NTI Bulletin 2006,42/3&4, 55 - 62

Sample size determination in health studies

*VK Chadha**

Summary

One of the most important factors to consider in the design of an intervention trial is the choice of an appropriate study size. Studies that are too small may fail to detect important effects on the outcomes of interest, or may estimate those effects too imprecisely. Studies that are larger than necessary are a waste of resources. Statistical methods are available for estimation of appropriate sample size depending upon the type of outcome measure, expected disease rates or size of effects, study design and the requirements of confidence interval/precision or power. These concepts along with the methods of estimating sample size in varied situations are presented in this article.

Key words: Sample size, proportion, mean, confidence interval, precision, null hypothesis, power, Type I and Type II errors.

Introduction

One of the most important factors to consider in the design of an intervention trial is the choice of an appropriate study size. Studies that are too small may fail to detect important effects on the outcomes of interest, or may estimate those effects too imprecisely. Studies that are larger than necessary are a waste of resources.

Before calculating sample size one has to decide on the following:

Study design

Types of outcome measure

Guess at likely result

Required level of significance

Required precision / power

The procedures for sample size estimation provide a rough estimate of the required study size, as they are often based on approximate estimates of expected disease rates and subjective decisions about the size of effects. However, a rough estimate of the necessary size of a study is generally all that is needed.

One should be familiar with the following concepts before embarking on the estimation of sample size.

Types of outcome measures

The statistical methods for sample size determination depend on which type of outcome is expected. The 3 most common types of outcomes in case of surveys / studies / trials are:

i. Proportions : For example, in a trial of a new measles vaccine, an outcome measure of interest may be the proportion of vaccinated subjects who develop high levels of antibodies.

ii. Means : For example, in a trial of an anti-malarial intervention, it may be of interest to compare the mean packed cell volume (PCV) at the end of the malaria season among those in the intervention group and those in the comparison group.

iii. Rates : For example, in a trail of multi-drug therapy for leprosy, the incidence rates of relapse following treatment may be compared in the different study groups under consideration.

* Sr. Epidemiologist, National Tuberculosis Institute, Bangalore

For a quantitative variable, it is only the rough estimates of the proportion, means or rates that are required for estimating sample size.

Sampling error

An estimate of an outcome measure calculated in an intervention study is subject to sampling error, because it is based on a sample of individuals and not on the whole population of interest. The term does not mean that the sampling procedure was applied incorrectly, but that when sampling is used to decide which individuals are in which group, there will be an element of random variation in the results. Sampling error is reduced when the study size is increased and vice versa.

Confidence Interval

The methods of statistical inference allow the investigator to draw conclusions about the true value of the outcome measure on the basis of the information in the sample. In general, the observed value of the outcome measure gives the best estimate of the true value. In addition, it is useful to have some indication of the precision of this estimate, and this is done by attaching a confidence interval to the estimate. The confidence interval is a range of plausible values for the true value of the outcome measure. It is conventional to quote the 95 percent confidence interval (also called 95 percent confidence limits). This is calculated in such a way that there is a 95 percent probability that it includes the true value.

If the outcome measure is a proportion estimated from the sample data as $\grave{o}$. The 95 percent confidence intervals to be presented here are = $\grave{o}$ ± 1.96 x SE, where SE denotes the standard error of the estimate. Similarly, if the outcome measure is a mean, 95 percent of the values derived from different samples are expected to fall within 1.96 standard deviations of the mean.

One of the factors influencing the width of the confidence interval is the sample size. The larger the sample size, the narrower is the confidence interval.

The multiplying factor 1.96 is used when calculating the 95 percent confidence interval. In some circumstances, confidence intervals other than 95 percent limits may be required and then values of the multiplying factor are as under:

Confidence interval (%)	Multiplying factor
90	1.64
95	1.96
99	2.58
99.9	3.29

Confidence intervals and their corresponding multiplying factors, based on the Normal distribution

Precision of effect measures – The narrower the confidence interval, the greater the precision of the estimate.

