**British Standards Institution Study Day**

# Sample size for clinical trials

Martin Bland
Prof. of Health Statistics
University of York
http://martinbland.co.uk

## Outcome variables for trials

An outcome variable is one which we hope to change, predict or estimate in a trial.

Examples:

- Systolic blood pressure in a hypertension trial
- Caesarean section rate in an obstetric trial
- Survival time in a cancer trial

How many outcome variables should I have?

If we have many outcome variables:

- all possibilities would be covered,
- we would be less likely to miss something,
- the risk of false positives, finding an apparent effect when there is none in reality, would be increased.

If we have few outcome variables:

- the trial would be easier and quicker to check and analyse,
- the trial would be cheaper,
- the trial would be easier for research subjects,
- we would avoid multiple testing and the high risk of a false positive result.

We get round the problem of multiple testing by having one outcome variable on which the main conclusion stands or falls, the primary outcome variable. If we do not find an effect for this variable, the study has a negative result.

Usually we have several secondary outcome variables, to answer secondary questions. A significant difference for one of these would generate further questions to investigate rather than provide clear evidence for a treatment effect.

The primary outcome variable must relate to the main aim of the study. Choose one and stick to it.

## How large a sample should I take?

A significance test for comparing two means is more likely to detect a large difference between two populations than a small one.

The probability that a test will produce a significant difference at a given significance level is called the **power** of the test.

The power of a test is related to:

- the postulated difference in the population,
- the standard error of the sample difference (which depends on the sample size)
- the significance level, usually 0.05, chosen in advance. .

Relationship between:

- power of the test, $P$
- postulated difference in the population, $\delta$ (delta)
- standard error of the sample difference, SE($d$)
- significance level, $\alpha$ (alpha)

These quantities are connected by an equation: $\delta^2 = f(\alpha, P)\mathrm{SE}(d)^2$. If we know three of these quantities we can calculate the fourth.

The function $f(\alpha, P)$ depends on power and significance level only. The following table shows values of $f(\alpha,P)$ for different $P$ and $\alpha$:

| | $\alpha$ | |
|---|---|---|
| $P$ | 0.05 | 0.01 |
| 0.50 | 3.8 | 6.6 |
| 0.70 | 6.2 | 9.6 |
| 0.80 | 7.9 | 11.7 |
| 0.90 | 10.5 | 14.9 |
| 0.95 | 15.2 | 20.4 |
| 0.99 | 18.4 | 24.0 |

Usually we choose the significant level, $\alpha$, to be 0.05 and the power, $P$, to be 0.80 or 0.90, so the numbers we actually use from this table are 7.9 and 10.5. We can then choose the difference we want the trial to detect, $\delta$, and from this work out what sample size we need.

## Example: trial to reduce blood pressure.

Suppose we want to compare two treatments designed to reduce blood pressure.. We decide that a clinically important difference would be 10 mm Hg.

Where does this difference come from? It could be

- Clinical judgement as to what would be important. This might come from a focus group of clinicians or patients, for example.
- What the treatment might achieve. This might come from pilot studies, or other trials of similar treatments.
- Back calculation from what is feasible. We start with the sample size we think we can get, then work from there back to the difference this sample size could detect. This is rightly frowned on by referees.

SE($d$) depends on sample size ***and*** variability

SE($d$) depends on the particular sample size problem.

## Comparison of two means

Compare the means of two samples, sample sizes $n_1$ and $n_2$, from populations with means $\mu_1$ and $\mu_2$, with the variance of the measurements being $\sigma^2$. We have $\delta = \mu_1 - \mu_2$ and

$$\mathrm{SE}(d) = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$$

so the equation becomes:

$$(\mu_1 - \mu_2)^2 = f(a, P)\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

For equal sized groups, $n_1 = n_2 = n$, the equation becomes:

$$(\mu_1 - \mu_2)^2 = f(a, P)\frac{2\sigma^2}{n}$$

For example, consider a hypothetical trial to reduce blood pressure. The primary outcome will be fall in systolic blood pressure. We decide that a clinically important difference would be 10 mm Hg. From a pilot study, the standard deviation of the difference between successive systolic readings = 16 mm Hg. We choose power $P$ = 0.90 = 90%, $\alpha$ = 0.05 =5%, an arbitrary but conventional choice. Then $f(\alpha,P)$ = 10.5.

