Sample size and power calculations made simple

Evie McCrum-Gardner

Background: *Sample size estimation is an important concern for researchers as guidelines must be adhered to for ethics committees, grant applications and publications. Studies may be underpowered (too few participants) or overpowered (too many participants) so it is important to achieve the correct balance.*

Content: *The process of sample size calculations, including relevant definitions, is explained and clear examples for different study designs are provided for illustration. A range of software packages and websites are discussed and evaluated.*

Conclusions: *The information regarding sample sizes and statistical programs should be useful for researchers in order to perform sample size calculations for research projects.*

Key words: ■ minimal clinically important difference ■ power ■ sample size ■ significance level ■ type I error ■ type II error

Submitted 10 July 2009, sent back for revisions 23 September 2009; accepted for publication following double-blind peer review 2 November 2009

Sample size estimation is an important concern for researchers undertaking research projects, but is often misunderstood or even ignored completely. Guidelines must be adhered to for ethics committees, grant applications and publications. Studies may be underpowered (too few participants) or overpowered (too many participants) so it is important to achieve the correct balance. The aim of this article is to explain the process of sample size calculations, including the purpose, relevant definitions and clear examples for different study designs. It will also discuss and evaluate a range of software packages and websites that are currently accessible, but which were not available at the time of earlier publications (e.g. Florey, 1993).

It is not assumed that the reader has previous experience of statistical tests. However, the choice of the correct statistical test is vitally important at the statistical analysis stage. When choosing the appropriate statistical test, the first step is to decide what scale of measurement your data is, as this will affect your decision (nominal, e.g. gender; ordinal, e.g. Likert scale; interval/ratio, e.g. weight). The next stage is to consider the analysis required – for example, are you comparing independent or paired groups? If an incorrect test is used, then invalid results and misleading conclusions may be drawn from the study. This is described in detail elsewhere (e.g. McCrum-Gardner, 2008).

Evie McCrum-Gardner *is Lecturer in Health Statistics, Schools of Nursing and Health Sciences, Health and Rehabilitation Sciences Research Institute, University of Ulster, Newtownabbey, Northern Ireland*

Correspondence to: E McCrum-Gardner Email: ee.gardner@ulster.ac.uk

RESEARCH ETHICS COMMITTEES GUIDELINES

Guidance on sample size by the Central Office for Research Ethics Committees (COREC) (2007) requires that 'the number should be sufficient to achieve worthwhile results, but should not be so high as to involve unnecessary recruitment and burdens for participants'.

SAMPLE SIZE CALCULATION

Sample size is a function of three factors – the significance level, power, and magnitude of the difference (effect size). These three factors are defined below with further definitions also provided by Devane et al (2004). Before explaining these factors, hypothesis testing is described.

Hypothesis testing

Researchers begin with a research hypothesis, for example, that treatment A is better than treatment B. Hypothesis testing involves expressing this as a null hypothesis and performing the appropriate statistical test to investigate whether the null hypothesis can be accepted/rejected. An example of a null hypothesis is: 'there is no difference (in the mean outcome measure) between treatments A and B'. The researcher wants to be able to reject the null hypothesis and to show that differences in the treatment outcomes are not a result of chance, i.e. that the alternative hypothesis is that there is a difference between the two treatments.

Significance level (*P* value)

Significance level is the probability cut-off (usually 0.05 or 5%) used. It is chosen in advance of performing the test, and the cut-off level depends on how much safeguard is required against accidentally rejecting the null hypothesis when it is in fact true.

Power

Power is the probability of rejecting the null hypothesis when the alternative hypothesis is true. It measures the ability of a test to reject the null hypothesis when it should be rejected. At a given significance level, the power of the test is increased by having a larger sample size. The minimum accepted level is considered to be 80%, which means there is an eight in ten chance of detecting a difference of the specified effect size.

Type I and type II errors

Performing the appropriate statistical test (McCrum-Gardner, 2008) will result in either rejecting or not rejecting the null hypothesis. The definitions of type I and II errors are given in *Table 1*.

Type I errors occur if the null hypothesis is rejected when it is true. From the definition of the significance level, this will occur one in 20 times if the test is at the 5% significance level. Probability of a type I error is therefore the significance level of the test, and is denoted by alpha (α).

Type II errors occur if the null hypothesis is not rejected when the alternative hypothesis is true. Probability of a type II error is denoted by beta (β). The 'power' of a test is the probability of not making a type II error and is therefore 1-β.