Significance tests & P value

In some instances, before calculating a confidence interval to indicate a range of plausible values of the outcome measure of interest, it may be appropriate to test a specific hypothesis about the outcome measure. In the context of an intervention trial, this will often be the hypothesis that there is no true difference between the outcomes in the groups under comparison- null hypothesis. The objective is thus to assess whether any observed difference in outcomes between the study groups may have occurred just by chance due to sampling error. The methods for testing the null hypothesis are known as significance tests. The sample data are used to calculate a quantity (called a statistic) which gives a measure of the difference between the groups with respect to the outcome(s) of interest. Once the statistic has been calculated,

its value is referred to an appropriate set of statistical tables, in order to determine the p – value (probability value) or 'significance' of the results.

For example, suppose a difference in mean PCV of 1.5 percent is observed at the end of the malaria season between two groups of individuals, one of which was supplied with mosquito-nets. A p-value of 0.03 would indicate that, if nets had no true effect on PCV levels, (if null hypothesis was true) there would only be a 3 percent chance of obtaining an observed difference of 1.5 percent or greater.

The smaller the p-value, the less plausible the null hypothesis seems as an explanation of the observed data. For example, a p-value of 0.001 means that the null hypothesis is highly implausible, and this can be interpreted as very strong evidence of a real difference between the groups. On the other hand, a p-value of 0.20 means that a difference of the observed magnitude could quite easily have occurred by chance, even if there was no real difference between the groups. Conventionally, p-values of 0.05 and below have been regarded as sufficiently low to be taken as reasonable evidence against the null hypothesis, and have been referred to as indicating a 'statistically significant difference'.

While a small p-value can be interpreted as evidence for a real difference between the groups, a larger 'non-significant' p-value must not be interpreted as indicating that there is no difference. It merely indicates that there is insufficient evidence to reject the null hypothesis, so that there may be no true difference between the groups.

Power of study

The concept of power comes into play when the focus of the study is to find out whether a significant difference exists between the two groups. Because of the variations resulting from sampling error, we cannot always be certain of obtaining a significant result of a study, even if there is a real difference. It is necessary to consider the probability of obtaining a statistically significant result in a trial, and this probability is called the power of the study. In other words, if the true difference exits, power of the study indicates the probability of finding a statistically significant difference between the two groups. Thus a power of 80 percent to detect a difference of a specified size means that if the study were to be conducted repeatedly, a statistically significant result would be obtained four times out of five if the true difference was really of the specified size. When designing a study, the objective is to ensure that the study size is large enough to give high power if the true effect of the intervention is large enough to be of practical importance.

The power of a study depends on:

1. The value of the true difference between the study groups; in other words, the true effect of the intervention (effect size). The greater the effect, the higher the power to detect the effect as statistically significant for a study of a given size.
2. The study size; The larger the study size, higher is the power.
3. The probability level at which a difference will be regarded as 'statistically significant'.

The power also depends on whether a one-sided or two sided significance test is to be performed and on the underlying variability of the data.

One sided and Two sided tests

If it is accepted that the null hypothesis is false, that means alternate hypothesis is true. For example if the claim is about superiority of a new drug, this is a one-sided alternative. If the claim is not of superiority or inferiority but only that they are different, the alternate hypothesis is two sided. This means that when

the p-value is computed, it measures the probability (if the null hypothesis is true) of observing a difference as great as that actually observed in either direction (i.e. positive or negative). It is usual to assume that tests are two-sided.

Wrongly rejecting a true null hypothesis is called type I error. The probability of this error is referred as P value as already discussed. The maximum P value allowed in a problem is called the level of significance (α). In a diagnostic set up, this is the probability of declaring a person sick when he is actually not. In a clinical trial set up, P value is the probability that the drug is declared effective or better when it is actually not.

Diagnosis	**Disease actually present**	
	No	**Yes**
Disease present	Mis-diagnosis	✔
Disease absent	✔	Missed diagnosis

In a court set up, this corresponds to convicting an innocent.