We want to detect difference $\mu 1 - \mu 2$ = 10 mm Hg.

$$2^2 = 10.5 \times \frac{2 \times 7^2}{n}$$

$$n = 10.5 \times \frac{2 \times 7^2}{2^2} = 257.25$$

Hence we need 258 patients in each group.

That’s the hard way. We can use:

- Software, e.g. nQuery Advisor
- Graphics, e.g. Altman’s nomogram (Altman 1991)
- Tables, e.g. Machin et al. (1998) *Statistical Tables for the Design of Clinical Studies, Second Edition*

Software is now the usual choice so we shall stick to that here.

## Software

There are several programs available. I often use “PS: Power and Sample Size”, which is a free program by William Dupont and Walton Plummer, Jr. There are many available on the World Wide Web. PS can be downloaded from www.mc.vanderbilt.edu/prevmed/ps/.

If you download and start PS, you can click continue to get into the program proper. To compare two means, click “t test”. For “What do you want to know” click “Sample size”. For “Paired or independent” click “Independent”. Now put in α = 0.05, power = 0.90, δ = 10, and sigma = 16. The rather mysterious “m” is the number of subjects in the second group for each subject in the first group, in enables you to have unequal group sizes. Put in “1”. Now click “Calculate” and you should see the sample size per group, 55.

The reason PS gets a larger sample size than my calculation is that it allows for the degrees of freedom in a t test. My formula is for a large sample Normal test and so does not allow for this. The smaller the indicated sample is, the further apart the formula and PS will be. PS is better.

Not too hard, is it, really? You can also put in the difference and sample size and calculate the power, or the sample size and power and calculate the difference you could detect. This is the method I usually use for sample size calculations.

## Other sample size considerations:

The decisions about difference to be detected and power, then the actual calculations, are not the only problems with sample size. Having estimated it, we might then find that

- Eligible patients disappear like snow in August.
- Eligible patients may refuse.
- Clinicians may forget or refuse to enrol eligible patients.
- Patients may drop out.

We often allow for this by increasing the sample size. Inflations between 10% and 20% are popular. We should also make sure that the new sample size is much smaller than the predicted number of eligible patients. Getting patients into clinical trials is NOT easy.

## Comparing two proportions:

This is the most frequent sample size calculation. We have two proportions *p*1 and *p*2. Unlike the comparison of means, there is no standard deviation. SE(*d*) depends on *p*1 and *p*2.

The equation becomes

$$(p_1 - p_2)^2 = f(\alpha, P)\left(\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

There are several different variations on this formula. Different books, tables, or software may give slightly different results. If we have two equal groups, it becomes:

$$(p_1 - p_2)^2 = f(\alpha, P)\left(\frac{p_1(1-p_1) + p_2(1-p_2)}{n}\right)$$

For an example, consider a trial of an intervention to reduce the proportion of births by Caesarean section. We will take the usual power and sample size $P = 0.90$, $\alpha = 0.05$, so $f(\alpha, P) = 10.5$.

From clinical records we observe 24% of births are by Caesarean section and decide that a reduction to 20% would be of clinical interest. We have $p_1 = 0.24$, $p_2 = 0.20$.

The calculation proceeds as follows:

$$(0.24-0.20)^2 = 10.5\times\left(\frac{0.24(1-0.24)+0.20(1-0.20)}{n}\right)$$

$$n = 10.5\times\left(\frac{0.24(1-0.24)+0.20(1-0.20)}{(0.24-0.20)^2}\right) = 2247$$

So $n$ = 2,247 in each group.

Detecting small differences between proportions requires a very large sample size.

