Mathematics Learning Support Centre

Statistics: An introduction to sample size calculations

Rosie Cornish. 2006.

1 Introduction

One crucial aspect of study design is deciding how big your sample should be. If you increase your sample size you increase the precision of your estimates, which means that, for any given estimate / size of effect, the greater the sample size the more "statistically significant" the result will be. In other words, if an investigation is too small then it will not detect results that are in fact important. Conversely, if a very large sample is used, even tiny deviations from the null hypothesis will be statistically significant, even if these are not, in fact, practically important. In practice, this means that before carrying out any investigation you should have an idea of what kind of change from the null hypothesis would be regarded as practically important. The smaller the difference you regard as important to detect, the greater the sample size required.

Factors such as time, cost, and how many subjects are actually available are constraints that often have to be taken account of when designing a study, but these should not dictate the sample size — there is no point in carrying out a study that is too small, only to come up with results that are inconclusive, since you will then need to carry out another study to confirm or refute your initial results.

There are two approaches to sample size calculations:

- **Precision-based**
  With what precision do you want to estimate the proportion, mean difference ... (or whatever it is you are measuring)?
- **Power-based**
  How small a difference is it important to detect and with what degree of certainty?

2 Precision-based sample size calculations

Suppose you want to be able to estimate your unknown parameter with a certain degree of precision. What you are essentially saying is that you want your confidence interval to be a certain width. In general a 95% confidence interval is given by the formula:

$$\text{Estimate} \pm 2(\text{approx})^{1} \times \text{SE}$$

where SE is the standard error of whatever you are estimating.

1. This is because 95% confidence intervals are usually based on the normal distribution or a t-distribution — for a normal distribution the value is 1.96; for t-distributions the value is generally just over 2.

The formula for any standard error always contains n, the sample size. Therefore, if you specify the width of the 95% confidence interval, you have a formula that you can solve to find n.

Example 1
Suppose you wish to carry out a trial of a new treatment for hypertension (high blood pressure) among men aged between 50 and 60. You randomly select $2n$ subjects. n of these receive the new treatment and n receive a the standard treatment, then you measure each subject's systolic blood pressure. You will analyse your data by comparing the mean blood pressure in the two groups — i.e. carrying out an unpaired t-test and calculating a 95% confidence interval for the true difference in means.

You would like your 95% confidence interval to have width 10 mmHg (i.e. you want to be 95% sure that the true difference in means is within $\pm$ 5 mmHg of your estimated difference in means. How many subjects will you need to include in your study?

We know that the 95% confidence interval for a difference in means is given by

$$(\bar{x}_1 - \bar{x}_2) \pm 2(\text{approx}) \times s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Hence, we want $2 \times s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ to be equal to 5 $\Rightarrow s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = s_p \sqrt{\frac{2}{n}} \approx 2.5$
(since we are aiming for groups of the same size).

In order to work out our sample sizes we therefore need to know what s_p is likely to be. This is either known from (a) previous experience (i.e. knowledge of the distribution of systolic blood pressure among men with hypertension in this age group), (b) using other published papers on blood pressure studies in a similar group of people or (c) carrying out a pilot study. I have used option (b) to get a likely value for s_p of 20 mmHg.

This gives

$$2.5 = 20\sqrt{\frac{2}{n}} \Rightarrow \frac{n}{2} = \left(\frac{20}{2.5}\right)^2 \Rightarrow n = 128 \qquad \text{(in each group)}$$

If you wanted your true difference in means to be within ± 2.5 mmHg rather than ± 5 mmHg of your estimate, this would become

$$\frac{n}{2} = \left(\frac{20}{1.25}\right)^2 \Rightarrow n = 512$$

i.e. if you want to increase your precision by a factor of 2, you have to increase your sample size by a factor of 4. In general, if you want to increase your precision by a factor k, you will need to increase your sample size by a factor k^2. This applies across the board — i.e. whether you are estimating a proportion, a mean, a difference in means, etc. etc.

Example 2
Supposing you are investigating a particular intervention to reduce the risk of malaria mortality among young children under the age of five in The Gambia, in West Africa. You know that the risk of dying from malaria in this age group is about 10% and you want the risk difference to be estimated to within $\pm 2\%$. A 95% confidence interval for a difference in proportions is

given by

$$(p_1 - p_2) \pm 1.96\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = (p_1 - p_2) \pm 1.96\sqrt{\frac{p_1(1-p_1) + p_2(1-p_2)}{n}}$$

if the sample size in each group is the same. As stated previously, we normall approximate 1.96 by 2. We therefore want

$$\sqrt{\frac{p_1(1-p_1) + p_2(1-p_2)}{n}} \approx 0.02/2 = 0.01$$

To work out the required sample size, we usually take $p_1 = p_2 =$ the value closer to 0.5, since this would give rise to a larger standard error and therefore a larger sample size (it is always better to err on the side of caution in sample size calculations because (a) you often get drop-outs, so it's better to have too many rather than too few in your sample to start with and (b) they are never 100% exact anyway, since you base them on estimates of the standard error, not on known values.

So, in this case we have

$$\sqrt{\frac{2(0.1)0.9)}{n}} = 0.01 \Rightarrow n = \frac{2(0.1)(0.9)}{0.01^2} = 1800 \qquad \text{(in each group)}$$

To summarise, in order to carry out any precision-based sample size calculation you need to decide how wide you want your confidence interval to be and you need to know the formula for the relevant standard error. Putting these together will give you a formula which can be rearranged to find n.