Judgment	**Assumption of innocence**	
	true	**false**
Pronounced guilty	Serious error	✔
Pronounced not guilty	✔	error

In a clinical trial set up, P value is the probability that the drug is declared effective or better when it is actually not.

Statistical decision	**Null hypothesis**	
	True	**False**
Rejected	Type I error	✔
Not rejected	✔	Type II error

This wrong conclusion can allow an ineffective drug to be marketed as being effective. In a court set up, this corresponds to convicting an innocent. This clearly is unacceptable and needs to be guarded against. For this reason, P value is kept at a low level (<5%). When P value is small, it is safe to conclude that groups are different. This threshold, 0.05 is the level of significance.

The second type of error is failure to reject null hypothesis when it is actually false. This corresponds to missed diagnosis as also to pronounce a criminal not guilty. The probability of this error is denoted by β. In a clinical trial set up, this is equivalent to declaring a drug ineffective when it is actually effective. The complimentary probability of type II error is the statistical power ($1-\beta$). Thus the power of a statistical test is a probability of correctly rejecting a null hypothesis when it is false.

Choice of criterion

The choice of which of the above two criteria (precision or power) should be used in any particular instance depends on the objectives of the study. If it is known unambiguously that the intervention has some effect, it makes little sense to test the null hypothesis, and the objective may be to estimate the magnitude of the effect, and to do this with acceptable precision.

In studies of new interventions, it is often not known whether there will be any impact at all on the outcomes of interest. In these circumstances, it may be sufficient to ensure that there will be good chance of obtaining a significant result if there is indeed an effect of some specified magnitude. It should be emphasized, however, that if this course is adopted, the estimates obtained may be very imprecise.

Usually, it is more important to estimate the effect of the intervention and to specify a confidence interval around the estimate to indicate the likely range, than to test a specific hypothesis. Therefore, in many situations it may be more appropriate to choose the sample size by setting the width of the confidence interval, rather than to rely on power calculations.

Allowances of losses

Losses to follow up occur in most longitudinal studies. Individuals may be lost because of various reasons like refusals, migration, death from cause unrelated to the outcome of interest, etc. Such losses may produce bias as the individuals who are lost often differ in important respects from those who remain in the study. The losses also reduce the size of the sample available for analysis, and this decreases the precision or power of the study.

For these reasons, it is important to make every attempt to reduce the number of losses to a minimum. However, it is rarely possible to avoid losses completely. The reduced power or precision resulting from losses may be avoided by increasing the initial sample size in order to compensate for the expected number of losses. A 20% allowance is generally considered appropriate.

Practical constraints

Resources in terms of staff, vehicles, laboratory capacity, time or money may limit the potential size of a study, and it is often necessary to compromise between the estimated study-size and what can be managed with the available resource. Trying to do a study that is beyond the capacity of the available resources is likely to be unfruitful, as data quality is likely to suffer and the results may be subject to serious bias, or the study may even collapse completely, thus wasting the effort and money that has already been put into it.

If the calculations indicate that a study of manageable size will yield power and/or precision that is unacceptably low, it is probably better not to conduct the study at all.

Consequences of studies those are too small

Suppose first that the intervention under study has little or no effect on the outcome of interest. The difference observed in the study is therefore likely to be non-significant. However, the width of the confidence interval for the effect measure, for example relative risk, will depend upon sample size. If the sample is small the confidence interval will be very wide and even though it will probably include the null value, it will extend to include large values of the effect measure. In other words, the study will have failed to establish that the intervention has no appreciable effect.

In case the intervention does have an appreciable effect, a study that is too small will have low power i.e it will have little chance of giving a statistically significant difference.

FORMULAE FOR SAMPLE SIZE ESTIMATION

I. Estimating a population proportion: With specified absolute precision

Required information for estimating the sample size is as under:-

- Anticipated population proportion: P, a rough estimate of P is sufficient.
- Desired confidence level
- Absolute Precision: (d) - total percentage points of the error that can be tolerated on each side of the figure obtained. For example, if anticipated prevalence of infection of TB in a population is 10%, we would be satisfied if our sample gives a figure of 8-12%. In this case, d = 2% ~ 0.02.

