Available online at www.sciencedirect.com

SCIENCE DIRECT®

Physical Therapy in Sport 6 (2005) 153–163

Physical Therapy in sport

www.elsevier.com/locate/yptsp

Research without tears

How big does my sample need to be? A primer on the murky world of sample size estimation

Alan M. Batterham[a,*], Greg Atkinson[b]

[a]School for Health, Sport and Exercise Science Research Group, University of Bath, Bath BA2 7AY, UK
[b]Research Institute for Sport and Exercise Sciences, Liverpool John Moores University, Henry Cotton Building, Webster Street, Liverpool L3 2ET, UK

Received 21 April 2005; revised 25 May 2005; accepted 26 May 2005

Abstract

Background: An explicit justification of sample size is now mandatory for most proposals submitted to funding bodies, ethics committees and, increasingly, for articles submitted for publication in journals. However, the process of sample size estimation is often confusing.
Aim: Here, we present a primer of sample size estimation in an attempt to demystify the process.
Method: First, we present a discussion of the parameters involved in power analysis and sample size estimation. These include the smallest worthwhile effect to be detected, the Types I and II error rates, and the variability in the outcome measure. Secondly, through a simplified, example 'dialogue', we illustrate the decision-making process involved in assigning appropriate parameter values to arrive at a ballpark figure for required sample size. We adopt a hypothetical, parallel-group, randomized trial design, though the general principles and concepts are transferable to other designs. The illustration is based on a traditional, power-analytic, null hypothesis-testing framework. In brief, we also address sample size estimation methods based on the required precision of the mean effect estimate.
Conclusion: Rigorous sample size planning is important. Researchers should be honest and explicit regarding the decisions made for each of the parameters involved in sample size estimation.

Keywords: Sample size; Power; Minimum clinically important difference

1. Introduction

How many statistical advisors does it take to change a light bulb? Three, one to change the bulb, one to check the power, and one to assess the goodness-of-fit. Alternative answers to this question include 'two, plus or minus one'. Clearly, humour and statistics are not comfortable bedfellows. Indeed, few subject areas strike more fear into the hearts of novice and experienced researchers—and research consumers—alike than the murky area of sample size estimation and power analysis. In a recent editorial in this journal, Zoë Hudson stated that (Hudson, 2003, p. 105):

> There has been an ongoing debate between editors and editorial boards of peer reviewed journals whether to accept articles with no or low statistical power. Indeed, there has been a call by some journals to refuse articles that do not contain a power analysis for the sample size required to show a significant difference.

Statistical power is defined as the probability of detecting as statistically significant a clinically or practically important difference of a pre-specified size, *if such a difference truly exists*. Formally, power is equal to 1 minus the Type II error rate (beta or β). The Type II error rate is the probability of obtaining a non-significant result when the null hypothesis is false—in other words failing to find a difference or relationship when one exists. In sample size planning, beta is fixed in advance to ensure an adequate probability of detecting a true, clinically relevant effect of a given size. These issues are discussed in detail subsequently. The aim of this primer article is to present a sketch of the theory and practice of power analysis and sample size estimation and, hopefully, help to demystify the process and alleviate some of the attendant trepidation.

[*] Tel.: +44 1225 383448; fax: +44 1225 383275.
E-mail address: a.m.batterham@bath.ac.uk (A.M. Batterham).

1466-853X/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ptsp.2005.05.004

From the outset, we would like to emphasise our deliberate use of the term 'sample size *estimation*', rather than 'sample size *calculation*'. Although the arrival at a number for the required sample size is invariably based on (often complex) formulae, the term 'calculation' implies an unwarranted degree of precision. Indeed, as noted by Williamson, Hutton, Bliss, Campbell, and Nicholson (2000, p. 10):

> Their (*sample size formulae*) purpose is not to give an exact number, say 274, but rather to subject the study design to scrutiny, including an assessment of the validity and reliability of data collection, and to give an estimate to distinguish whether tens, hundreds, or thousands of participants are required.

Such sentiments echo those of biostatistician and clinical trials expert Stephen Senn (1997), who described power calculations as "a guess masquerading as mathematics".

Pocock (1996) commented that sample size estimations are "a game that produces any number you wish with manipulative juggling of the parameter values" (as we demonstrate subsequently in this article). Unfortunately, in our experience this 'game' is played all too frequently. A common scenario is the following. A researcher or research team decide, on practical grounds, on the maximum number of participants that can be recruited and measured. Later, when faced with the increasingly common demands (from ethics committees, grant awarding bodies, journal editors, and the like) for a fully justified written section on sample size estimation, they approach a statistical advisor for assistance. As we discuss later in this article, one of the key parameters in sample size estimation is the minimum clinically important difference (MCID)—the smallest effect worth detecting that is of clinical significance. In our 'common scenario', a relatively large MCID may be selected that 'justifies' the sample size chosen (smaller sample sizes are required to detect larger effects). We believe that this manipulative rearrangement of the sample size estimation equations is unethical. In the profession, this approach is said to involve replacing the clinically important difference with the *cynically* important difference.

Use of the cynically important difference in sample size justifications may lead to underpowered studies and the increased probability that some clinically beneficial interventions will be dismissed as 'non-significant' (a Type II error). To return briefly to Zoë Hudson's editorial comments on this matter, where does this leave us? Everitt and Pickles (2004) argue that the case against studies with low numbers of participants is strong, though they concede that with the growing use of meta-analysis there may still be a place for smaller studies that are otherwise well-designed and executed. We agree with the opinions of Williamson et al. (2000, p. 10) that "all proposals should include an honest assessment of the power and effect size of a study, but that an ethics committee need not automatically reject studies with low power". However, proper sample size estimation is often regarded as an ethical *sine qua non*, helping to avoid a waste of resources and/ or the subjecting of participants to potentially ineffective (and possibly harmful) interventions due to samples that are too small or, less frequently, larger than necessary. Moreover, the process of sample size estimation helps to clarify one's thoughts at the outset with respect to what is the central research question, what is the primary outcome variable, what are the secondary outcome variables, and what is the proposed analysis strategy?