Significance and power have accepted levels or minimum criterion levels in research. However, effect size, which is also a factor in sample size estimation, is more problematic.

EFFECT SIZE

The effect size quantifies the difference between two or more groups. It is a measure of the difference in the outcomes of the experimental and control groups, i.e. a measure of the effectiveness of the treatment. Cunningham and McCrum-Gardner (2007) provide formulae for effect size in different situations, but sample size software can be used, which requests the required information (e.g. absolute mean difference and standard deviation) and performs the calculations. The effect size can often be estimated from previous publications and based on the scientific knowledge of the researcher.

TABLE 1.
Definition of type I and type II errors

	Null hypothesis rejected	**Null hypothesis not rejected**
Null hypothesis true	Type I error	–
Alternative hypothesis true	–	Type II error

Further definitions with regard to clinical significance, clinically important differences, the primary outcome measure, and response rate are outlined below.

CLINICAL VERSUS STATISTICAL SIGNIFICANCE

Just because a result is statistically significant, this does not mean it is substantive in effect. For example, two treatments could be statistically significantly different, but their clinical effects may be so small as to be unimportant.

Minimal clinically important difference (MCID)

The MCID can be defined as the 'smallest meaningful change score' or 'the smallest (absolute) difference in score which patients perceive as beneficial, and which would mandate, in the absence of troublesome side-effects and excessive cost, a change in the patient's management. Therefore, differences in scores smaller than the MCID are considered not important independent of their statistical significance' (Ruperto, 2007).

It has also been described as a much sought after, but elusive figure. The MCID for many outcome measures is difficult to estimate and yet, if known, makes sample size calculations easier and more accurate.

PRIMARY OUTCOME MEASURE

Sample estimation size should be based on the primary outcome measure, but if there is more than one outcome then the largest sample size should be chosen so that all the outcome measures are fully powered.

RESPONSE RATE

After the sample size has been calculated, it will need to be increased depending on the expected response rate. This can be estimated from previous publications or a pilot study.

TABLE 2.
Website addresses for software packages

Software package	Website address
Minitab	www.minitab.com/en-GB/support/downloads/
PS	biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize
GPower	www.psycho.uni-duesseldorf.de/aap/projects/gpower/
Epi-info	www.cdc.gov/epiinfo/downloads.htm

A selection of software packages and websites which can be used to perform the calculations are described below.

SOFTWARE

There are many software packages for performing sample size/power calculations. Four useful ones are described in this article:

- Minitab
- PS
- GPower
- Epi-info/StatCalc

Minitab is simple to use for a beginner but has a limited range of study designs; it necessitates purchasing a licence, although a 30-day free trial is available (see *Table 2* for list of website addresses). PS, GPower and Epi-info can be downloaded for free.

GPower is more complicated to use for a novice, but does cover a wider range of study designs. GPower is useful if using effect sizes (small/moderate/large as discussed in Cunningham (2007)) rather than absolute values, but can be complicated if using absolute values to calculate the effect size.

PS is a good choice of software, although the use of mathematical symbols can make it appear more complicated and inaccessible. However, it has a help facility to provide definitions. The correct format for inputs in PS is not obvious – *power* should be input as e.g. 0.8 (for 80% power) and *alpha level* as e.g. 0.05. It also covers a wide range of study designs including survival analysis, linear regression, case-control and cohort studies, as well as the more commonly used designs in the examples given above. The most up-to-date version (V3) creates a very useful text description based on the input values, which could be copied/pasted in to a publication or grant application.