We can do the same calculation using PS. We choose “dichotomous”, because our outcome variable, Caesarean section, is yes or no. We then proceed as for the comparison of two means, for “What do you want to know click “Sample size”, for “Paired or independent” click “Independent”. We now have a new question “Case control?”, to which we answer “Prospective”, because a clinical trial is a prospective study. The alternative hypothesis is expressed as two proportions. To correspond to the large sample formula above, the testing method would be uncorrected chi-squared. You could choose “Fisher’s exact test” if you were feeling cautious. There is a lot of statistical argument about this choice and we will not go into it here. Now put in $\alpha$ = 0.05, power = 0.90, $p_1$ = 0.24, $p_2$ = 0.20, and m = 1. Click “Calculate” and you should see the sample size per group, 2253. Not quite the same as my formula gives, but very similar.

## Expressing differences between two proportions

As percentages, the proportions are $p_1$ = 24%, $p_2$ = 20%, so the difference = 4. Are we looking for a reduction of 4% in Caesarean sections? No we are not! A reduction of 4% would be 4% of 24% = 4×24/100 = 0.96, i.e. from 24% down to 23.04%. The reduction from 24% to 20% is 4 percentage points, **NOT** 4%. This is important, as I see trials described as aiming to detect a reduction of 20% in mortality, when this is from 30% down to 10%. That is a reduction of 20 percentage points, or 67%.

### Other factors affecting power

Power may be *increased* by adjustment for baseline and prognostic variables.

Power may be *reduced* by cluster randomisation.

If in doubt, consult a statistician.

## Allowing for adjustment

Power may be increased by adjustment for baseline and prognostic variables. We need to know the reduction in the standard deviation produced by the adjustment. Researchers often simply say: “Power will be increased by adjustment for . . .”, so that things will actually be better than they estimate, but they cannot say by how much.

However, we can estimate it if we know the strength of the relationship between the outcome and the adjustment variables. The proportion of variation explained by regression = $r^2$. The standard deviation after regression is $\sigma\sqrt{}(1 - r^2)$.

For example, suppose we want to do a trial of a therapy programme for the management of depression. We will measure depression using the PHQ9 scale, 0 to 27, high score meaning depression. From an existing trial, we know that people identified with depression in primary care had a baseline PHQ9 score with mean = 18 and SD = 5. After four months of being given treatment as usual, they had mean 13, SD = 7. The correlation between baseline and outcome PHQ9 score was $r = 0.42$.

We want to detect a difference in mean PHQ9 = 2 points. The standard deviation of PHQ9 after regression will be $\sigma\sqrt{}(1 - r^2) = 7\times\sqrt{}(1- 0.42^2) = 6.35$.

We want to design a trial to detect a difference in mean PHQ9 = 2 points. We choose power = 0.90, significance level = 0.05. With unadjusted SD = 7, this would require $n = 258$ per group. With adjusted SD = 6.35, this would require $n = 213$ per group.

If we have a good idea of the reduction in the variability that adjustment will produce, we can use this to reduce the required sample size. The effect is not usually very great.

For example, to halve the required sample size, we must halve the standard deviation so we must have $\sqrt{}(1 - r^2) = ½ \Rightarrow r^2 = ¾ \Rightarrow r = 0.87$. 0.87 is a pretty big correlation coefficient.

## Confidence intervals

There has been a movement to present results of trials in the form of confidence intervals rather than P values (Gardner and Altman 1986). This was motivated by the difficulties of interpreting significance tests, particularly when the result was not significant. This campaign was very successful and many major medical journals changed their instructions to authors to say that confidence intervals would be the preferred or even required method of presentation. This was later endorsed by the wide acceptance of the Consort standard for the presentation of clinical trials (CONSORT). We insist on interval estimates and rightly so.

If we ask researchers to design studies the results of which will be presented as confidence intervals, rather than significance tests, I think that we should base our sample size calculations on confidence intervals, rather than significance tests. It is inconsistent to say that we insist on the analysis using confidence intervals but the sample size should be decided using significance tests (Bland 2009).