The steps involved in the sample size estimation process can, therefore, help develop and refine the research design and methods for the study. The theory and practice underlying these steps is outlined in the first substantive section of this primer. In the second section, we illustrate the 'dialogue' and decision-making involved in arriving at a sample size estimation using a worked example. We restrict our discussion to estimations carried out before the study is conducted. Although not uncommon, we believe that conducting power analyses once the data have been collected is largely redundant. At this stage, power is appropriately and more effectively illustrated by the calculation and presentation of confidence intervals for the effect of interest (Wilkinson, 1999).

2. Considerations for a statistical power analysis

In the first part of this primer, we concentrate on the factors that influence statistical power and required sample size. We will not delve too much into the underlying mathematics in view of the availability of specialist sample size estimation programs such as nQuery Advisor® (Statistical Solutions, Cork, EIRE), sample size and power options in popular software packages including Stata®, SAS®, and StatsDirect®, as well as published tables and nomograms (Machin, Campbell, Fayers, & Pinol, 1997). Rather, we consider each factor in turn with the aid of Table 1, which is designed to illustrate the impact of changing various study factors on required sample size. Our 'baseline' hypothetical situation for comparison is detailed in column A of Table 1. In column A, we start with a hypothetical two-sample design, which might involve the comparison of mean changes in pain scores between an intervention and a control group (e.g. measured using a continuous or categorical Visual Analogue Scale). With this design and using an independent t-test with a 0.05 two-sided significance level, a sample size of 23 in each group will have 90% power to detect a difference in mean change in pain of 1 unit, assuming that the common standard deviation is also 1 unit. We emphasise at this stage that a difference of one standard deviation in mean pain score change is a relatively large effect. A larger sample size would be needed to detect smaller, potentially clinically important, effects.

Table 1
The effects of changing various terms in a statistical power calculation

	A	B	C	D	E	F
Test significance level (alpha)	0.05	0.05	0.05	0.05	0.05	0.05
1 or 2 sided hypothesis	2	2	1	2	2	2
Difference in means ($_{\mathrm{d}}$)	1.00	1.00	1.00	**2.00**	1.00	1.00
Common standard deviation (SD)	1.00	**0.45**	1.00	1.00	**2.00**	1.00
Effect size (*d*/SD)	1.00	**2.22**	1.00	**2.00**	**0.50**	1.00
Statistical power (%)	90	90	90	90	90	**80**
Sample size per group	23	5	18	7	86	17
95% CI assuming a constant sample size of 23 subjects	0.33–1.67	0.70–1.30	0.44–1.56	1.33–2.67	−0.12–2.12	0.46–1.54

Changed terms are shown in bold. A, Independent t-test for two-group comparison; B, The effect of changing design to a repeated measures (paired *t*-test); C, The effect of changing to a one-sided hypothesis of interest; D, The effect of increasing the difference in means to be detected; E, The effect of increasing the data variability; F, The effect of reducing the required statistical power (an increase in beta).

2.1. Research design

Generally, research designs involving correlated data (e.g. repeated measures or crossover designs) are associated with greater statistical power than those involving separate samples allocated to different treatment groups. This relative improvement in statistical power is mediated by the magnitude of the 'variability' (standard deviation) term, which is entered into the calculations. For example, the data in column B of Table 1 represent what would happen if the two-group design were replaced with a repeated measures analysis involving the paired *t*-test. Assuming that such a design generates data that are correlated to a degree of $r=0.9$, then the standard deviation of the pairs of changes would be 0.45 instead of 1 unit. This would increase the effect size (in this case defined as the mean difference divided by the standard deviation of the paired differences) to 2.22, and hence fewer subjects ($n=5$) are required to detect this larger effect size. Similarly, with a repeated measures design, the smaller standard deviation associated with correlated data would lead to a narrower (more precise) 95% confidence interval for the mean difference between pain change scores (Table 1).

The decreased variability associated with a repeated measures design suggests that it may be sensible to adopt this approach when such a design matches the research question. Nevertheless, it is often difficult to do this, since a treatment might have long-term residual effects on the primary outcome variable. For example, it is virtually impossible to adopt a repeated measures design to investigate the efficacy of a physical therapy intervention on functional outcomes with participants who have a pre-existing injury. A research design that is worth considering by physical therapists in such situations is the matched-pairs approach, since these designs might generate correlated data for analysis, as well as generally reduce variability in the data (Atkinson & Nevill, 2001). A matched-pairs design might involve matching participants in a treatment group and a control group for any intervening variables, such as age or body mass, or the baseline measurements might be used to match participants. If matching of participants for intervening variables or pre-intervention measurements is not possible, one can still improve statistical power by entering these variables as covariates in the analysis (Vickers & Altman, 2001).

It is also worth mentioning that an 'unbalanced' research design might require a larger total sample size, all other factors being equal. That is, if it was only possible (or was desirable) for some reason to recruit half as many participants for the intervention group than the control group (or vice versa), then statistical power would be lower (for the same total sample size) compared to the scenario in which group sizes were equal (Whitley & Ball, 2002). For example, consider a two-group trial in which 50 participants were required in each group (total $N=100$) to obtain the desired power to detect the smallest worthwhile effect. If this total sample size of 100 were maintained, but there were 70 participants in one group and 30 in the other, then the power would be lower than that in the 'balanced' design. Together with this statistical power issue, one should also scrutinise whether sampling bias is apparent if groups are unequal in size; i.e. any systematic factor which has led to the ratio of sample sizes being other than one should be considered.

