Statistical Notes: Sample size calculation

A Cook, A Sheik

INTRODUCTIO

The accurate determination of sample size is a crucia aspect of study design. If the study sample is to sm small, a real effect may exist, and be observed, ye lac ack statistical significance thus resulting in a fals ne negative conclusion. On the other hand, a study that i lar larger than it needs to be will absorb funding tha co could have been better used elsewhere, and possibl delay the release of important results. Sample size calculations need to be presented in funding ap applications where they are rightly subjected to clos sc scrutiny, and they should also be clearly reported whe publishing results

Es estimating sample size is complex and for all but th simplest studies a statistician should be consulted Ho However, the questions that a statistician will ask ar la largely predictable by someone with an elementar ug understanding of the principles involved in calculatin sa sample size. In this paper, we describe the method an and the information required for a simple calculation mo moving on to consider some aspects of more comple study designs that may influence sample size.

COMPONENTS OF SAMPLE SIZE CALCULATION

Difference to be detecte

The process of estimating the optimal number of su subjects for a particular study requires researchers t th think forward to the data analysis stage. Significanc tests, comparing two or more groups, seek to de determine whether an observed difference is a chanc fi finding or whether it represents a true differenc be between the groups. When planning the study, it i ne necessary to decide on the size of difference that th study will be able to detect. This should be the sm smallest change that is considered to be clinicall im important. Ideally, we would want to detect an improvement, no matter how small, but because sa sample size increases sharply when seeking smal changes, pragmatic considerations become important

Mo Most studies gather data on several outcomes; fro these the single most important one should be identified, commonly referred to as the 'primary ou outcome'. Sample size should then be based on th an anticipated difference in primary outcome between th gr groups under study. In the case of several outcome ha having equal importance, so no primary outcome ca be identified, sample size should be calculated fo ea each outcome of interest and the largest resultant valu used

In the context of a clinical trial in which the primar outcome measure is a categorical variable (i exacerbations of asthma within a predetermined per - od, for example), baseline disease prevalence, (i.e. fr frequency of exacerbation's in the control group needs to be known, together with the change in di disease prevalence that the study seeks to detect. If th primary outcome is a continuous variable (eg FEV1 o PE PEF variability), the expected difference in the mea value is required, together with the standard deviatio of the data in the control group.
So Sometimes the required values will be obtainable fro the literature or from a pilot study; if this is not th ca case and reliable estimates are unavailable 'best guess clinical judgement must suffice

Significance and powe

Significance tests can result in Type I or Type I errors. A Type I error, or false positive, occurs when ne new treatment is declared to be better than the contro bu but is in fact no different. Conversely, to conclude tha the new treatment is no different when in fact it is co constitutes a Type II error or false negative result. Th risk of such errors is directly related to sample size wh when planning the study, it is therefore necessary t decide what risk of error is considered acceptable.

Table 1. Fbrmulae to calculate number in each arm of a two group tria

1. Continuous primary outcom	$n \geq \left\{ \frac{T_1 + T_2}{diff/sd} \right\}^2$	*diff* - difference in mean value *sd* - standard deviation in control grou
2. Categorical primar outcom	$n > \left\{ \frac{T_1\sqrt{2\bar{p}(1-\bar{p})} + T_2\sqrt{p_1(1-p_1)+p_2(1-p_2)}}{p_1-p_2} \right\}^2$	$\bar{p}$ - prevalence overal p_1 - prevalence, control grou p_2 - prevalence, intervention grou

T_1 and T_2			
Probability(Type I error)	1%	5%	10%
T_1	2.576	1.960	1.645
Power	90%	80%	70%
T_2	1.282	0.842	0.52

Adrian Cook
Statisticia

Aziz Sheik
NHS R&D National Primar Care Training Fello
Department of Primary Healt Care & General Practic
Imperial College School o Medicin

