Sample size

Martin Bland

Emeritus Prof. of Health Statistics
University of York

http://martinbland.co.uk/

Outcome variables

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

Examples:

- Systolic blood pressure in a hypertension trial
- Cæsarean section rate in an obstetric trial
- Survival time in a cancer trial
- Presence of asthma in a respiratory disease risk study

Number of outcome variables

Should I have few or many?

Many outcome variables

- Cover all possibilities
- Less likely to miss something
- Risk of false positives

Few outcome variables

- Easier and quick to check and analyse
- Cheaper
- Easier for research subjects
- Avoid multiple testing

Primary and secondary outcome variables

Get round the problem 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 several secondary outcome variables, to answer secondary questions.

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 $\alpha = 0.05$.

Relationship between:

- power of the test, P
- postulated difference in the population, δ
- standard error of the sample difference, $\mathrm{SE}(d)$
- significance level, α

$$\delta^2 = f(\alpha, P)\mathrm{SE}(d)^2$$

If we know three of these we can calculate the fourth.

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

Relationship between:

- power of the test, P
- postulated difference in the population, δ
- standard error of the sample difference, $\mathrm{SE}(d)$
- significance level, α

We choose δ

$\mathrm{SE}(d)$ depends on sample size ***and*** variability

$$\delta^2 = f(\alpha, P)\mathrm{SE}(d)^2$$

$f(\alpha, P)$ depends on power and significance level only

If we know three of these we can calculate the fourth.

$\mathrm{SE}(d)$ depends on the particular sample size problem.

$$\delta^2 = f(P, \alpha)\mathrm{SE}(d)^2$$

P = power of the test, α = significance level.

Values of $f(\alpha,P)$ for different P and α

	α	
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

Example: trial to reduce blood pressure.

Decide a clinically important difference would be 10 mm Hg.

Where does this come from?

- Clinical judgement as to what would be important – focus group of clinicians.
- What the treatment might achieve – pilot studies, other trials of similar treatments.
- Back calculation from what is feasible – frowned on by referees.

Relationship between:

- power of the test, P
- postulated difference in the population, δ
- standard error of the sample difference, $\mathrm{SE}(d)$
- significance level, α

We choose δ

$\mathrm{SE}(d)$ depends on sample size ***and*** variability

$$\delta^2 = f(\alpha, P)\mathrm{SE}(d)^2$$

$f(\alpha, P)$ depends on power and significance level only

If we know three of these we can calculate the fourth.

Comparison of two means

Compare the means of two samples, sample sizes n_1 and n_2, from populations with means μ_1 and μ_2, with the variance of the measurements being σ^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:

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

Comparison of two means

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

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

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

Example: trial of an intervention for depression in primary care

Primary outcome: PHQ9 depression score after 4 months. PHQ9 scale from 0 to 27, high score = depressed.

Decide a clinically important difference would be 2 points.

Pilot study: standard deviation of PHQ9 after treatment = 7

Choose power P = 0.90 = 90%, α = 0.05 =5%.

Choose power P = 0.90 = 90%, α = 0.05 = 5%:

$f(\alpha,P)$

	α	
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

Detect difference $\mu_1 - \mu_2$ = 2 points on PHQ9.

Pilot study: σ = 7.

P = 0.90 = 90%, α = 0.05 =5%, $f(\alpha,P)$ = 10.5

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

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

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

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

Tables,

e.g. Machin, et al. (1998) *Statistical Tables for the Design of Clinical Studies, Second Edition*

nQuery Advisor, http://www.statistical-solutions-software.com/
(commercial)

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

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

PS,
http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize
(free Windows program)

Effect size

Effect size or standardized difference is the difference in standard deviations.

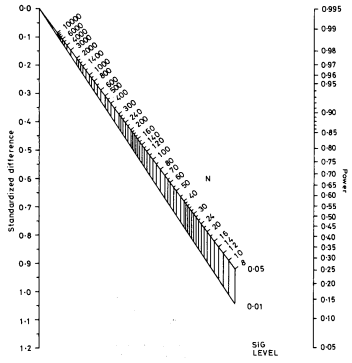
Difference divided by the standard deviation.

Example: depression measured by PHQ9.

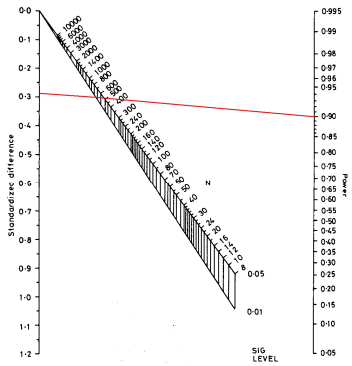
Difference to be detected = 2, SD = 7.

Effect size = 2/7 = 0.286

Altman's nomogram



Standardized difference = 0.286, $P = 0.90$, $\alpha = 0.05$



Total sample size N = 520.

Gives $n = 260$ in each group, as before.