2.2. Question of interest

Researchers seldom rationalise the choice of a one-tailed or two-tailed hypothesis or confidence interval (Atkinson & Nevill, 2001; Knottnerus & Bouter, 2001). One-tailed analyses are selected when the hypothesis or question of interest is directional (e.g. a decrease in rating of pain in response to some intervention is hypothesised), whereas two-tailed analyses are chosen when the hypothesis of interest centres on a change, irrespective of direction. Atkinson and Nevill (2001) argued that a one-tailed analysis might be employed when the researcher is only interested in enhancement of a functional performance outcome per se, and that performance outcome is directly measurable. The rationale is that a certain treatment would not be adopted if it *either* did not change, or actually decreased, functional performance. Therefore, a directional alternative statistical hypothesis would be appropriate in this case (Knottnerus & Bouter, 2001). However, Altman (1991) stated that in the vast majority of cases the two-tailed procedure is the more correct one, and that even when we strongly suspect that a treatment can only lead to changes in one direction, we cannot be certain. Furthermore, knowledge of significant and substantial changes in the opposite direction ('harm' as opposed to benefit in an intervention study) is also important.

Importantly, one-tailed inferential statistics offer a gain in statistical power over the corresponding two-tailed analysis (Peace, 1988; Rice & Gaines, 1994). One can see from column C in Table 1 that a one-tailed rather than a two-tailed analysis has been selected. This change is estimated to result in five fewer participants being required for the research. The 95% confidence interval also becomes narrower. However, we believe that there must be a compelling rationale for conducting one-sided tests of significance, and that this decision must be made before the data are analysed. Conducting a one-tailed test after failing to find significance with a two-sided test represents an unethical and largely pointless fishing expedition.

2.3. Effect size

Generally, the bigger the size of effect to be detected, the greater the statistical power is for a given sample size and within-subject variability. In column D of Table 1, the difference between treatment and control groups in the mean change of pain ratings has been doubled to two units. The standardised effect size also doubles leading to a substantial reduction of estimated sample size. Note that the lower limit of the confidence interval is higher than previous lower limits. This is interpreted as evidence that the population mean change is greater than 1 unit. Therefore, although the width of the confidence interval has not changed, the fact that the hypothesised mean change is two rather than 1 unit leads to greater probability that the population mean is not zero.

One of the most difficult, and yet critical, aspects of sample size estimation is the a priori selection of effect size (Atkinson, 2003). The most accurate predictions of effect size are obtained from past and related studies involving a similar intervention and the same outcome variable or from one's own preliminary studies or pilot work. It is also good practise to choose an effect size on the basis of expert opinion or data on the minimum clinically important difference or correlation. For example, Hopkins, Hawley, and Burke (1999) delimited worthwhile effect sizes for athletic performance variables on the basis of the likelihood of winning a medal at major championships. Approaches linking magnitudes of effect with associated clinical or practical endpoints are known as 'anchor-based' methods. It is also important to appreciate that delimited effect sizes may differ according to the type of intervention that is introduced to the participants. For example, a very invasive, time consuming or expensive intervention might lead to the selection of a larger effect size than a simpler intervention. This idea of weighing up the cost of treatment to the magnitude of benefit is encapsulated in the "Number Needed to Treat" philosophy of effect size and statistical power estimations (Dalton & Keating, 2000).

In the absence of any robust anchor-based information on the clinically or practically worthwhile effect size, one can turn to 'distribution-based' methods using generalised 'cut-off' values for effect size. Cohen (1988) suggested standardised effect sizes (mean difference divided by the between-subject standard deviation) of 0.2, 0.5 and 0.8 as representing 'small', 'moderate' and 'large' effects, respectively. An oft-quoted recommendation is to default to a Cohen effect size (d) of 0.2 for the smallest worthwhile effect, in the absence of robust evidence. Powering a study to detect this small effect has a dramatic impact on the estimated sample size required. Assuming two groups with an alpha of 0.05 and 90% power in a two-sided test would require 527 participants in each group. It should be noted that such generalised effect sizes were formulated with the social sciences in mind, for which there may be no directly measurable variable that can be used to appraise practical significance.

It is also worth noting that in repeated measures designs, the effect size generated by dividing the mean difference by the standard deviation of the change scores is not strictly interpretable according to the thresholds for Cohen's d of 0.2, 0.5, and 0.8 described previously. These thresholds are based on fractions of a between-subject standard deviation, not within-subjects variability. As discussed in detail by Cumming and Finch (2001), to assess the size of an effect by standardising the effect size by the standard deviation of the change scores does not make substantive sense. Arguably, the natural frame of reference for thinking about a change in pain scores, for example, is the estimate of the population standard deviation indicated by the between-subject variance in the sample at baseline (post-intervention variability may well be inflated due to individual differences in response to treatment). The standard deviation of the change scores is required, however, for sample size estimations based on a paired t-statistic.

2.4. Variability

The variability in the primary outcome measurement influences sample size estimations. For the two-group comparison example in Table 1, if the between-subjects standard deviation is increased from 1 to 2 units, then sample size increases to 86 participants per group (column E). The philosophy is that the more 'noise' there is in the data, the more subjects are required in order to detect the 'signal' of a given mean difference.

For 'within-subjects' or 'repeated measures' designs, the standard deviation of the changes or the differences is the important statistic (Atkinson & Nevill, 2001). In this respect, there is a direct link between statistical power and the test–retest variability of the outcome measurement (Atkinson & Nevill, 1998). Poor test–retest repeatability for

