Sample size for beginners

Charles du V Florey

Department of Epidemiology and Public Health, University of Dundee, Ninewells Hospital and Medical School, Dundee DD1 9SY
Charles du V Florey, *head of department*

BMJ 1993;306:1181-4

The common failure to include an estimation of sample size in grant proposals imposes a major handicap on applicants, particularly for those proposing work in any aspect of research in the health services. Members of research committees need evidence that a study is of adequate size for there to be a reasonable chance of a clear answer at the end. A simple illustrated explanation of the concepts in determining sample size should encourage the faint hearted to pay more attention to this increasingly important aspect of grantsmanship.

The question of how many subjects are needed in a study arises in the planning of most projects but is often left inadequately answered in grant proposals. Although the answer requires statistical reasoning as well as informed guesswork, all researchers should be able to grasp the concepts used in calculations of sample size and many should be able to do the calculations for themselves. This paper is concerned with the simplest aspects of the concept of sample size and is primarily for those with little experience of the subject.

The paper is divided into three main sections. The first explains the principles underlying the estimation of sample size. It is intended for readers who want to grasp only the basic ideas on which calculations of sample size rely to understand the questions a collaborating statistician will ask. The second section describes the practical aspects of calculation in two simple examples for those who want to see what is concerned. It also gives the rationale for the inclusion in grant applications of certain numerical information. Readers should understand the first section before attempting the second. The third section summarises the elements of the calculation of sample size which should appear in a grant application.

The reader will need some knowledge of statistical concepts such as given in the book *Medical statistics on microcomputers*.[1]

The principles

Calculations of sample size depend on four factors: the variance of the variable being studied, the size of the effect of interest, the level of significance, and the power of the test. In this section we look at the relations between these factors by first considering how the means from samples taken from the same population vary both one from another and according to sample size.

SAMPLING VARIATION AND VARIANCE

The central problem of using samples to tell us about the group they represent is that the results are influenced by the play of chance. This is called sampling variation. To illustrate this let us set up an experiment to estimate the percentage of boys among newborn infants in Scotland. From statistics published by the government, 52% of babies born in Scotland are known to be male. This percentage comes from counting all 53 000 births in a year according to sex. How precisely can an estimate of this population percentage be made by taking a small sample of say 100 births? (Box 1.)

Box 1

- *Population* refers to the universe of items being sampled, in this case all infants in Scotland born in a particular year
- *Sample* refers to a selection of items from the population, in this case the 100 infants chosen randomly

From a single random sample of 100 infants the percentage of boys is found to be 54%. This is an estimate of the population value. A second sample would in all likelihood give a slightly different result even though it had been taken from the same population as the first. If a very large number of samples each of 100 infants were taken the percentages from all the samples could be plotted as a histogram. The plot in figure 1 shows what the distribution would look like.

Figure 1 was made by simulation on a computer. Twenty thousand random samples of 100 were selected and the frequency of samples was plotted according to the percentages of boys observed. In the simulation all the samples were taken from the same population so all the variation from sample to sample in the percentage of boys is due only to chance. There was no other influence on the result of each sample of 100 infants except chance, but even under these conditions a few samples had very extreme values, a long way from the population value. The variation from one sample to another is measured as the variance (box 2).

The distribution is normal, centred around the population value of 52%. Figure 2 shows the theoretical shape of the curve for samples of size 100 with the population value of 52% and the ranges within which 68% and 95% of the distribution lie.

It is conventional to call any observed proportion unusual or surprising if it occurs outside the central 95% of the distribution. Our sample value of 54% is not unusual so we might be willing to accept it as entirely consistent with a true (population) value of 52%. In fact, given this interpretation of the central 95% interval, any value between 42% and 62% would not be surprising.

Even though we have used what might seem to be a

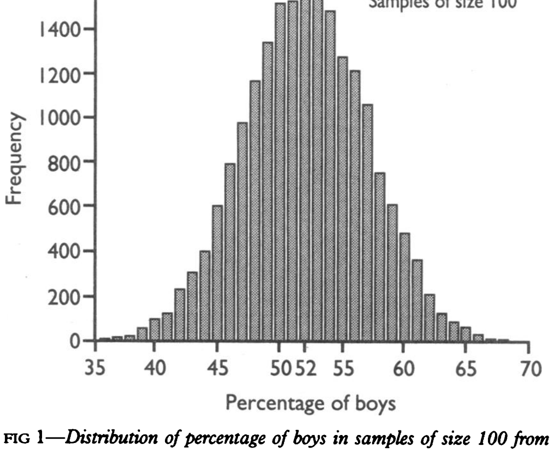

FIG 1—*Distribution of percentage of boys in samples of size 100 from population with true value of 52%*