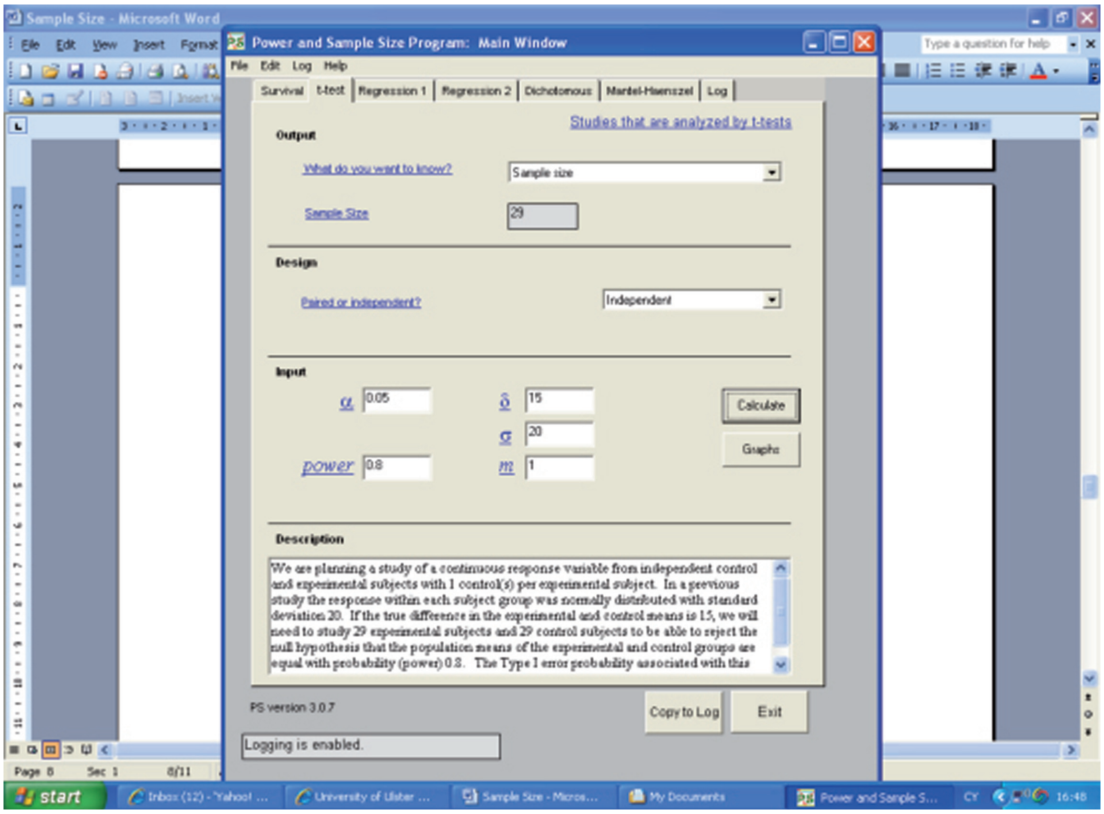
Epi-info is useful for population surveys, case-control and cohort studies.

In summary, PS is probably the best choice of software, as it covers the most commonly used study designs, is relatively easy to use and importantly, is easy and free to download. However, the user may discover his/her own personal preference.

WEBSITES

There are many websites which can be used to perform sample size/power calculations. Some

Figure 1. Using PS software for the sample size calculation for two independent groups