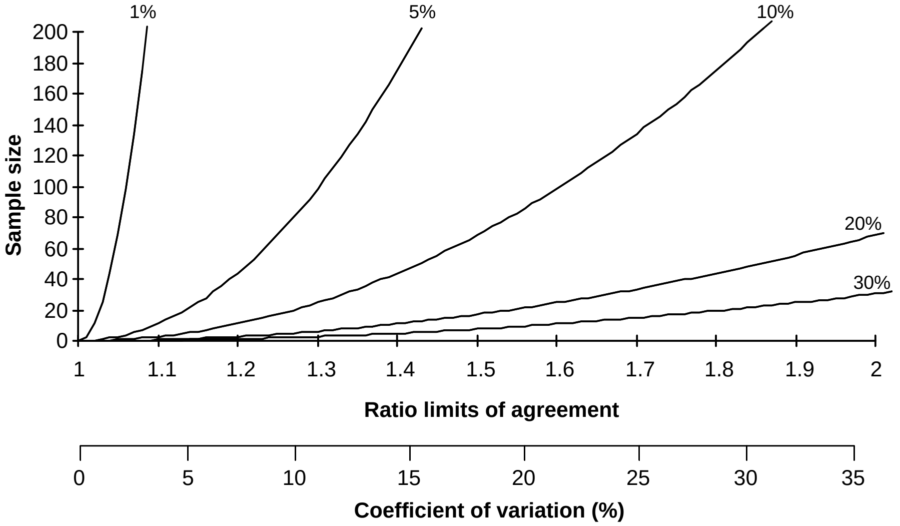

Fig. 1. A nomogram to estimate the effects of measurement repeatability error on whether 'analytical goals' are attainable or not in exercise physiology research. Statistical power is 90%. The different lines represent different worthwhile changes of 1, 5, 10, 20 and 30% due to some hypothetical intervention. The measurement error statistics, which can be utilised are the LOA and CV. For example, a physiological measurement tool, which has a repeatability CV of 5% would allow detection of a 5% change in a pre-post design experiment (using a paired *t*-test for analysis of data) and with a feasible sample size (approximately 20 participants).

the outcome variable would mean, again, that a greater sample size is required in order to detect a given change over time. In Fig. 1, a nomogram is presented to show the relationship between test and retest variability (described with the coefficient of variation or limits of agreement statistics) and required sample size. With the nomogram, one can predict whether the test–retest variability of the outcome variable is so high that the required sample size for a simple, two-condition repeated measures experiment becomes impractical.

A common sample size calculation involves the paired t-test as the choice of statistical analysis (e.g. for comparing functional performance between two experimental conditions). The statistic, which represents the variability of measurements and which would be used in the sample size calculations in this case is the standard deviation of the differences. Table 2 presents the mathematical relationships between the standard deviation of the differences and various popular test–retest error statistics. Readers might like to refer to Table 2 when they are attempting to arrive at a general value for the variability component in the sample size calculation, and if several studies have cited different measurement error statistics.

Altman (1991) noted that an estimate of the standard deviation of the differences is frequently not available. As noted, a handle on this variability may be gained by examination of reported test–retest measurement error statistics, though technically it is the standard deviation of the changes expected (treatment condition minus control condition) that should form the denominator for the effect size informing the sample size estimation. Ideally, this is best estimated from prior substantive studies, or preliminary or pilot studies.

2.5. Alpha and beta

Consider a scenario in which we compare measurements for a primary outcome variable in two independent groups using, say, an independent t-statistic. The observed mean difference between groups is 10 units, with a *P*-value for the t-statistic of 0.001. Formally, this *P*-value is the a posteriori likelihood that we would have observed an effect as large (or larger) than 10 units *under the assumption that the null hypothesis is true* (Devane, Begley, & Clarke, 2004). This *P*-value indicates that we would observe a difference of 10 units or greater only one time in a thousand, assuming that the null hypothesis is true. Therefore, in this instance the null hypothesis is *implausible*, because the effect we observed would occur by chance only very rarely. Sterne and Smith (2001) suggest that *P*-values measure the strength of the evidence against the null hypothesis—the smaller the *P*-value, the stronger the evidence. However, the *P*-value *does not* provide the probability that the null hypothesis is

Table 2
The mathematical relationships between the standard deviation of the differences and various measurement error statistics

Statistic	Formula
Standard error of measurement (SEM)	$\mathrm{SD_{diff}}=\sqrt{2}\times\mathrm{SEM}$
Coefficient of variation (CV) and grand mean (×) of data	$\mathrm{SD_{diff}}=\sqrt{2}\times(\mathrm{CV}\div100\times\text{grand mean})$
Limits of agreement (LOA)	$\mathrm{SD_{diff}}=\mathrm{LOA}\div1.96$
Pearson's correlation coefficient (r) and between-subjects standard deviation (SDB)	$\mathrm{SD_{diff}}=\sqrt{(2\times\mathrm{SDB}^2–2\times r\times\mathrm{SDB}^2)}$
Mean square error (MSE) term from repeated measures analysis	$\mathrm{SD_{diff}}=\sqrt{(2\times\mathrm{MSE})}$

true, because in standard frequentist statistics it can only be calculated by assuming that the null is true. Whether or not the observed difference is clinically important as well as statistically significant is another issue that cannot be solved by interpretation of *P*-values alone.

How does one decide whether the a posteriori *P*-value for the effect indicates a result that is sufficiently rare to warrant rejection of the null hypothesis? The statistician and geneticist R.A. Fisher was the first to advocate a cut-off level of significance ('alpha', α) of 5% ($P=0.05$) as a standard criterion for suggesting that there is evidence against the null hypothesis. Fisher (1950, p. 80) suggested that:

> If *P* is between 0.1 and 0.9, there is certainly no reason to suspect the hypothesis tested. If it is below 0.02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at 0.05.

Although it was not Fisher's intention, his approach was subsequently modified into the ubiquitous, inflexible, 'yes/no' type decision making for statistical hypothesis tests based on an arbitrary alpha value. If a priori, we choose to adopt an alpha value of 0.05 then we are willing to accept a 5% probability of falsely rejecting a true null hypothesis (a Type I error or 'false alarm'). We reiterate that there is nothing special about an alpha level of 5%. It was Fisher's belief that the researcher should critically interpret the obtained *P*-value and not use the a priori alpha value as an absolute decision rule (Sterne & Smith, 2001).

Like effect size, the choice of alpha for sample size estimations may depend on the research circumstances. For example, it might be preferable in some circumstances to adopt a more cautious approach to rejecting the null hypothesis by selecting a smaller value of alpha. Such a decision would mean that a larger sample would be needed (all other factors being equal). Conversely, a larger value of alpha might be selected when false rejection of the null hypothesis is not as serious an error, for example, if the intervention is easily adopted by patients and there is little or no risk of harm. A related issue in the confidence interval approach to arriving at study conclusions is the choice of coverage. A common coverage for a confidence interval is 95% (0.95), but this probability is again arbitrary. The adoption of a 99% confidence interval may be warranted if precision of knowledge about the population difference or change is more important (e.g. for very invasive or expensive interventions, or those with the potential for harmful side-effects).