Sample size can be estimated using the following formula :-

$$n = \frac{Z^2_{1-\frac{\alpha}{2}}(1-P)}{d^2} \quad \text{——} \quad (1)$$

P=anticipated proportion, d=absolute precision required on either side of the proportion. p and d are expressed in fractions, z is a constant, its value for a two sided test is 1.96 for 95%

confidence, 1.645 for 90% confidence and 2.576 for 99% confidence.

Example: For a survey aimed at estimating vaccination coverage, P is usually anticipated at 0.5, d=0.1 and the level of significance = 5%. The estimated sample size using the above formula is 96 for random sampling. The estimated sample size is applicable only in case of simple random sampling (SRS). If another sampling method is used, a larger sample size is likely to be needed because of design effect. For cluster sampling strategy, the estimated sample size as above is multiplied by design effect, which is defined as the ratio of variance obtained in cluster survey to the variance for the same sample size adopting SRS.

$$\text{Design effect} = \frac{\text{Sample design variance}}{\text{variance using SRS}}$$

In most sample surveys adopting cluster-sampling strategy, a design effect of 2 is taken. This means twice as many individuals would have to be studied to obtain the same precision as with SRS. Thus, the estimated sample size in the above example would be 96x2=192. However, design effect should ideally be estimated from previously available data or from pilot studies.

II. Estimating a population proportion: With specified relative precision

Required information :-

- Anticipated population proportion : P
- Confidence level
- Relative Precision: ε – The sample result should fall within ε % of the true value.

Sample size can be estimated using following formula :

$$n = \frac{Z^2_{1-\frac{\alpha}{2}}(1-P)}{\varepsilon^2 P} \quad \text{——} \quad (2)$$

Example : For estimating prevalence of tuberculous infection among children of 0-9 years of age in a locality, how many should be included in the sample so that prevalence may be obtained within 10% of true value with 95% confidence. The anticipated P is 5-10%

P= 0.05 (lower limit is to be taken to have larger sample & better precision)

Confidence level is 95%

Relative Precision (å) = 10% = 0.1

The estimated sample size using formula 2 is 3457. For cluster sampling the sample size will be 3457x D (design effect).

III. Estimating the difference between two population proportions with specified absolute precision (Two – sample situations)- equal n in the two groups

Required information :-

- Anticipated population proportions = P_1 & P_2
- Confidence level = 95%
- Absolute precision required on either side of the true value of the difference between proportions = d

Sample size can be estimated using following formula:

$$n = \frac{Z^2_{1-\frac{\alpha}{2}}[P_1(1-P_1) + P_2\ (1-P_2)]}{d^2} \quad \text{——} \quad (3)$$

P_1, P_2 = anticipated value of the proportions in the two populations.

Example : What sample size to be selected from each of two groups of people to estimate a risk difference to be within 3 percentage points of true difference at 95% confidence when anticipated P_1 & P_2 are 40% & 32% respectively.

$$n = \frac{(1.96)^2\ [(0.4\times 0.6) + (0.32\times 0.68)]}{(0.03)^2} = 1953$$

IV. Hypothesis testing for two population

proportions

Required information :-

- Anticipated values of the population proportions: P_1 & P_2
- Level of significance
- Power of the test: 100 (1-β) %

Sample size can be estimated using following formula :

$$n = \frac{\left\{Z_{1-\frac{\alpha}{2}}\sqrt{2\bar{P}(1-\bar{P})} + Z_{1-\beta}\sqrt{P_1(1-P_1)+P_2(1-P_2)}\right\}^2}{(P_1-P_2)^2} \quad -(4)$$

Values of Z for 90% power = 1.28

80% power = 0.82

V. Estimating a population mean: With specified absolute precision

Required information :-

- Variance : σ^2 , known or can be estimated from a pilot study
- Absolute precision : d