What do you want to know? Sample size

Sample Size 29

Paired or independent? Independent

α 0.05 δ 15

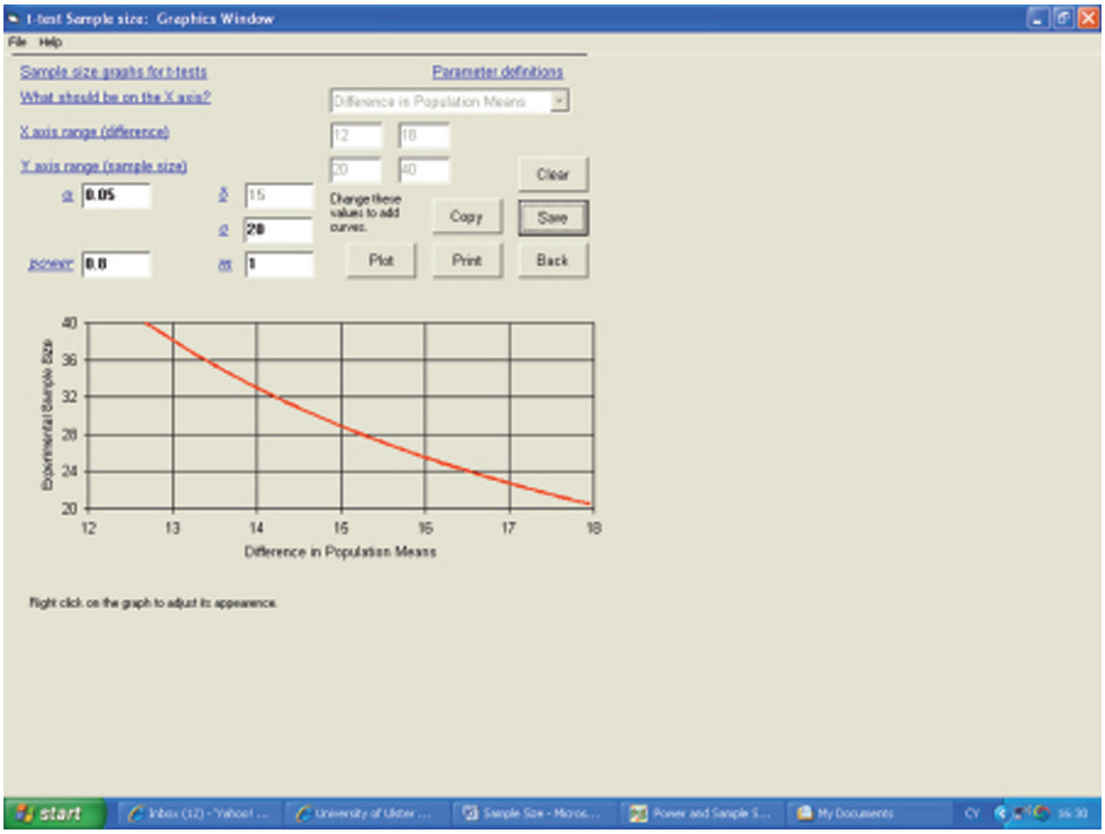
σ 20

power 0.8 m 1

We are planning a study of a continuous response variable from independent control and experimental subjects with 1 control(s) per experimental subject. In a previous study the response within each subject group was normally distributed with standard deviation 20. If the true difference in the experimental and control means is 15, we will need to study 29 experimental subjects and 29 control subjects to be able to reject the null hypothesis that the population means of the experimental and control groups are equal with probability (power) 0.8. The Type I error probability associated with this

PS version 3.0.7

Logging is enabled.



What should be on the X axis? Difference in Population Means

X axis range (difference) 12 18

Y axis range (sample size) 20 40

α 0.05

δ 15

σ 20

power 0.8

m 1

Change these values to add curves.

Right click on the graph to adjust its appearence.

Figure 2. Using PS software to produce a graph

useful ones are described below. The study designs covered vary from website to website.

The Statpages website (http://statpages.org/#Power) has a very wide range of options, although the reader should be aware that some links have not been updated and no longer work.

Russ Lenth's website (http://tinyurl.com/yhrr695) covers a range of study designs, and allows the user to adjust the levels of power, mean difference etc, by dragging the cursor along a continuum. However, this can sometimes be awkward, depending on the range of values required. Lenth's (2001) article provides some practical guidelines for sample size calculations using this website.

The Obstetrics and Gynaecology site of the Chinese University of Hong Kong (http://tinyurl.com/yep8bfz) has a 'statistics toolbox', which includes a sample size menu. It includes a range of study designs including population surveys, equivalence studies and ordinal data, as well as the commonly-used designs in the examples given above. It also includes examples to facilitate ease of use.

Power calculations for population surveys can be performed using the Raosoft website (http://www.raosoft.com/samplesize.html).

Overall, the Obstetrics and Gynaecology site of the Chinese University of Hong Kong is a particularly good choice as it covers a range of study designs and is relatively easy to use.

EXAMPLES

The examples below demonstrate what information is required for the sample size calculation for a range of study designs. To aid understanding they also demonstrate the increase/decrease in sample size when factors are modified, e.g. increasing the mean difference, decreasing the standard deviation, and increasing the power level. The PS software has been used for these examples, but alternatively, the other software packages or websites described above could be used. More information about the statistical tests mentioned can be obtained from McCrum-Gardner (2008).

The independent samples t-test is used to compare sample means from two independent groups for an interval–scale variable when the distribution is approximately normal; the paired t-test is used to compare two sample means for an interval–scale variable where there is a one-to-one correspondence (or pairing) between the samples and the distribution of within-pair differences is approximately normal; the chi-square (χ^2) test is used to compare proportions between two or more independent groups or investigate whether there is any association between two nominal-scale variables. Sample size calculations are given below for each of these three study designs.

Example 1 – two independent groups (independent t-test)

A proposed study wishes to investigate the effects of a new hypertensive drug (experimental group) compared to a conventional treatment (control group). Previous studies show that the minimum clinically important difference is 15 mmHg and that the pooled standard deviation (SD) is 20 mmHg.