This is not difficult to do. For example, the International Carotid Stenting Study (ICSS) was designed to compare angioplasty and stenting with surgical vein transplantation for stenosis of carotid arteries, to reduce the risk of stroke. We did not anticipate that angioplasty would be superior to surgery in risk reduction, but that it would be similar in effect. The primary outcome variable was to be death or disabling ipsilateral stroke after three years follow-up. There was to be an additional safety outcome of death, stroke, or myocardial infarction within 30 days and a comparison of cost. The sample size calculations for ICSS were based on the earlier CAVATAS study (CAVATAS investigators 2001), which had the 3 year rate for ipsilateral stroke lasting more than 7 days = 14%. The one year rate was 11%, so most events were within the first year. There was very little difference between the treatment arms. The width of the confidence interval for the difference between two very similar percentages is given by

$$\text{observed difference} \pm 1.96\sqrt{2p(100-p)/n}$$

where $n$ is the number in each group and $p$ is the percentage expected to experience the event. If we put $p = 14\%$, we can calculate this for different sample sizes, as shown in the Table. Similar calculations were done for other dichotomous outcomes.

For health economic measures, the difference is best measured in terms of standard deviations. The width of the confidence interval is expected to be

$$\text{observed difference} \pm 1.96\sigma\sqrt{2/n}$$

where $n$ is the number in each treatment group and $\sigma$ is the standard deviation of the economic indicator.

If we put $n = 740$, we can calculate this for the chosen sample size: $\pm 1.96\sigma\ \surd(2/750) = \pm 0.10\sigma$. This was thought to be ample for cost data and any other continuous variables.

These calculations were subsequently amended slightly as outcome definitions were modified. This is the sample size account in the protocol:

> ‘The planned sample size is 1,500. We do not anticipate any large difference in the principal outcome between surgery and stenting. We propose to estimate this difference and present a confidence interval for difference in 30-day death, stroke or myocardial infarction and for 3-year survival free of disabling stroke or death. For 1,500 patients, the 95% confidence interval will be the observed difference ± 3.0 percentage points for the outcome measure of 30-day stroke, myocardial infarction and death rate and ± 3.3 percentage points for the outcome measure of death or disabling stroke over 3 years of follow-up. However, the trial will have the power to detect major differences in the risks of the two procedures, for example if stenting proves to be much more risky than surgery or associated with more symptomatic restenosis. The differences detectable with a power of 80% are 4.7 percentage points for 30-day outcome and 5.1 percentage points for survival free of disabling stroke. Similar differences are detectable for secondary outcomes.’ (Featherstone *et al.* 2004)

Despite my best attempts, power calculations could not be excluded completely. However, the main sample size calculation was based on a confidence interval and the study was funded.

## References:

Altman, D.G. (1991) *Practical Statistics for Medical Research.* Chapman and Hall, London. (nomogram)

Bland JM. (2009) The tyranny of power: is there a better way to calculate sample size? *BMJ* 2009; **339**: b3985.

Bland M. (2000) *An Introduction to Medical Statistics.* Oxford University Press.

CAVATAS investigators (2001). Endovascular versus surgical treatment in patients with carotid stenosis in the Carotid and Vertebral Artery Transluminal Angioplasty study (CAVATAS): a randomised trial. *Lancet* **357**: 1729-37.

CONSORT. http://www.consort-statement.org/Featherstone RL, Brown MM, Coward LJ. International Carotid Stenting Study: Protocol for a randomised clinical trial comparing carotid stenting with endarterectomy in symptomatic carotid artery stenosis. *Cerebrovasc Dis* 2004; 18: 69-74.

Gardner MJ and Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J* 1986; 292: 746-50.

Machin, D., Campbell, M.J., Fayers, P., Pinol, A. (1998) *Statistical Tables for the Design of Clinical Studies, Second Edition.* Blackwell, Oxford.

nQuery Advisor, www.statsol.ie/nquery/nquery.htm
(commercial)

PS, www.mc.vanderbilt.edu/prevmed/ps/
(free Windows program)