Journal Title
XX(X):2–6

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Robust, value-based sample size determination for clinical trials when nuisance parameters are unknown

Duncan T. Wilson[1]

Abstract

Background: The conventional approach to determining the sample size of a clinical trial is to choose the smallest value such that the power of the trial is above some nominal threshold, often 0.8 or 0.9. This sample size can be highly sensitive to nuisance parameters, such as the variance of a continuous primary outcome. Sample size re-estimation methods allow these parameters to be estimated at an interim analysis to allow the sample size to be adjusted accordingly. However, these methods do not formally account for the associated costs of increased sampling, and as a result can lead to incoherent decisions.

Methods: We present an alternative model for sample size determination which explicitly balances costs and benefits by introducing a value function to be maximised. We explore the implications of the model and argue it provides a better representation of sample size determination in practice than the conventional approach. We show the method is significantly less sensitive than the conventional approach to nuisance parameters, to the point where a fixed design with no interim sample size adjustment can be near-optimal for large regions of the nuisance parameter space. We propose a criterion for choosing an optimal fixed sample size, considering the range of nuisance parameter values for which the value of the fixed design is within a tolerable distance of the value of the best possible design.

Results: We illustrate our approach by applying it to two trial design problems: choosing the accrual and follow-up times for a parallel group trial comparing overall survival, where the median survival time in the control arm is unknown; and choosing the number of clusters in a cluster randomised trial with unknown variance components at both the individual and cluster levels.

Conclusion: Accounting for the costs of sampling when determining the sample size of a clinical trial, we can find simple, fixed sample size designs which are highly robust to nuisance parameter uncertainty.

Prepared using sagej.cls

Keywords
Clinical trials, sample size, power, interim analysis

1 Introduction

Prepare and motivate.

Need to know: - Standard SSD procedure (target difference, nuisance parameters, constrained optimisation) - Implication of ignoring costs, especially as nuisance params vary - The sample size samba, and how this accounts for cost informally - Inability to do the samba in an SSR setting

Motivation: - Conventional SSR leads to incoherent decisions, specifically to trials which are very inefficient in the sense that they are too large for the benefit they bring - Need a new approach for SSD which can be appied in SSR coherently and transparently and which leads to more efficient trials

No real distinction between SSD and SSR, under the ideal case of perfect estimates. Question is just what sample size should be chosen for a given parameter value. We want a system for making that decision that's coherent. The current convention isn't, because it doesn't formally account for costs. It survives because people game the process, but this gaming relies on adjusting the target difference and we can't do that in the SSR setting. So, applying conventional SSR, we get interim decisions that do not cohere with the initial SSD decisions (e.g. deciding the same power is suddenly worth a lot more sampling).

We propose a simple method for accounting for cost and benefit through defining a value function and maximising it. We show how this leads to coherent decision making across nusiance parameter values, and therefore how it provides a basis for SSR. We then show that under our formulation fixed designs can be highly robust to nuisance parameter values and as such SSR can be avoided altogether, saving logisitcal headaches, money, keeping analyses simple (no multiple testing adjustments), and in many cases making the trial feasible (since the parameter can't be estimated with mch precision, e.g. hazard rate or ICC). We go on to suggest criteria for finding the best fixed design. We illustrate the general approach by applying to two examples.

2 Body

First, introduce some notation and describe conventional SSD and SSR more formally. We will consider a specific example, of a two-sample t-test with size α

[1]Leeds Institute of Clinical Trials Research, University of Leeds, Leeds, UK

Corresponding author:
Duncan T. Wilson, Clinical Trials Research Unit, Leeds Institute of Clinical Trials Research, University of Leeds, Leeds, LS2 9JT, UK
Email: d.t.wilson@leeds.ac.uk

Prepared using sagej.cls *[Version: 2017/01/17 v1.20]*

of a normally distributed outcome, with a common sd σ, balanced sample size n, and a difference in means to be detected of size μ. We would like to minimise n subject to the constraint $\beta(n, \mu, \sigma, \alpha) \leq \beta^*$, where $\beta(n|\mu, \sigma, \alpha)$ is the type II error rate for the trial with n, and β^* is the constraint level, typically 0.1 or 0.2.