Using the PS software (*Figure 1*) it is estimated that 29 subjects would be needed in each of the control and experimental groups (α level P=0.05, power 80%) in order to detect a statistically significant difference in mean blood pressure between the two groups (if it exists). Increasing the mean difference to 20 mmHg would require 17 participants per group, i.e. a substantial decrease. Decreasing the SD to 10 mmHg would require only five participants in each group.

Example 2 – two proportions (chi-square test)

A health promotion intervention to reduce smoking is going to be introduced in a country with a high prevalence of smoking. The current smoking rate in a recently published report is 65% and it is estimated that the intervention will reduce the smoking levels by 30%. A sample size of 42 in each group is required (P=0.05, power 80%). Increasing the power level to 90% requires 56 per group.

Example 3 – two paired groups (paired t-test)

Is there evidence that clofibrate changes the mean cholesterol level? Cholesterol is to be measured before and after receiving clofibrate. From previous studies, a mean difference of 40 mg/dl is deemed clinically significant, with standard deviation of 50. A sample size of 14 is required (P=0.05, power 80%), P=0.01 requires 21 subjects, i.e. an increase.

KEY POINTS

- Sample size estimation is a significant concern for researchers as guidelines must be adhered to for ethics committees, grant applications and publications.
- Studies may be underpowered (too few participants), or overpowered (too many participants) so it is important to achieve the correct balance.
- Relevant definitions of power, significance level, effect size etc. are provided in order to understand the process of performing sample size calculations.
- Clear examples of sample size calculations for three commonly-used study designs are provided for illustration and to aid understanding.
- A range of software packages and websites are available to perform sample size calculations and a selection of them are discussed and evaluated.

POST-HOC POWER CALCULATIONS

An *a priori* analysis calculates sample sizes for given effect sizes, alpha levels and power values. A post-hoc/retrospective analysis computes power values for given sample sizes, effect sizes, and alpha levels. Both GPower and PS can perform these analyses.

The previously mentioned cholesterol study (example 3) was carried out but as a result of the poor response/dropout rate, only 12 patients were recruited. A mean difference of 50 and SD of 60 mg/dl were found. What is the power for this study? Carrying out a retrospective analysis indicates that the power is 75% i.e. the study is under powered as it is less than 80%.

GRAPHICAL PRESENTATION

Both PS and GPower can produce graphs to explore the relationships between power, sample size and effect size, e.g. the difference in population means. It is often helpful to hold one of these variables constant, and plot the other two against each other. For example, a plot of difference in means against sample size for the two treatment groups in example 1 is shown in *Figure 2*.

CONCLUSIONS

The reader must be aware that the choice of the correct statistical test is vitally important at the statistical analysis stage, and this is described elsewhere (McCrum-Gardner, 2008). This article has described the process of sample size calculations including relevant definitions and examples for different study designs are provided for illustration. A range of software packages and websites have also been discussed and evaluated. This information should be useful for researchers in order to perform sample size calculations for their research projects. IJTR

Conflict of interest: none

Central Office for Research Ethics Committees (2007) Question specific guidance on NHS REC application form (2007). Available from: http://tinyurl.com/yjogwn9 (accessed 17 December 2009)

Cunningham JB, McCrum-Gardner E (2007) Power, effect and sample size using GPower: practical issues for researchers and members of research ethics committees. *Evidence Based Midwifery* **5**(4): 132–6

Devane D, Begley C, Clarke M (2004) How many do I need? Basic principles of sample size estimation. *Journal of Advanced Nursing* **47**(3): 297–302

Florey CV (1993) Sample size for beginners. *BMJ* **306:** 1181–4

Lenth RV (2001) Some Practical Guidelines for Effective Sample Size Determination. *The American Statistician* **55:** 187–93. Available from http://www.stat.uiowa.edu/techrep/tr303.pdf (accessed 17 December 2009)

McCrum-Gardner E (2008) Which is the correct statistical test to use? *Brit J Oral Maxillofacial Surgery* **46:** 38–41

Ruperto N (2007) Is Minimal Clinically Important Difference Relevant for the Interpretation of Clinical Trials in Pediatric Rheumatic Diseases? *The Journal of Rheumatology*. http://tinyurl.com/yjst6ct (accessed 17 December 2009)