Box 2

Measures of variability

- *Variance* is a measure of the variability of any characteristic which varies from sample to sample. It is calculated in squared units. Its square root is called either a standard deviation or standard error
- *Standard deviation* (SD) is a measure of how variable individual observations are within a sample. It is applied to continuously distributed variables such as height, weight, blood pressure, haemoglobin concentration and so on but not to percentages
- *Standard error* (SE) is a measure of how variable the mean or a percentage is from one sample to another. In normal distributions the range mean −1 SE to mean +1 SE includes 68% of the distribution and the mean − 1·96 SE to mean + 1·96 SE includes 95%. The SE of continuously distributed variables is equal to $SD/\sqrt{n}$, where n is the sample size.

fairly large sample of 100 observations the interval is rather wide and does not give much indication of the population value. The interval is critically dependent on sample size. Figure 3 shows the normal distribution for samples of size 500 superimposed on the distribution for samples of size 30. The small samples give an extremely wide interval (34% to 70%), so much so that any sample result will give little clue as to the likely value of the population mean. The larger samples have a much narrower interval (47·6% to 56·4%). With larger sample sizes the population value will be estimated with increasing precision. Figure 3 shows why a sample may be so small that it cannot provide a useful answer. Estimation of the required sample size before collection of data will go a long way towards avoiding this pitfall.

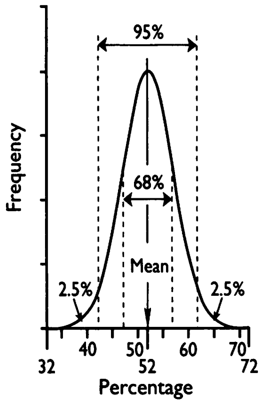

FIG 2—*Normal distribution with areas shown for central 68% and 95% and two tails with 2·5% of distribution*

EFFECT SIZE

The sample size determines the precision with which a population mean can be estimated. Few studies, however, set out simply to estimate population means or prevalence or incidence rates of disease; most test hypotheses that there are differences between groups of people who are exposed to different experiences. For example, consider a hypothetical trial of treatment for ovarian cancer. For an estimation of sample size we must have some idea of how big an effect of treatment can be expected based on previous experience or from published reports or what would be of scientific or clinical interest. The five-year mortality of women receiving conventional care is known to be 73% from experience in the hospital where the trial is to take place. A new drug is available for assessment. A reduction in mortality to 50% would be of considerable clinical interest. The postulated effect of the new drug is the difference between these mortalities (23%) and is known as the effect size. The question is: how big a sample is needed to be reasonably certain of distinguishing a true effect of this size from no effect at all? We are interested in detecting either an advantage or a disadvantage of the new treatment so must compare the effect size observed in the trial with zero effect. To test the hypothesis a controlled trial is used in which patients are randomly allocated to two groups, one receiving conventional care (controls) and one the new drug (the experimental group). The hypothesis may be rephrased to state that "there is no difference between the mortalities of the experimental group and the controls." Because the hypothesis is in terms of no difference it is known as the null hypothesis.

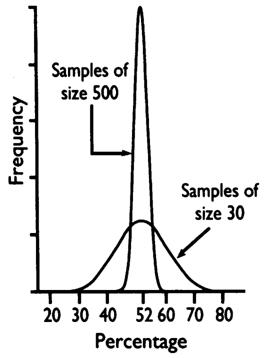

FIG 3—*Overlapping distributions of samples of sizes 30 and 500 to show effect on spread of observations with change in sample size*

In place of the distributions of percentages from single samples, as in figures 1-3, we are now concerned with the distributions of the differences between mortalities derived from pairs of samples. Figure 4 shows the distribution of the differences in mortalities obtained from pairs of samples of size 100 drawn from experimental and control groups with identical population mortalities. As the difference between the population mortalities must be zero the distribution is centred around this population value.