Generally, σ is not known and is replaced by an estimate $\hat{\sigma}$, which may be from external data, internal pilot data, or expert judgement. The difference to be detected, μ, should be the MCID.

We assume that all other things being equal, a higher powered trial is preferred to a lower powered trial; similarly, a trial of small sample size is preferred to a larger one. Under these assumptions, the constrained optimisation approach for choosing n under different estimates $\hat{\sigma}$ is incoherent.

2.1 Value-based design

An alternative approach to SSD is to plot the power of the trial as a function of its sample size and to choose the point deemed to best balance cost (i.e. sample size) and benefit (i.e. power). To help describe this decision-making process formally, we can introduce a *value* function $v(n, \beta) : \mathbb{N} \times [0, 1] \to \mathbb{R}$ which describes our preferences $\succ$ between all possible pairs of sample size and type II error. Specifically, v is such that

$$(n, \beta) \succ (n', \beta') \Leftrightarrow v(n, \beta) > v(n', \beta').$$

Now, if we have a function $f(n)$ that returns the type II error rate obtained for any given choice of sample size (conditional on the remaining aspects of the trial design) then we can define our optimal sample size as

$$n^* = \arg\min_{n \in \mathbb{N}} v(n, f(n)).$$

What might the value function look like? Consider first a slice of the function with power fixed at some value β_1, denoted $v_{\beta_1}(n)$. Clearly this will be a strictly decreasing function in n, for all β_1. We propose that it will also be linear, and that the gradient is the same for all β_1 - that is, the cost of increasing the sample size by some amount Δ is independent of the starting point (n_1, β_1). If a similar linear structure holds for $v_{n_1}(\beta)$, then we have a constant rate of substitution between n and β. In this case a suitable value function will be of the form

$$v(n, \beta) = \beta + \lambda n,$$

corresponding to linear indifference curves. The optimal sample size can then be obtained by plotting the power curve for the problem at hand and finding the point where the tangent has gradient of λ (approximately). For example, consider a two-sample t-test to detect a difference in means of size 1, an outcome standard deviation of $\sigma = 1$, and a value function with $\lambda = 0.025$. The power curve is plotted in Figure 1, and the optimal sample size is $n = 17$ (giving a power of 0.807).

As the figure suggests, if our assumption about the form of the value function holds then we can determine the value of λ by simply plotting the power

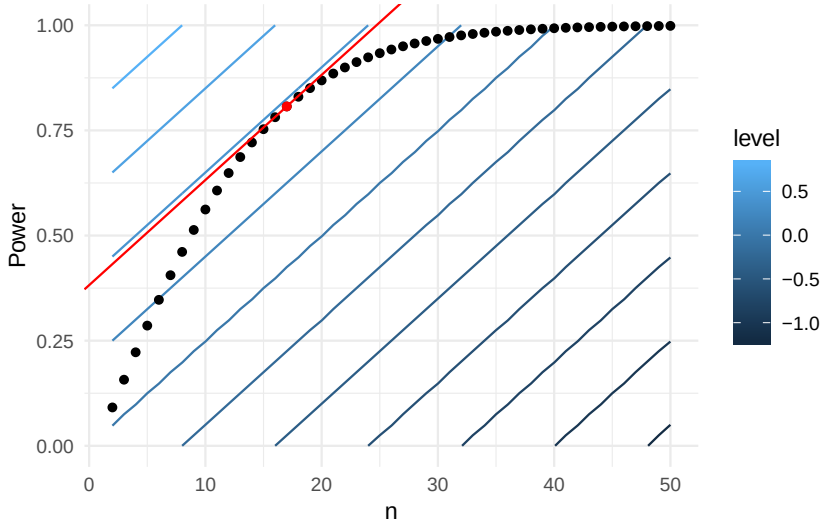

Figure 1. .