Fisher's approach was concerned primarily with the risk of Type I errors. In sample size estimation, the probability ('beta') of a Type II error—failing to reject the null hypothesis when the treatment truly works (a 'failed alarm')—is also a required parameter. Statistical power is calculated by subtracting beta from 1. There is less convention surrounding the required power than for alpha values, but a now commonly selected value of beta is 10% (0.1), giving statistical power of 90% (0.9). A power of 80% is usually regarded as the minimum acceptable. Note the value of beta is conventionally larger than alpha, a situation ascribed to the natural conservatism of scientists (Machin et al., 1997). A researcher might be willing to be less conservative in falsely concluding that there are no differences or changes present by increasing beta to 20% and hence decreasing the selected level of statistical power to 80% (column F of Table 1). A reduction in sample size results in this situation. The aim of fixing alpha and beta in advance is to decrease the number of mistakes made. In the planning of clinical trials, appropriate selection of alpha and beta help to ensure that the study is large enough to minimise the risk of recommending ineffective interventions (Type I error) and of rejecting interventions that may well be beneficial (Type II error). Kelley, Maxwell, and Rausch (2003) argued that the widespread inattention to these issues has led to a situation in which the probability of a Type I error (say 5%) is only slightly lower than the probability of correctly rejecting the null hypothesis (power). Combined with the well-documented publication bias effect this suggests that a substantial proportion of significant findings in the published literature may be Type I errors. We recommend that a range of powers, alphas, and effect sizes be used to estimate required sample size to overcome the tendency for a single estimate being regarded as absolutely definitive (Wilkinson, 1999).

3. Illustrative examples of the sample size estimation decision-making process

In this section, we present a brief example of a simplified 'dialogue' illustrating aspects of the decision-making process involved in estimating the required sample size in a typical study scenario (though the fundamental concepts and principles extend to any research design). The design presented is a parallel-group randomised controlled trial. We issue the caveat that the scenario is for illustrative purposes only and is not based on a systematic review of the literature in the topic area presented.

3.1. The sample size estimation dialogue

Alan has approached Greg for advice regarding the planning of a trial examining the effect of a complex intervention (synergistic vitamin and mineral supplement combined with physiotherapy) on recovery from a particular musculoskeletal injury. Before investing time on the sample size estimation, the primary question of interest is "is the study worth conducting?" Hopefully, the research team will have already firmly established the rationale and justification for the proposed work, and answered this "so what?" question. If not, it is worthy of discussion at this point before proceeding further.

Greg. So, Alan, you say you are interested in recovery from injury in this trial. Could you be a little more specific? What aspect of recovery are you most interested in—in other words what's your primary outcome variable?

Alan. In this study, we are mainly concerned with self-reported pain, which we are convinced we can measure validly and reliably using a Visual Analogue Scale before and after the intervention.

Greg. OK, let's assume that pain score is the primary outcome variable. It's customary to power the study on that variable. We'll also need to discuss the length of the intervention, and when the post-intervention measures are being taken, that is, immediately following or after some follow-up period. For now, though, what is the effect of the intervention being compared with or to?

Alan. We are planning to have a control group receiving so-called 'usual care' for the injury, that is, standard physiotherapy. We want to see if our specially formulated nutritional supplement combined with physiotherapy is superior to standard care.

Greg. OK, so we have a two-group controlled trial, with injured participants randomly assigned to a combined supplement and physiotherapy intervention or to 'usual care'. We have not addressed this yet, but we will also need to discuss who the participants are, that is, where are they coming from and what are the eligibility criteria for the trial? Let's review, where we are at present. I now have essential background information according to what is sometimes known as the PICO framework for trial planning—Participants, Interventions, Comparisons, and Outcomes. There are a few more decisions we need to make before we can get an estimate of the numbers required for this trial. Do you have a feel for what you would consider to be the smallest clinically worthwhile effect? In other words, what difference in mean pain scores would you regard as clinically important and why? One of my colleagues, Professor Will Hopkins, remarked in a symposium presentation at the American College of Sports Medicine Annual Meeting 2004, that "if you can't answer this question, quit the field"! For some outcome variables, though, pinning down the smallest worthwhile effect is not that easy.

Alan. OK. That's a difficult one. This intervention is relatively novel, so there isn't much specific literature out there on which to base a decision. There has been some work on the improvement in self-reported pain scores that people consider meaningful, that is, a change that makes a difference to their quality of life or ability to undertake routine activities that they can detect. This change is typically about 10 units on our scale. We have also conducted a preliminary study on a small sample of injured participants, which confirms this figure So, I would say a change of 10 units, compared to usual care would be our minimum clinically important difference. Our discussions with colleagues and clinicians also support the validity of this figure.

Greg. OK, good. Now we need some handle on the within-group variability in pain scores for this primary outcome variable. What does the literature and your preliminary data suggest as an estimate for the between-subject standard deviation for pain score?

Alan. The SD is about 20 units, typically, in studies of this type.

Greg. OK, so our standardised effect size that we consider the smallest effect worth detecting clinically is about half a standard deviation. We need a few more bits of information and we're ready to go. We need to decide in advance on the risk we are prepared to accept of making Type I and II errors, and on whether we are adopting a one- or two-sided test of significance. First, is there any compelling rationale for adopting a one-sided test?

Alan. None that we can think of. This is a novel, complex intervention and we would not be certain that it could only result in an improvement compared to conventional care. To play it safe we believe that a two-sided test is more appropriate in this instance. We are also interested in the mechanisms of action of the nutrient supplement and feel that if the intervention actually worsens pain rather than improves it, then we need to investigate this further from a mechanisms perspective.