SIGNIFICANCE

The significance tells us how likely it is that an observed difference is due to chance when the true difference is zero. We can arbitrarily specify surprising results as those outside the 95% interval. In figure 4 such results would be those in which the differences between the experimental group and controls were equal to or greater than 13% (that is, in favour of treatment) or equal to or less than −13% (the negative value implies an increased mortality in the experimental group). In this trial 13% and −13% are the upper and lower critical values—if the observed result lay beyond them the null hypothesis would be rejected, even though this would mean taking a 5% chance of being wrong because the result was simply an extreme one. This error of rejecting the null hypothesis when it is true is known as a type I or α error. The error is the one usually referred to as the level of significance in reports of statistical analyses. In this example the conventional value of 5% for α has been chosen, but this is quite arbitrary. A more restrictive value might be taken, such as 1% or even 0·1% to reduce the possibility of accepting the treatment as effective when the result is simply due to chance.

POWER

The power tells us how likely we are to detect an effect for a given sample size, effect size, and level of significance. In the treatment trial we believe that an effect or change in mortality of 23% can be obtained. Thus our alternative hypothesis (alternative to the null) is that there is a true effect of 23%. The results from samples taken to estimate this value can be expected to be centred about the value 23%.

Figure 5 shows two distributions. On the left is the theoretical distribution for the null hypothesis which underlies figure 4. The upper critical value of 13% is marked by the vertical line. On the right is the distribution for the alternative hypothesis that the true effect size is 23%, similarly based on samples of size 100. The areas where the two curves overlap on either side of the upper critical value of 13% are also marked. Quite a few samples within the distribution for the alternative hypothesis lie in the shaded area, labelled β, where the null hypothesis would be accepted. In this example, about 10% of the alternative hypothesis samples would be accepted as being consistent with the null hypothesis. This is the converse of the type I error and is known as the type II or β error. It occurs when the null hypothesis is accepted when in fact it is wrong.

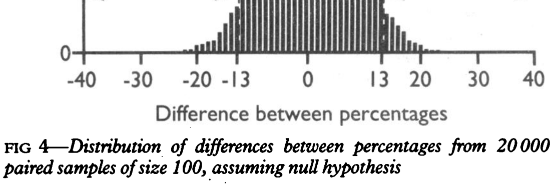

FIG 4—*Distribution of differences between percentages from 20 000 paired samples of size 100, assuming null hypothesis*

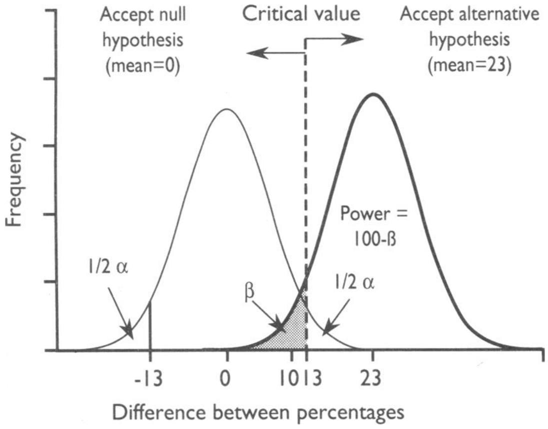

FIG 5—*Overlap between two distributions, assuming one with true value for difference between samples of zero and the other a true difference of 23/100*

The power of the sample to detect a true difference in mortality of 23% is the likelihood of a sample estimate occurring above the critical value. As 10% (β) of the distribution lies below this value 90% must lie above it. Thus there is a 90% (100−β) chance of detecting a true difference of 23% with samples of size 100 given a significance level of 5% (α).

We have already seen that the spread of the distribution of sample estimates can be altered by changing the number of observations used in calculating each estimate; the more observations the narrower the distribution (fig 3). Figure 6 shows simulated distributions of sample estimates of differences for samples of size 500 instead of 100. Both distributions are centred around the same population values as in figure 5, but the spread of sample values is now so much less that for all intents and purposes there is no overlap. If from a single pair of samples we obtained a difference of 13% or more it would be an extremely surprising result were the null hypothesis true as it would occur in only about 1 in 1000 samples. We would have no hesitation in rejecting the hypothesis. In fact with these large samples we should be able to detect as statistically important much smaller differences than 23%. Simply by increasing the number of observations we have improved our ability to discriminate between our null and alternative hypotheses.

Practice

BASIC CALCULATIONS

To calculate the sample size for a trial or a comparison between two groups with effects measured in percentages we must have the following information:

- The likely value for the outcome variable given conventional treatment and the effect size which will be of clinical or biological interest (in the trial example these are the likely death rate in the control group—73% and the effect size of interest—23%)
- The value for α (in the example we used the conventional 5%)
- The value for β (in the example we used 10%).