Sample size can be estimated using formula:

$$n = \frac{Z^2_{1-\frac{\alpha}{2}} \times \sigma^2}{d^2} \quad \text{——} \quad (5)$$

σ = standard deviation estimated from a pilot study

VI. Estimating population mean: With specified relative precision

Required information :-

- Population mean: μ
- Variance : σ^2
- Relative precision: (ε)

Sample size can be estimated using formula :

$$n = \frac{Z^2_{1-\frac{\alpha}{2}} \times \sigma^2}{\varepsilon^2\mu^2} \quad \text{——} \quad (6)$$

VII. Estimating difference between means of two populations with specified precision.

Difference in means = $\mu_1 - \mu_2$

Standard Deviation (SD) = s_1, s_2

Absolute precision: d

First calculate pooled variance and estimate sample size using formula

Pooled variance (σ^2) = $\frac{(n_1-1)s_1^2-(n_2-1)s_2^2}{(n_1-1)+(n_2-1)}$

$$n = \frac{Z^2_{1-\frac{\alpha}{2}}[2\sigma^2]}{d^2} \quad \text{——} \quad (7)$$

Example : Suppose you want to know the sample sizes (equal groups) required for detecting a mean difference of 0.3 mg / ml between well-nourished and under-nourished children. This difference is considered to be clinically important. Suppose the mean and SD of Hemoglobin (Hb) levels available from an earlier study in a random sample of well-nourished and under-nourished groups were as follows:

Well-nourished (group-1) :

n_1 = 100, $\bar{x}_1$=10.1 and SD_1= 0.9

Under-nourished (group-2) :

n_2 = 70, $\bar{x}_2$= 9.7 and SD_2= 1.1

The SDs do not differ too much and we can pool them. Thus,

Pooled variance (σ^2)

= (99 x 0.9^2 +69 x 1.1^2) / (100 + 69)

= 0.97

$$n = \frac{1.96^2\,[2\times(0.97)^2]}{(0.3)^2} = 80$$

VIII. Hypothesis testing for two population means

Required information :-

Anticipated values of the population means : μ_1 & μ_2

Standard deviation: s_1, s_2

Level of significance

Power of the test: 100 (1-β) %

Sample size can be estimated using formula:

$$\text{Pooled variance } (\sigma^2) = \frac{s_1^2 - s_2^2}{2}$$

$$n = \frac{2\sigma^2\left[Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right]^2}{(\mu_1 - \mu_2)^2} \quad \text{——} \quad (8)$$

STUDIES WITH MULTIPLE OUTCOMES

In most studies, several different outcomes are measured. For example, in a study of the efficacy of insecticide-treated mosquito-nets on childhood malaria, there may be interest in the effect of the intervention on deaths, deaths attributable to malaria, episodes of clinical malaria over a period of time, spleen sizes at the end of the malaria season, packed cell volumes at the end of the season.

The investigator should first focus attention on a few key outcomes. Calculate the required study size for each of these key outcomes. The outcomes that result in the largest study size would be used to determine the size.

SAMPLE SIZE ESTIMATION FOR OTHER SITUATIONS

For more complex study designs, for example two groups of unequal size, comparison of more than two groups, incidence studies, and interventions allocated to communities; the methods of sample size estimation are more complex and may be referred to in a standard statistical text book.

There are computer programs available that perform sample size calculations. In particular, this facility is available in the package 'Epi Info', though it does not cover the full range of possibilities.

SAMPLE SIZE IN QUALITATIVE RESEARCH

Sample size in qualitative research for e.g. Knowledge, Attitude and Practice (KAP) studies, will depend upon expected response, for e.g. the proportion of doctors using intermittent regimen. Based on the expected response, the usual method for estimating sample size can be employed. However, in general assessment of KAP cannot be performed on the basis of a single parameter. If we use approach based on proportions, then we need to calculate 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. Therefore, the usual approach for sample size estimation viz. estimating proportion(s) or mean (s) could be employed.