Correspondence to
Adrian Coo
Department of Primary Healt Care & General Practic
ICSM Charing Cross Campus
Reynolds Building
St Dunstan's Road,
London, W6 8R
a.d.cook@ic.ac.u

Date submitted: 05/01/0
Date accepted:28/02/0

Prim. Care Respir. J
2001 10(1):15-1

The probability of a Type I error is often referred to as the **significance level,** typically set at 0.05 (5%). The **power** of a study is one minus the probability of a Type II error, the most commonly accepted chance of a Type II error is 0.2 (20%), giving a power of 0.8 (80%). In some situations, other values of significance and power will be more appropriate. For example, if a new treatment has unpleasant side effects it may be appropriate to reduce the false positive risk to one-percent (0.01), or for a treatment that possibly represents a major therapeutic breakthrough, power may be increased to 90% or above

Example

A randomised-controlled trial is to be carried out to compare the efficacy of a new bronchodilator with salbutamol in patients with established asthma. The primary outcome is morning peak expiratory flow (PEF), and a difference of greater than 20 litres per minute would be considered clinically important. The trial is required to have 80% power at the five-percent significance level. From the literature, it is estimated that adult asthmatics have an average morning PEF of 400 l/min, with a standard deviation of 100 l/min.

From equation 1, with $_1$=1.96 and $_2$=0.842, the study would require at least 250 patients in each arm

Losses to follow-up

For a number of reasons participants pull out of longitudinal research studies and are lost to follow up. Usually this results in exclusion from the data analysis, effectively reducing the sample size. Since losses to follow up are almost inevitable it is wise to compensate at the design stage by calculating sample size normally and then multiplying up by an appropriate factor. The factor used depends on the expected losses to follow up, best estimated from previous studies of a similar nature. In the absence of any prior knowledge a dropout rate of 20% may be assumed, if the sample size is then increased by 25% it will be returned to the original value by a 20% dropout rate.

Example

A randomised-controlled trial designed to compare the efficacy of a new bronchodilator with salbutamol requires at least 250 patients in each arm to have 80% power of detecting a five-percent difference in PEF at the five-percent significance level. In the absence of data from pilot studies, it is assumed that up to 20% of subjects may be lost to follow-up. In order to account for dropouts, it is necessary to increase the sample size by 25%, thus requiring at least 313 patients to be recruited to each arm of the trial. Losing 20% of the 313 patients then returns the total to 250

UNEQUAL GROUP

The most statistically efficient study design will always have equally sized groups, although sometimes other considerations still make an unbalanced study the best option. If a case-control design is used to study a rare disease the number of available cases may be limited, or if a new treatment is particularly expensive, it may be cost effective to compare a small number of treated patients with a larger untreated number, for example. To compensate for the lower efficiency of an unbalanced design the overall sample size needs to increase - reducing the number of patients in one arm of a trial thus requires an increase of greater magnitude in the other arm. As the design moves further from a 1:1 ratio, to 2:1, 3:1 or 4:1 greater compensation is required. The statistical inefficiency of unbalanced trials should therefore be weighed against the practical considerations mitigating against a balanced study design

CONCLUSION

The final step of sample size calculation is a simple matter of entering numbers into a formula, but as with any formula it is very much a case of 'rubbish in, rubbish out'. This paper has described the principle pieces of information that researchers need to make available to statisticians: the effect size which needs to be reliably detected; a measure of dispersion for continuous outcomes; the acceptable risk of error (significance and power) and predicted losses to follow-up. Care taken at this stage will help ensure that the results obtained can be readily and reliably interpreted

ACKNOWLEDGMENT

We thank Claire Cook for her constructive comments on an earlier draft of this manuscript

FURTHER READIN

Mathers N, Williams M, Hancock B. *Statistical analysis in primary car* . Oxon: Radcliffe Medical Press, 2000

Pocock SJ. *Clinical trials: a practical approac* . Chichester: Wiley, 1996