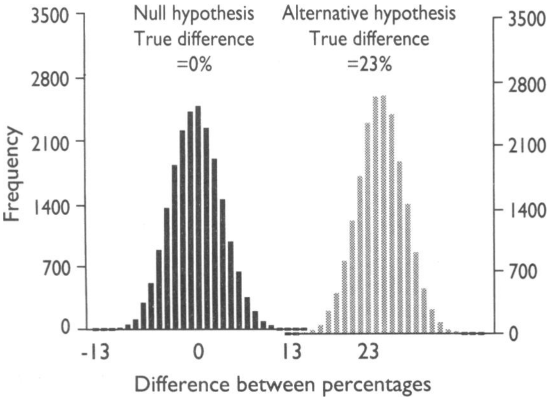

FIG 6—*Distributions of differences between percentages for sample size 500*

These choices will lead to rejection of the null hypothesis when it is true in 5% of samples and acceptance of the null hypothesis when it is not true in 10% of samples, given the assumed true difference of 23%. Other values commonly used are 1% for α and 20%, 5%, and 1% (giving powers of 80%, 95%, and 99% respectively) for β.

The critical value in figure 5 marks the upper limit of the 95% interval of the null distribution. Its position on the horizontal axis can be calculated from information both from the null distribution and from the alternative distribution.

Based on the properties of the normal distribution the critical value lies 1·96 standard errors above the true null value of 0 (zero). The standard error can be calculated from the formula:

$$SE_{diff}=\sqrt{(p_1q_1+p_2q_2)/n}$$

where p_1 is the percentage of subjects dying in the treated group and q_1 is the percentage surviving (equal to $100-p_1$) and similarly for p_2 and q_2 for the controls. The number of subjects in one of the groups is assumed equal to the number of subjects in the other groups and is represented by n. Thus the position of the critical value on the horizontal axis is:

$$\text{the true null value}+1{\cdot}96\times SE_{diff}$$
$$=0{\cdot}0+1{\cdot}96\sqrt{(p_1q_1+p_2q_2)/n}$$

The critical value can also be defined by using information about the alternative distribution. The standard error is calculated as above. When β=10%, as in the figure, the critical value lies 1·28 standard errors below the alternative true value of 23% (box 3).

As the critical value can be defined by two different expressions, the expressions must be equal to each other:

$$23-1{\cdot}28\sqrt{(p_1q_1+p_2q_2)/n}=0{\cdot}0+1{\cdot}96\sqrt{(p_1q_1+p_{2q_2})/n}$$

$$\therefore 23^2n=\left(1{\cdot}28\sqrt{(p_1q_1+p_2q_2)}+1{\cdot}96\sqrt{(p_1q_1+p_2q_2)}\right)^2$$

$$\therefore n=((1{\cdot}96+1{\cdot}28)^2(p_1q_1+p_2q_2))/23^2$$

Substituting our own figures into this equation we get:

$$p_1q_1+p_2q_2=50\times50+73\times27=4471$$
$$n=(1{\cdot}96+1{\cdot}28)^2(4471)/23^2$$
$$n=89 \text{ (to the nearest whole number)}$$

This equation is rather rudimentary and more sophisticated methods of estimation are available.[2-3] It shows, however, that by using the standard formulas we come up with more or less the same value as obtained from simulation—namely, about 100 observations in each of a pair of samples or a total study size of 200.

If you compare means of continuously distributed variables such as blood pressure, plasma glucose concentration, or height the calculation is slightly simpler. It requires the difference you think will be of interest between the means of the two groups (d), the likely standard deviation (s) of the variable (it should be the same for both groups), and the selected values of α and β. When α and β are 5% and 10% respectively, the formula for the sample size n is:

$$n=2(1{\cdot}96+1{\cdot}28)^2\,s^2/d^2$$

Box 3

Multipliers for conventional values of α

α	Multiplier
5%	1·96
1%	2·58

Multipliers for conventional values of β

β	Multipliers
20%	0·842
10%	1·28
5%	1·64
1%	2·33

The sample size is for each group so the total number of subjects in the study will be 2n.

This formula could be used, for example, in a comparison of blood pressures between control and treated groups. A reduction of 20 mm Hg would be of clinical interest. The standard deviation of diastolic blood pressure has been found to be 15 mm Hg in patients already measured in the hospital where the trial is to be carried out. The number of subjects required in each group, given that α and β have been chosen as 5% and 10% respectively, is:

$$2(1{\cdot}96+1{\cdot}28)\times 15^2/20^2=20{\cdot}995\times 225/400=12$$

This means that 12 controls and 12 patients treated with the new drug would be needed to have a 90% chance of detecting a true difference of 20 mm Hg or larger at the 5% level of significance.

