various effect sizes at a fixed power (80%) and level of significance (two-tailed, α = 0.05)

Effect size (Two sample proportions)	Sample size per group
30% vs. 40% (i.e. small effect size)	356
30% vs. 50% (i.e. intermediate effect size)	93
30% vs. 60% (i.e. large effect size)	47

Procedures for calculating the sample size

There are four procedures that could be used for calculating sample size: use of formulae, readymade tables, nomograms, and computer software.

Use of formulae for sample size calculation and power analysis

There are more than one hundred formulae for calculating the sample size and power in different situations for different study designs.[1] The following are two examples of their use in medical research.

To investigate the role of oral contraceptives (OC) in the etiology of cutaneous malignant melanoma in women, an unmatched case-control study is to be undertaken. For calculating the sample size for this study using formulae,[3] the following parameters have to be specified:

p_0 (Prevalence of exposure in control population) = 0.30 (approximately 30% women in the population are using OC (information obtained from literature)).

$\alpha = 0.05$ (two-sided), $Z_\alpha = 1.96$

$\beta = 0.10$, $Z_\beta = 1.28$ (power = 90%)

Odds ratio (OR) = 2 (information obtained from literature or from earlier studies in other population settings)

FORMULA[3]

$n = 2p' q' (Z_\alpha + Z_\beta)^2 / (p_1 - p_0)^2$

$p_1 = p_0 \text{ OR} / (1 + p_0 (\text{OR} - 1))$

$p' = (1/2) (p_1 + p_0)$, $q' = 1 - p'$, $q1 = 1 - p_1$, $q_0 = 1 - p_0$

Solution (by putting above specified values in the formula): n = 188 in each group.

If we decide to study only 50 cases and 50 controls, then with the other specifications unchanged, the power of the study would be as follows.

Formula:[3] $Z_\beta = \{[\text{sqrt}(n(p_1 - p_0)^2)] - [Z_\alpha \text{ sqrt } (2p'q')]\} / \{\text{sqrt } (p_1 q_1 + p_0 q_0)\}$

The power is determined from tables of the normal distribution by finding the probability with which the calculated value of Z_β is not exceeded.

Solution (by putting the above specified values in the formula): $Z_\beta = -1.13$.
From tables of the normal probability function, one finds, Power = p ($Z \leq -1.13$) = 0.13.
Thus if the odds ratio in the target population is 2, a case-control study of n = 50 per group has only a 13% chance of finding that the sample estimate will be significantly ($\alpha = 0.05$) different from unity.

Use of readymade tables for sample size calculation[1-5]

How large a sample of patients should be followed up if an investigator wishes to estimate the incidence rate of a disease to within 10% of its true value with 95% confidence? Here, the relative precision (ε) is 10% and the confidence level is 95%. Table 3 shows that for ε = 0.10 and a confidence level of 95% a sample size of 385 would be needed. This table can be used to calculate the sample size making the desired changes in the relative precision and confidence level, e.g. if the level of confidence is reduced to 90%, then the sample size required would be 271. Such tables that give ready made sample sizes are available for different designs and situations.

Table 3: Estimating an incidence rate with specified relative precision [Formula: $n = (Z_{1-\alpha/2} / \varepsilon)^2$]

Relative precision (ε)	Confidence level		
	99%	95%	90%
0.01	66358	38417	27061
0.02	16590	9605	6766
0.03	7374	4269	3007
0.04	4148	2402	1692
0.05	2655	1537	1083
0.06	1844	1068	752
0.07	1355	785	553
0.08	1037	601	423
0.09	820	475	335
0.10	664	385	271
0.12	461	267	188
0.14	339	197	139
0.16	260	151	106
0.18	205	119	84
0.20	166	97	68
0.22	138	80	56
0.24	116	67	47
0.26	99	57	41
0.28	85	50	35
0.30	74	43	31
0.32	65	38	27
0.34	58	34	24
0.36	52	30	21
0.38	46	27	19
0.40	42	25	17
0.42	38	22	16
0.44	35	20	14
0.46	32	19	13
0.48	29	17	12
0.50	27	16	11

Use of nomograms for sample size calculation[6,7]

For using a nomogram to calculate the sample size, one needs to specify the study (treatment/group 1) and the control groups (placebo/group 2). This could be arbitrary or based on the study design; the nomogram will work either way. The researcher should then decide the effect size that is clinically important to detect. This should be expressed in terms of the percentage change in the response rate compared with that of the control group. For example, if 40% of patients treated with the standard therapy are cured and one wants to know whether a new drug can cure 50%, one is looking for a 25% increase in the cure rates [((50% - 40%)/40%) = 25%].