Greg. OK, all that remains is to decide on the alpha values and the required power. I suggest that we conduct a range of estimations based on alpha values of $P=0.05$ and 0.01, and power of 80 and 90%, to get a feel for the numbers we might need. We can take our standardised effect size of half a standard deviation and use the nomograms, we discussed earlier. Alternatively, we can sit and input the parameter values we have agreed upon in a sample size and power software package. Finally, is there any reason why we can't aim for an equal number of participants in each group?

Alan. No, we don't anticipate any major obstacles. Our experience in this field suggests that when the control group is an 'active control', that is, they get some treatment in the form of 'usual care', then there are few problems with recruitment and obtaining consent for randomisation.

Greg. Excellent. Let's conduct the estimations assuming that the data are suitable for parametric analysis. The outcome of interest would seem to be the difference in post-intervention pain scores between the two arms of the trial, which we will plan to analyse with an independent t-statistic, together with a confidence interval for the mean difference. We will further assume that with a reasonable sample size, and effective participant sampling and randomisation, there will be no substantial difference between the groups at baseline. However, assuming there is no substantial regression to the mean, we will explore the potential merit in using a t-statistic for differences between groups in the change scores (post-intervention minus pre-intervention). An alternative analysis allowing for chance baseline imbalances would be an Analysis of Covariance (ANCOVA), with the post-intervention scores as the dependent variable, a nominal group variable as the independent variable, and the baseline scores as a covariate. When we have generated the sample size estimates, we can then discuss what evidence there is to help us predict the potential attrition or loss to follow-up in such a trial, plus any anticipated potential problems with compliance. Often, because of these factors we will need to recruit additional numbers to those estimated to ensure we have a sufficient sample size in the end. This is a very important issue, as heavy losses to follow-up can result in being unable to detect clinically worthwhile effects. In the absence of solid information, an arbitrary cushion of an additional 10% is often adopted, but researchers must employ a range of strategies to maximise compliance and minimise attrition. OK, let's come up with some numbers and then make some decisions on the targets for recruitment.

3.2. The results of the sample size estimations

The estimations were conducted using the nQuery Advisor 5.0 software package (Statistical Solutions, Cork, EIRE). For an alpha level of 0.05, sample sizes of 64 and 86 participants per group would provide 80 and 90% power, respectively, to detect as statistically significant a standardised effect size of 0.5 standard deviations. At an alpha level of 0.01, sample sizes of 96 (80% power) and 121 (90% power) participants per group would be required to distinguish the minimum clinically important difference from the statistical null (zero effect of the intervention compared to the control).

Greg. As you can see, we have estimates ranging from 64 participants per group to 121. Let's return to the alpha and power values discussion. As you know, these decisions are not etched in stone. From your experience of research ethics committees and funding bodies in this field, are there any firm conventions or expectations regarding the alpha and beta values?

Alan. In recent years, experience suggests that there is an increasing trend to regard power of 80% as insufficient to guard against the risk of a Type II error. I would feel more comfortable with 90% power, giving us only a 10% chance of rejecting a treatment that actually is beneficial. We believe that this type of intervention has enormous potential, so we want a reasonably low probability of missing a beneficial effect if it exists.

Greg. That sounds sensible. What about the alpha value?

Alan. We are happy with 0.05. If it was good enough for Fisher it's good enough for us! There is little or no risk of harm in this treatment, and the economic cost is reasonable compared with standard care, so a false positive result would not be a disaster. We are comfortable with only a 1 in 20 chance of finding the treatment to be beneficial when in fact it is not. An alpha of 0.01 would seem unnecessarily cautious given the nature of the intervention and the participants likely to be subjected to it.

Greg. So, that leaves us with an estimated sample size (two-sided test, alpha at 0.05, power at 90%) of 86 participants in each group. This will allow us to detect a pre-defined clinically important effect of 0.5 standard deviations.

Alan. Yes, and based on my knowledge of the field I think a 10% figure for potential loss to follow-up is realistic. In our previous work, we have not had any major issues with compliance or attrition. The participants tend generally to be highly motivated, typically. As a crude approximation, if I divide 86 by 0.9 to allow for 10% attrition, that gives me a target of 96 in each group.

Greg. Yes, it's only a ballpark figure, so let's round it up to a target of 100 participants in each group. As you have said in your justification for the study, the injury you are studying is fairly common in this population, and you have good access routes to the required sample. As planning progresses, if there are anticipated problems with accrual of participants we may need to consider using more recruitment centres. Incidentally, as a rough back-of-the-envelope calculation for these two-group designs (with alpha at 0.05 and power at 90%), the required total sample size is approximately 42 divided by the standardised effect size (ES) squared. So, in your case this would be $42/0.5^2$, which is 168 participants or 84 per group-close to the estimate we got from the software. For 80% power and alpha equal to 0.05 the formula is $32/\text{ES}^2$.

3.3. Assumptions for the examples presented

The sample size estimation examples provided in the two main sections of this primer are based on a planned parametric statistical analysis framework. All else being equal, parametric tests are more powerful than their nonparametric analogues. Formulae for sample size estimations for nonparametric tests are also available. For these 'distribution free' tests, essentially the same decisions are required regarding Types I and II error rates, one-sided versus two-sided tests, and the effect size worth detecting. For the scenario presented in the example dialogue, the effect size for a nonparametric counterpart of the independent *t*-test (Mann–Whitney test) is expressed as the probability that an observation in the intervention group is larger than an observation in the control group (Noether, 1987). The null hypothesis here is that this probability is 0.5, i.e. a 50/50 chance implying that the two groups are effectively from the same population with respect to the outcome variable.

Pain scores assessed through a Visual Analogue Scale (VAS) represent the primary outcome variable in the example presented. We acknowledge the debate surrounding whether assumptions of normality, and interval-ratio level of data hold for VAS data, and thus whether parametric tests are strictly valid. However, we side with the developing consensus that such data can be analysed appropriately using parametric statistical techniques

(Dexter & Chestnut, 1995; Kelly, 2001). This recommendation also applies to the examples presented in the previous main section, using unit changes in pain scores analysed using an independent t-statistic, as Dexter and Chestnut (1995) reported that the power to detect differences between groups was not less for a categorical VAS than for a continuous VAS.

