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.