The desired percentage change is located on a horizontal axis of the nomogram (x line, Figure 1). A vertical line is extended to intersect with the diagonal line corresponding to the response rate in the control

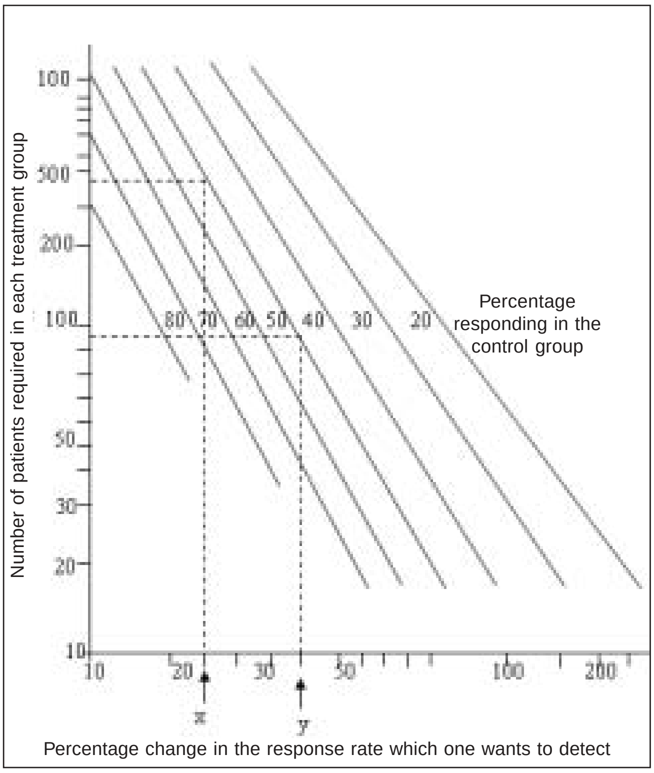

Figure 1: Nomogram for calculating sample size for studies using dichotomous variables. The nomogram is for a significance level of $\alpha = 0.05$ (two-sided), and $\beta = 0.20$ (one-sided). The points x and y may be commonly used, and one of them is used in the example in the text

group. If the appropriate diagonal line does not extend far enough to intersect with this vertical line, one can try using the other treatment group as the control group. The symmetrical design of the nomogram allows an arbitrary designation of control group. Finally, a horizontal line is extended from this point to the vertical axis, showing the sample size required for both the treatment and control groups.

EXAMPLE[6]

A study randomly allocates patients with an infectious disease to treatment with drug A or drug B. The study reports a 40% cure rate using drug A, the current standard therapy, and a 45% cure rate using drug B, a new drug. The study concludes that there is no statistically significant difference in response rates between the two drugs. There are 150 patients in each treatment group.

A researcher, who is reading this study, believes that previous studies suggest a better response rate in patients treated with drug B. He decides that a 25% improvement in the usual response rate from drug A, from 40% to 50%, would be important for him. He does not consider a smaller difference to be clinically important. Using the nomogram, he finds that the sample size required to detect a 25% difference in cure rate between drug A and drug B, assuming a control group cure rate of 40%, is about 370 (line x, Figure 1). This is the sample size that ensures an 80% chance of detecting this difference if it exists, assuming α of 0.05. Because there are only 150 patients in each treatment group, the sample size is clearly inadequate; it is not large enough to be sure that a clinically important 25% difference in cure rates does not exist. The researcher, therefore, feels justified in continuing to prescribe drug B since previous evidence suggests that it is more effective and the new study, despite its negative results, is too small to refute this evidence.

A separate nomogram is available for continuous variables.[6] Both these nomograms are intended to provide the clinician with a handy and easy-to-use reference for ascertaining whether an apparently negative study has a sample size adequate to detect reliably any important difference between treatment groups.

Use of computer software for sample size calculation and power analysis

The following software can be used for calculating sample size and power: STATA, Epi-Info, Sample, Power and Precision, and nQuerry Advisor.

REFERENCES

1. Zodpey SP, Ughade SN. Workshop manual: Workshop on Sample Size Considerations in Medical Research. Nagpur: MCIAPSM; 1999.
2. Arkin CF, Wachtel MS. How many patients are necessary to assess test performance? JAMA 1990;263:275-8.
3. Schlesselman JJ. Case-control studies – Design, conduct and analysis. 1st ed. New York: Oxford University Press; 1982.
4. Bach LA, Sharpe K. Sample size for clinical and biological research. Aust NZ J Med 1989;19:64-8.
5. Lwanga SK, Lemeshow S. Sample size determination in health studies – A practical manual. 1st ed. Geneva: World Health Organization; 1991.
6. Young MJ, Bresnitz EA, Strom BL. Sample size nomograms for interpreting negative clinical studies. Ann Int Med 1983;99: 248-51.
7. Altman DG, Gore SM. Statistics in practice. Harrow, Middlesex: BMJ 1982.