4. Estimation approaches to sample size planning: A thumbnail sketch

This section outlines an alternative approach to sample size planning, based on precision of estimation of experimental effects. It is intended to illustrate the fundamental principles of this strategy for those readers wishing to advance their knowledge. If preferred, readers may skip this section and proceed to the conclusion without loss of coherence.

The methods for sample size estimation discussed thus far in this primer are based on the traditional power analytic, hypothesis-testing approach. The aim of this method is to obtain sufficient power to distinguish a pre-specified effect from the null hypothesis. With respect to data analysis and presentation of results, readers will doubtless be aware that in recent years there has been an increasing trend towards adopting an 'estimation' approach, rather than an over-reliance on tests against the null hypothesis (see, for example, Altman, Machin, Bryant, & Gardner, 2000). This approach, which we applaud, involves the calculation and presentation of confidence intervals (usually 95 or 90%) interpreted generally as the likely range for the 'true' effect in the population from which the sample was drawn. The confidence interval approach facilitates the interpretation of the clinical, practical, or mechanistic significance of findings, depending on the context.

Due to the increasing popularity and relevance of this estimation-based method, attempts have been made to extend this thinking into sample size estimation. Instead of specifying power to distinguish a given effect size from the statistical null, these methods require the pre-specification of a target confidence interval width. For the example given in the previous section, of a minimum clinically important difference of 10 units on the pain scale, we could, for instance, specify a target 95% confidence interval width of 20 units. The sample size that would afford this degree of precision could then be estimated from standard formulae or appropriate software packages. Presuming that the observed difference between groups was also 10 units, the target width would be expected to span from zero difference (10−10) to 20 units (10+10). The researcher may thus believe that this sample size is just sufficient to detect the smallest worthwhile effect of 10 units (i.e. to distinguish the effect from no difference as the confidence interval does not overlap the null value of zero). The sample size estimation using this approach for our example requires the following information: the required width or half-width of the two-sided 95% confidence interval (20 units, or half-width of 10 units) and the variability for the measure (standard deviation of 20 units). This results in a required sample size of just 31 in each group, compared with 86 in each group estimated from the power analytic method.

Rearrangement of the conventional power analysis formula reveals that a sample size of 31 in each group would provide only 50% power to detect the smallest worthwhile effect—a beta value or Type II error rate of 0.5. Therefore, we would be accepting a risk of a 'failed alarm', or missing a truly beneficial intervention, of 50%. This apparent anomaly occurs, in part, due to the fact that the variability (standard deviation) inputted into the sample size equation a priori is only an estimate of the actual variability exhibited in the subsequent study. Therefore, the actual observed confidence interval—calculated from the study data—may be shorter or longer that the target width. On average, the observed confidence interval would be expected to be wider (and hence include the value of zero difference—the null) 50% of the time. Furthermore, the true location of the parameter of interest (the mean difference in pain scores between groups) is not accounted for. Various methods have been advanced in an attempt to address these perceived problems, but all still result in what many regard as unacceptably small sample sizes, providing only approximately 50–65% power to detect the smallest worthwhile effect. The reader is referred to Daly (2000), and associated references, for a fuller discussion. Of course, one may choose a target confidence interval width that is not based solely on distinguishing the experimental effect from the statistical null. For example, a target width of 10 units centred around the observed effect would require 123 participants in each group. As a rule of thumb, halving the confidence interval width (doubling the precision of estimation) requires approximately a four-fold increase in sample size (123 versus 31 participants per group).

In our worked example in the previous main section, we estimated the sample size required to detect a standardised effect of half a standard deviation with 90% power, using a two-sided test. This sample size is sufficient to distinguish this effect from the mean between-group difference of zero assumed under the null hypothesis. However, as suggested by Kelley et al. (2003), this sample size may not define the observed effect precisely. Assuming a noncentral t-distribution for the standardised effect size (Cumming & Finch, 2001) the 95% confidence interval for an effect of 0.5 standard deviations with a sample size of 86 participants in each group ranges from 0.2 to 0.8. Therefore, although we may observe an effect size that we would define as 'moderate' using Cohen's (1988) criteria, the 'true' effect in the population could be anything from small (0.2) to large (0.8).

It is possible, of course, to combine the power analytic and confidence interval width approaches to address such issues. Kelley et al. (2003) detail the accuracy in parameter estimation approach (AIPE). As noted, we could choose a sample size such that the expected width of the confidence interval was sufficiently narrow. Let us assume that we desire a confidence interval width of 0.4 (half-width of 0.2). If this interval were centred around an observed effect size of 0.5, the lower and upper bounds would be 0.3 and 0.7. Although still a relatively wide interval, we could arguably distinguish crudely our observed effect from either small (0.2) or large (0.8+) effects, as these values lie outside the 95% confidence interval. This degree of precision of estimation would require approximately 200 participants in each group. Even more participants would be needed if we wanted to address the issue of under- and over-coverage of the observed confidence interval described in the previous paragraph. This apparently large discrepancy between the estimates from the power analytic approach and the AIPE approach only occurs when we are trying to detect moderate to large effects. With small effects, the required sample size to distinguish the experimental effect from the null value with reasonable power is also more than adequate for defining the observed effect precisely. For example, detecting a standardised effect size of 0.2 (two-tailed $P=0.05$, power=90%) requires 527 participants per group. The 95% confidence interval for an effect size of 0.2 with this sample size is approximately 0.1–0.3—sufficiently precise to define the 'true' effect as 'small'.

5. Conclusions

It is beyond the scope of a single article to address all of the specific hurdles that may be encountered in sample size planning. Hopefully, we have provided general concepts and principles that transfer to many different scenarios and research designs. Our example illustrated by the dialogue was for a parallel group randomised controlled trial. The underlying messages, however, translate to other common designs. For instance, if the design is repeated measures or a crossover type, the measure of variability required for the estimates is the standard deviation of the expected differences, rather than the between-subject standard deviation. All other parameter value issues are the same.