Other values of α and β may be selected. the conventional values and their associated multipliers (standard normal variate values) are given in box 3.

What to put in a grant application

You should consider the following points and check whether you have included them in a paragraph devoted to the calculation of sample size.

Firstly, give the bare essentials for the calculation so that the reader can check your arithmetic. For differences in percentages the likely population value of p_1 and the effect size of interest should be given, from which p_2 and the standard error can be calculated. The variable you choose for the calculation should be the most important one to the interpretation of the study. Sometimes it is wise to do calculations for a number of outcome variables to see if they all give about the same sample size or if there are some variables which are more likely to yield precise results than others. For differences in means of continuously distributed variables the likely population mean of the variable of greatest interest for the controls, the effect size of interest, and the standard deviation for either the cases or controls should be given (the standard deviations for the two groups are assumed to be the same). Your choice of effect size and variance (standard error or deviation) should be justified—for example, because they have appeared in published reports or they make biological or clinical sense.

Secondly, state the selected values for α (usually 5% or 1%) and β (usually 20%, 10%, 5%, or rarely 1%). Justification should be given for your choice of the values of α and β, such as the unavailability of greater numbers of patients, the cost of increasing the sample size, the loss of control over quality of data if the sample size were increased, etc.

Finally, you should refer to the method of calculation of sample size.[2-9]

A word of caution

This discussion has barely touched on the entire subject of calculating sample size. There are many special circumstances when the formulas given here are inadequate or inappropriate.[8] Randomised controlled trials are better analysed by comparison of survival curves, an analysis which implies a somewhat different mathematical approach from the one I have given. Or the question asked might be whether the new treatment is not worse than the usual treatment because if it is no worse its other properties—lower cost for example—might make it preferable.[9] Crossover trials, trials with more than two treatments, and studies using ordinal data which require different statistical methods from those given above are among other topics to keep in mind when deciding exactly how to estimate sample size for the study in hand. These are not within the scope of this article. My aim has been to alert the novice to the underlying concepts and to suggest the very minimum of information to be offered to committees judging grant applications. The references point to further reading. I hope that the reader will appreciate the importance of joining forces with a statistician early in the planning of a study.

I thank Dr Iain Crombie, Dr Fiona Williams, Mr Huw Davies, and Mr Simon Ogston for their numerous very helpful suggestions for improving the manuscript. The manuscript was reviewed by many others whose ideas contributed to clarifying earlier drafts. I am grateful for their help and time.

1 Brown RA, Swanson Beck J. *Medical statistics on microcomputers.* London: BMJ, 1990.

2 Casagrande JT, Pike MC, Smith PG. The power function of the "exact" test for comparing two binomial distributions. *Applied Statistics* 1978;27:176-80.

3 Freedman L. Tables of the number of patients required in clinical trials using the logrank test. *Stat Med* 1982;1:121-9.

4 Lemeshow S, Hosmer DW, Klar J, Lwanga SK. *Adequacy of sampler size in health studies.* Chichester: John Wiley, 1990.

5 Gardner MJ, Altman DG. *Statistics with confidence—confidence intervals and statistical guidelines.* London: BMJ, 1989.

6 Daly L. Confidence intervals and sample sizes: don't throw out all your old sample size tables. *BMJ* 1991;302:333-6.

7 Miller DK, Homan SM. Graphical aid for determining power of clinical trials involving two groups. *BMJ* 1988;297:672-6.

8 Lachin JM. Introduction to sample size determination and power analysis for clinical trials. *Controlled Clin Trials* 1981;2:93-113.

9 Blackwelder WC. "Proving the null hypothesis" in clinical trials. *Controlled Clin Trials* 1982;3:345-53.

(Accepted 9 February 1993)

INSIGHTS

Tennyson looks at the psychological development of the infant

The baby new to earth and sky,
What time his tender palm is prest
Against the circle of the breast,
Has never thought that "this is I":

But as he grows he gathers much
And learns the use of "I" and "me,"
And finds "I am not what I see,
And other than the thing I touch."

So rounds he to a separate mind
From whence clear memory may begin,
As thro' the frame that binds him in
His isolation grows defined.

Alfred Lord Tennyson (1809-92), *In Memoriam*, section 45.