3 Power-based sample size calculations

We have seen above that precision-based sample size calculations relate to estimation. Power-based sample size calculations, on the other hand, relate to hypothesis testing. In this handout, the formulae for power-based sample size calculations will not be derived, just presented.

Definitions

Type I error (false positive)
Concluding that there is an effect (e.g. that two treatments differ) when they do not
α = P(type I error) = level of statistical significance $\quad [= P(\text{reject } H_0 \mid H_0 \text{ true})]$

Type II error (false negative)
Concluding that there is NO effect (e.g. that there is no difference between treatments) when there actually is.
β = P(type II error) $\quad [= P(\text{accept } H_0 \mid H_1 \text{ true})]$

Power
The (statistical) power of a trial is defined to be $1 - \beta$ $\quad [= P(\text{reject } H_0 \mid H_1 \text{ true})]$

3.1 Power calculations: quantitative data

Suppose you want to compare the mean in one group to the mean in another (i.e. carry out an unpaired t-test). The number, n, required in each group is given by

$$\boldsymbol{n = f(\alpha, \beta) \cdot \frac{2s^2}{\delta^2}}$$

Where:
α is the significance level (using a two-sided test) — i.e. your cut-off for regarding the result as statistically significant.
$1 - \beta$ is the power of your test.
$f(\alpha, \beta)$ is a value calculated from α and β — see table below.
δ is the smallest difference in means that you regard as being important to be able to detect.
s is the standard deviation of whatever it is we're measuring — this will need to be estimated from previous studies.

$f(\alpha, \beta)$ for the most commonly used values for α and β

	β			
α	0.05	0.1	0.2	0.5
0.05	13.0	10.5	7.9	3.8
0.01	17.8	14.9	11.7	6.6

Example

Returning to the blood pressure example. Suppose we want to be 90% sure of detecting a difference in mean blood pressure of 10 mmHg as significant at the 5% level (i.e. power = 0.9, $\beta = 0.1$, $\alpha = 0.05$). We have, from above, $s = 20$ mmHg. Using the table, we get $f(\alpha, \beta) = 10.5$. This gives

$$n = f(\alpha, \beta) \cdot \frac{2s^2}{\delta^2} = 10.5 \cdot \frac{2(20)^2}{10^2} = 84$$

You would need 84 subjects in each group.

Obviously, if you increase the power or want to use a lower value for α as your cut-off for statistical significance, you will need to increase the sample size.

3.2 Power calculations: categorical data

Suppose we are comparing a binary outcome in two groups of size n. Let
$p_1 =$ proportion of events (deaths/responses/recoveries etc.) in one group
$p_2 =$ proportion of events in the other group

We need to choose a value for $p_1 - p_2$, the smallest practically important difference in proportions that we would like to detect (as significant). We also need to have some estimate of the proportion of events expected. This can often be obtained from routinely collected data or previous studies.

The number of subjects required for each group is given by

$$\boldsymbol{n = \frac{p_1(1-p_1) + p_2(1-p_2)}{(p_1 - p_2)^2} \cdot f(\alpha, \beta)}$$

Example

A new treatment has been developed for patients who've had a heart attack. It is known that 10% of people who've suffered from a heart attack die within one year. It is thought that a reduction in deaths from 10% to 5% would be clinically important to detect. Again we will use $\alpha = 0.05$ and $\beta = 0.1$. We have p_1 = proportion of deaths in placebo group = 0.1, p_2 = proportion of deaths in treatment group = 0.05. This gives

$$n = \frac{0.1(0.9) + 0.05(0.95)}{(0.1 - 0.05)^2} \cdot 10.5 = 578$$

Thus, 578 patients would be needed in each treatment group to be 90% sure of being able to detect a reduction in mortality of 5% as significant at the 5% level.

The size of difference regarded as clinically important to be able to detect has a strong effect on the required sample size. For example, to detect a reduction from 10% to 8% as significant (using the same α and β as above) 4295 patients would be needed in each treatment group. This may be clinically reasonable (since any reduction in mortality would probably be regarded as important) but perhaps too expensive. Conversely, to detect a reduction from 10% to 1% as significant, 130 patients would be needed in each group. This kind of reduction would perhaps be over-optimistic and therefore a trial of this size would be unlikely to be conclusive.

4 Further notes

1. Sample size formulae in other situations
   In this handout I have only presented sample size calculations for certain types of analysis. Similar formulae can be obtained for other types of analysis by reference to appropriate texts.

2. Increasing power or precision without having to increase sample size
   As indicated previously, certain constraints (cost etc.) may mean that the required sample size is impossible to achieve. In these situations it may be possible to increase precision or power by changing the design of the study. For example, using a paired design can often be more efficient since you are considering individual (within-subject) differences rather than between-subject differences; the former are usually less variable than the latter (i.e. the standard error is lower) and therefore the sample size required to detect a given difference will be lower.

3. Allowing for non-response / withdrawals
   In most studies, particularly those involving humans, there is likely to be a certain amount of data "lost" (or never gathered) from the original sample. This could be for a variety of different reasons: non-response (e.g. to a survey); subjects deliberately withdrawing from a study or getting "lost" in some other way (e.g. cannot be traced); subjects in a clinical trial not following their allocated treatment; or missing data (e.g. on a questionnaire). Allowance should be made for this when determining the sample size — i.e. the sample size should be increased accordingly. The extent to which this is needed should be guided by previous experience or a pilot study.