Journal Title
XX(X):2–5
©The Author(s) 0000
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Robust, value-based trial design when nuisance parameters are unknown

Duncan T. Wilson[1]

Abstract

Background:
Methods:
Results:
Conclusion:
Motivating problem is uncertainty in nuisance parameters, particularly those which are hard to get interim estimates of.
SSD / SSR with the conventional contrained formulation hgighly sentive to the estimates, so large variability in the final n (and when estimates are poor, large variability in the final power too). Can lead to cases where a very large sample is requested at the interim. Need to decide how large a sample is tollerated, but no guidance on how to do so. Lack of consistency, in the sense that we can "pay" vastly different costs (sample sizes) to obtain the same value (power).
We suggest an alternative fomulation of the SSD / SSR problem, based on maximising value rather than minimising cost s.t. constrained power, where value is a weighted sum of sample size and power.
We argue that this is a better model of SSD, both in a normative and descriptive sense (including the "sample size samba" of MCID adjustment), compared to the constrained optimisation model.
We show that under this model, in many problems there will be little justification for SSR - the initial fixed design will not be very sub-optimal.
The implication is that trial designs robust to variation in nuisance parameters can be obtained without the practical and logistical trouble associated with interim analyses.
We illustrate the approach by applying it to two SSD problems: choosing the number of clusters in a cluster randomised parellel group trial with uncertain variance components at both the individual and cluster levels; and choosing the accrual and follow-up time periods in a parellel group trial with an overall survivial endpoint and uncertain event rate in the control arm.

Keywords
Clinical trials, simulation, sample size, power

1 Introduction

? - 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

2 Value-based trial design

Running example throughout - two sample t-test of means with outcome sd being the nuisance parameter.

The dominant paradigm in clinical trial sample size determination is based on power, the probability of rejecting a null hypothesis conditional on an alternative hypothesis being true. Assuming that the type I error rate is fixed, the two objectives of minimising sample size and maximising power are in conflict. The conventional solution to resolving this is to declare some threshold level of power which must be met, and then find the smallest sample size that satisfies this criteria. As has been previously discussed by others, a major flaw of this approach is that the cost associated with sampling is ignored.

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 $v(n, \beta) : \mathbb{N} \times [0,1] \rightarrow \mathbb{R}$ 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

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

[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

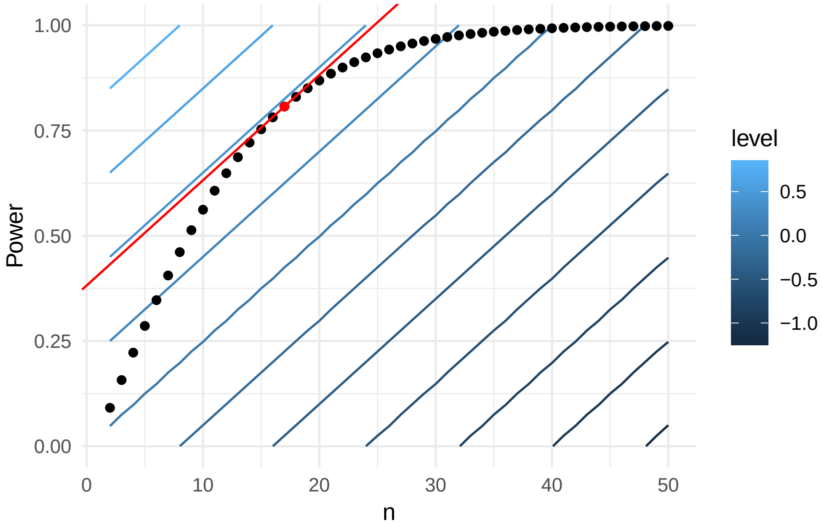

Figure 1. .

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 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

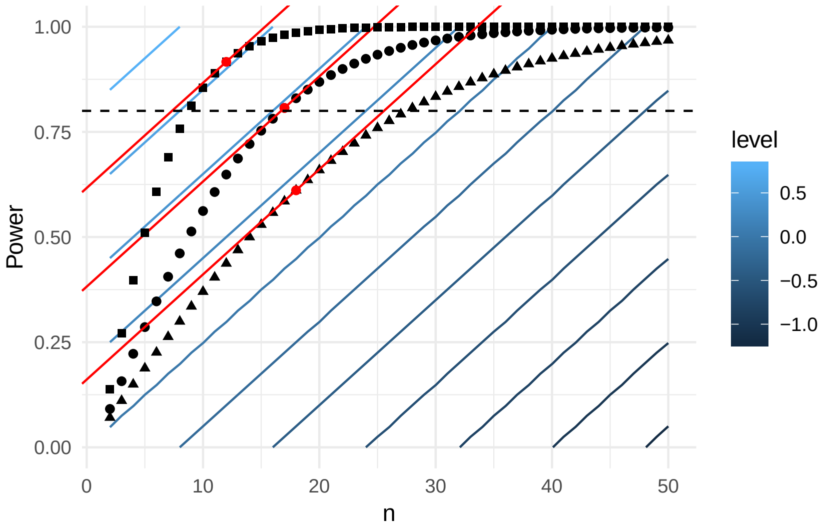

Figure 2. .

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.

2.1 Known nuisance parameters

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

2.2 Unknown nuisance parameters

3 Examples

3.1 Cluster RCT

3.2 Survival

4 Discussion

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].