Other sample size considerations:

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

Allow for this by increasing the sample size and making sure that the new sample size is much smaller than the predicted number of eligible patients.

Comparing two proportions:

Proportions p_1 and p_2. SE(d) depends on p_1 and p_2.

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

Several different variations on this formula. Different books, tables, or software may give slightly different results.

Two equal groups:

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

Example: Cæsarean section.

P = 0.90, α = 0.05, so $f(\alpha,P)$ = 10.5.

From clinical records we observe 24%. A reduction to 20% would be of clinical interest.

We have p_1 = 0.24, p_2 = 0.20.

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

$$(0.24 - 0.20)^2 = 10.5 \times \left(\frac{0.24(1-0.24) + 0.20(1-020)}{n}\right)$$

$$n = 10.5 \times \left(\frac{0.24(1-0.24) + 0.20(1-020)}{(0.24 - 0.20)^2}\right) = 2247$$

Example: Cæsarean section.

P = 0.90, α = 0.05, p_1 = 0.24, p_2 = 0.20.

$$n = 10.5 \times \left(\frac{0.24(1-0.24) + 0.20(1-020)}{(0.24 - 0.20)^2}\right) = 2247$$

So n = 2,247 in each group.

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

Using PS . . .

Expressing differences between two proportions

p_1 = 24%, p_2 = 20%, so difference = 4.

Are we looking for a reduction of 4% in Cæsarean sections?

NO.

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%.

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 be better than they estimate, but they cannot say by how much.

Allowing for adjustment

Proportion of variation explained by regression = r^2.

Standard deviation after regression is $\sigma\sqrt{(1-r^2)}$.

Example:

We want to do a trial of a therapy programme for the management of depression.

Measure depression using the PHQ9 scale, 0 to 27, high score = depression.

From an existing trial we know that people identified with depression in primary care and given treatment as usual have baseline PHQ9 score with mean = 18 and SD = 5. After four months they had mean 13, SD = 7.

We want to detect a difference in mean PHQ9 = 2 points.

Allowing for adjustment

Standard deviation after regression is $\sigma\sqrt{(1-r^2)}$.

Example:

We want to design a trial to detect a difference in mean PHQ9 = 2 points.

Power = 0.90, significance level = 0.05, difference = 2, SD = 7: n = 258 per group.

We also know $r = 0.42$.

Standard deviation after regression = $\sigma\sqrt{(1-r^2)}$ $= 7\times\sqrt{(1-0.42^2)} = 6.35$.

Power = 0.90, significance level = 0.05, difference = 2, SD = 7: n = 213 per group.

Allowing for adjustment

If we have a good idea of the reduction in the variability that reduction 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 have $(1-r^2) = ½ \rightarrow r^2 = ½ \rightarrow r = 0.71$.

0.71 is a pretty big correlation coefficient.

Confidence intervals

Movement to present results of trials in the form of confidence intervals rather than P values.

Motivated by the difficulties of interpreting significance tests, particularly when the result was not significant.

Major journals changed their instructions to authors to say that confidence intervals would be the preferred or even required method of presentation.

Endorsed by the wide acceptance of the Consort standard for the presentation of clinical trials.

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

Confidence intervals

We ask researchers to design studies the results of which will be presented as confidence intervals, rather than significance tests.

We should base our sample size calculations on confidence intervals, rather than significance tests.

How do we do this?

We need a formula for the confidence interval for the treatment difference in terms of the expected parameters and sample size.

We must decide how precisely we want to estimate the treatment difference.

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

Confidence intervals

For a large study, the 95% confidence interval will be 1.98 standard errors on either side of the observed difference.

For a trial with equal sized samples, the 95% confidence interval for the difference between two means will be

$$\pm 1.96\sigma\sqrt{(2/n)}$$

and for two proportions it will be

$$\pm 1.96\sqrt{(p_1(1-p_1)/n + p_2(1-p_2)/n)}$$

Confidence intervals

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.

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.

Confidence intervals

The sample size calculations for ICSS were based on the earlier CAVATAS study, which had the 3 year rate for death or ipsilateral stroke lasting more than 7 days = 14%.

This was an equivalence trial, no difference was anticipated, $p_1 = p_2 = 0.14$.

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

Confidence intervals

For two proportions, width of the 95% confidence interval for the difference =

$$\pm 1.96\sqrt{(p_1(1-p_1)/n + p_2(1-p_2)/n)}$$

If we put $p = 0.14$, we can calculate this for different sample sizes:

Sample size	Width of 95% confidence interval
250	±0.061
500	±0.043
750	±0.035
1000	±0.030

We chose 750 in each group.

Confidence intervals

For two means, width of the 95% confidence interval for the difference = $\pm 1.96\sigma \sqrt{(2/n)}$.

If we put $n = 740$, we can calculate this for the chosen sample size: $\pm 1.96\sigma \sqrt{(2/750)} = \pm 0.10\sigma$.

This was thought to be ample for cost data and any other continuous variables.