curve and choosing n; λ is then the gradient of the tangent of the power curve at that point. Is this an accurate representation of our preferences? Do we aggree with its implications? Note that, for example, $\lambda = 0.025$ implies that $(n = 0, \beta = 0) \sim (n = 40, \beta = 1)$, and so 40 is the maximum sample size we would ever consider. When it comes to determining λ, we could use any hypothetical "power" curve and if our assumptions hold we should always choose the sample size that gives the same tangent gradient. For example, consider two other power curves obtained by changing the standard deviation from 1 to 0.7 and 1.3:

On the face of it, this approach seems to be totally inconsistent with the usual method of setting a nominal power to detect our MCID of $\delta = 1$ at (say) 80%. Under that model, an initial guess of $\sigma = 1$ gives a sample size of $n = 17$, and if we were told our initial estimate of $\sigma = 1$ was incorrect and that actually $\sigma = 1.3$, we should then revise the sample size to $n = 27$ - not $n = 18$ as suggested by our model. But, the common practice of revising the MCID to fit the sample size that is considered 'feasible' suggests that the two methods may actually agree. So, we might reason that $n = 27$ is not feasible and unlikely to be funded, and pretend that our MCID is actually $\delta = 1.23$. Then, $n = 18$ will give us 80% power for this new MCID, but 61% for the original.

2.2 Known nuisance parameters

Normative model is also descriptive, justifying the 'sample size samba'.

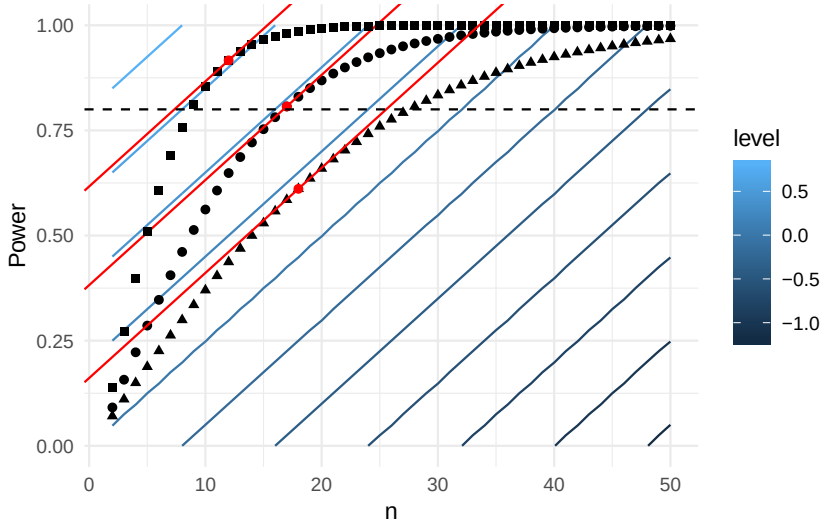

Figure 2. .

2.3 Unknown nuisance parameters

3 Examples

3.1 Cluster RCT

3.2 Survival

4 Discussion

Ethics of underpowered and overpowered studies; the threshold myth.

[?] - Sample size calculations: should the emperor's clothes be off the peg or made to measure?

[?] - Why "underpowered" trials are not necessarily unethical

[?] - Sample-size calculations for trials that inform individual treatment decisions: a 'true-choice' approach

[?] - The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies

[?] - Blinded and unblinded sample size reestimation procedures for stepped-wedge cluster randomized trial

[?] - Statistics And Ethics In Medical Research: III How Large A Sample?

[?] - Conservative Sample Size Estimation in Nonparametrics

Acknowledgements

Acknowledgements.

Declaration of conflicting interests

The Authors declare that there is no conflict of interest.

Funding

This work was supported by the Medical Research Council [grant number xxx].