What if your outcome variable is a dichotomous percentage or proportion, such as the percentage 'in pain' versus 'pain free', in place of a continuous primary outcome? In this instance, we would need to specify the percentage 'in pain' in the control group and define the minimum clinically important effect as some difference from this proportion. For example, 'on *usual care* we expect 50% of the control group to remain 'in pain' at follow-up, compared to only 30% of the intervention group'. Other than this, decisions regarding alpha, beta, and one-sided versus two-sided tests are essentially equivalent to the example presented. For more complex designs including, for example, cluster randomised trials and trials with pre-planned sub-group analyses sample size estimation is yet more complex, and we would advise that expert assistance be sought in these instances. Indeed, whatever the research design, it is prudent to have one's sample size estimations checked and verified by an experienced biostatistician or measurement specialist.

In this primer, we have attempted to demystify the murky world of sample size estimation. Rather than present formulae and number-crunch through hand-worked examples, we elected to focus instead on the *decision-making process*. Nomograms, tables, and dedicated statistical software for sample size estimation are widely available. In our experience the confusion, particularly for novice researchers, lies in the decisions and assumptions that have to be made before the formulae can be employed. We hope that increased understanding of these broad issues will help researchers and research consumers alike. We believe that a thorough and honest approach to sample size estimation is vital in planning all research, perhaps particularly health intervention research. However, we urge researchers, expert peer—reviewers of articles and research funding proposals, journal editors, ethics committee representatives, funding body committee members—indeed, all of the 'gatekeepers of knowledge'—to maintain a balanced view of sample size estimation. As Bachetti (2002, p. 1271) remarked:

> Because of uncertainties inherent in sample size planning, reviewers can always quibble with sample size justifications-and they usually do. The information needed to determine accurately the "right" sample size (a murky concept in itself) is often much more than available preliminary information.

References

Altman, D. G. (1991). *Practical statistics for medical research*. London: Chapman & Hall.

Altman, D. G., Machin, D., Bryant, T. N., & Gardner, M. J. (2000). *Statistics with confidence* (2nd ed.). Bristol: BMJ Books.

Atkinson, G. (2003). Does size matter for sports performance researchers? *Journal of Sports Sciences*, *21*, 73–74.

Atkinson, G., & Nevill, A. M. (1998). Statistical methods in assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Medicine*, *26*, 217–238.

Atkinson, G., & Nevill, A. M. (2001). Selected issues in the design and analysis of sport, performance research. *Journal of Sports Sciences*, *19*, 811–827.

Bachetti, P. (2002). Peer review of statistics in medical research: The other problem. *British Medical Journal*, *324*, 1271–1273.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). London: Lawrence Erlbaum Associate Publishers.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*, 532–574.

Dalton, G. W., & Keating, J. L. (2000). Number needed to treat: a statistic relevant for physical therapists. *Physical Therapy*, *80*, 1214–1219.

Daly, L. E. (2000). Confidence intervals and sample sizes. In D. G. Altman, D. Machin, T. N. Bryant, & M. J. Gardner (Eds.), *Statistics with confidence* 2nd ed. (pp. 139–152). Bristol: BMJ Books.

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

Dexter, F., & Chestnut, D. H. (1995). Analysis of the statistical tests to compare visual analog scale measurements among groups. *Anesthesiology*, *82*, 896–902.

Everitt, B. S., & Pickles, A. (2004). *Statistical aspects of the design and analysis of clinical trials*. London: Imperial College Press.

Fisher, R. A. (1950). *Statistical methods for research workers*. London: Oliver and Boyd.

Hopkins, W. G., Hawley, J. A., & Burke, L. M. (1999). Design and analysis of research on sport performance enhancement. *Medicine and Science in Sports and Exercise*, *31*, 472–485.

Hudson, Z. (2003). The research headache—answers to some questions (editorial). *Physical Therapy in Sport*, *4*, 105–106.

Kelley, K., Maxwell, S. E., & Rausch, J. R. (2003). Obtaining power or obtaining precision: Delineating methods of sample-size planning. *Evaluation and the Health Professions*, *26*, 258–287.

Kelly, A. M. (2001). The minimum clinically significant difference in visual analogue scale pain score does not differ with severity of pain. *Emergency Medicine Journal*, *18*, 205–207.

Knottnerus, J. A., & Bouter, L. M. (2001). The ethics of sample size: Two-sided testing and one-sided thinking. *Journal of Clinical Epidemiology*, *54*, 109–110.

Machin, D., Campbell, M., Fayers, P., & Pinol, P. (1997). *Sample size tables for clinical studies* (2nd ed.). Oxford: Blackwell Science.

Noether, G. E. (1987). Sample size determination for some common nonparametric statistics. *Journal of the American Statistical Association*, *82*, 645–647.

Peace, K. E. (1988). Some thoughts on one-tailed tests. *Biometrics*, *44*, 911–912.

Pocock, S. J. (1996). In P. Armitage, & H. A. David, *Clinical trials: A statistician's perspective*. *Advances in biometry* (pp. 405–422). Chichester: Wiley.

Rice, W. R., & Gaines, S. D. (1994). Heads I win, tails you lose-testing directional alternative hypotheses in ecological and evolutionary research. *Trends in Ecology and Evolution*, *9*, 235–237.

Senn, S. (1997). *Statistical issues in drug development* pp. 169–186. Chichester: Wiley.

Sterne, J. A. C., & Smith, G. D. (2001). Sifting the evidence—what's wrong with significance tests? *British Medical Journal*, *322*, 226–231.

Vickers, A. J., & Altman, D. G. (2001). Analysing controlled trials with baseline and follow up measurements. *British Medical Journal*, *323*, 1123–1124.

Whitley, E., & Ball, J. (2002). Statistics review 4: Sample size calculations. *Critical Care*, *6*, 335–341.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

Williamson, P., Hutton, J. L., Bliss, J., Blunt, J., Campbell, M. J., & Nicholson, R. (2000). Statistical review by research ethics committees. *Journal of the Royal Statistical Society A*, *163*, 5–13.