Discussion Papers in Economics

Power Analysis and Sample Sizes: A Binding Frontier Approach.

Subrato Banerjee

November 2015

Discussion Paper 15-04

Indian Statistical Institute, Delhi
Economics and Planning Unit
7, S. J. S. Sansanwal Marg, New Delhi 110016, India

Power Analysis and Sample Sizes: A Binding Frontier Approach*

Subrato Banerjee†

November 2015

Abstract

I introduce a completely new (non-parametric) approach to determine (satisficing) sample sizes, when dynamic sampling and stopping rules are not feasible, and when no assumption can be made on the underlying (and unknown) distribution(s) that is (are) believed to generate observed experimental data. The method proposed, that relies on the construction of a 'binding function', is shown to be a general solution for a class of problems associated with decision functions that frequently interest (lab and field) experimental researchers in the areas of economics and psychology.

Keywords :Sampling theory, Hypothesis testing, Power analysis.

JEL Classification : C12, C18, C83, C91, C92.

*I would like to thank Prof. Bharat Ramaswami for all his support and guidance throughout my research. This paper has also benefited immensely from the comments and feedback received by Profs. K. R. Parthasarathy, Isha Dewan, Probal Chaudhuri, and Arindam Chatterjee.

†Research scholar at the Indian Statistical Institute, SJS Sansanwal Marg, New Delhi - 110016, India. I can be contacted at (email) subrato8r@isid.ac.in.

1 Introduction

The motive of this paper is to construct a function that is an 'upward' (that is, from the North-East direction) bound for an infinite family of functions that explain how Type I errors respond to Type II errors (or conversely). The purpose behind such a construction (which we call as the *Binding Frontier*) is to invert this function and fix the probabilities of Type I and Type II errors to implicitly solve for a satisficing sample size. This will ensure that our 'errors are contained'. Methods of sample size determination that weigh Type I errors against Type II errors almost always rely on assumptions on the functional forms of densities under the null and the alternative hypotheses, or some methods to estimate those densities (or use Bayesian priors). I will demonstrate that with binding frontiers, one neither needs to assume any density functional form, nor does one need to estimate any density (or use Bayesian priors).

Adaptive/dynamic sampling procedures that rely on stopping rules have important applications to surveys of animal, plant, mineral, and fossil-fuel resources and may also have applications to other fields such as epidemiology and quality control (see Thompson (2012)).[1] These methods, however, have limited application in the fields of experimental economics and psychology, where the nature of experiments are such that the entire focus of the researcher is required by the process of data generation,[2] so that any analysis of the said data can only be done at a later stage. In short, it is practically impossible to generate data and do the analysis at the same time, given the human and managerial limitations. It is thus, often advisable to be 'prepared' with several pilot experiments before conducting the final experiment to generate the required data. Clearly, the sampling requirements for experiments in economics and psychology are often different from those of the *purer* sciences.

Researchers frequently impose assumptions (mostly that of normality) in the distribution of outcome variables (in the parametric approaches), or rely on the conditions of asymptotic normality (in many non-parametric approaches) to determine sample sizes (see Beal (1989); Cochran (1977; 2009); Chow et al. (2008); Kraemer and Thiemann (1987); Noether (1987);

[1] These processes of determination of (expected) sample sizes have been discussed in Wald (1947), and Berger (1985), and are observed to have advantages over non-sequential methods in terms of lower expected sample sizes, and more controllable precision (see Thompson (2012)). Chow et al. (2008) report that sample-size calculations based on the above processes (see Pocock (1977); O'Brien and Fleming (1979); and Chick and Frazier (2012)) have functional forms very similar to the central expression proposed in this paper (although they are not free from terms involving $z_{\alpha/2}$, and z_{β}, of the normal variate).

[2] The process of data generation includes (carefully) reading out experimental instructions to each group (often large enough) of subjects, interviewing subjects, making them fill questionnaires (this is also true for field data collection required by household surveys for income, consumption and other data) among still other facets of organization that require constant attention. Simply put, organizing experiments and ensuring their smooth running, often tend to be very involving processes. Given the already heavy managerial requirements on data generation, the additional task of on-the-spot sample size determination could be daunting.

and Thompson (2012)). Alternatively, researchers also frequently allocate subjects equally (uniformly) among the different treatment groups. In this paper, we ask if it is possible to work out a feasible sample-size when no assumption can be made about the distribution of the outcome variable in question. I propose a non-parametric approach that does not rely on asymptotic normality in the determination of sample sizes.

2 Tests Concerning Means

Let $Z_1, ..., Z_n$, constitute a random sample from a population given either by a common density (or distribution) f_0, with the parameter vector $\mathbf{\Phi}_0$ and mean μ_0 (under the null hypothesis, H_0), or f_1, with the parameter vector $\mathbf{\Phi}_1$ and mean μ_1 ($> \mu_0$, under our alternate hypothesis, H_1). Further, while $\mathbf{\Phi}_0$ and $\mathbf{\Phi}_1$, need not have the same dimension, we assume that the variances of the two population densities are identical (although higher order moments defining skewness, kurtosis and so on, may significantly differ). The random variable $\bar{Z} = \frac{Z_1+...+Z_n}{n}$, is accordingly assumed to have mean μ_0 (under H_0, and associated with the density $\hat{f}_0$, with the parameter vector $\hat{\mathbf{\Phi}}_0$) or μ_1 (under H_1, and associated with the density $\hat{f}_1$, with the parameter vector $\hat{\mathbf{\Phi}}_1$). For example, if $f_0(z; \mathbf{\Phi}_0)$ is the normal density with $\mathbf{\Phi}_0 = (\mu_0, \sigma_0)$, then $\hat{f}_0(\bar{z}; \hat{\mathbf{\Phi}}_0)$ is the normal density with $\hat{\mathbf{\Phi}}_0 = (\mu_0, \sigma_0/\sqrt{n})$. Our decision rule is: do not reject H_0, if $\bar{z} \leq \mu_0 + c$, our critical value (and reject otherwise). Thus, our decision function assumes the following (simple) form

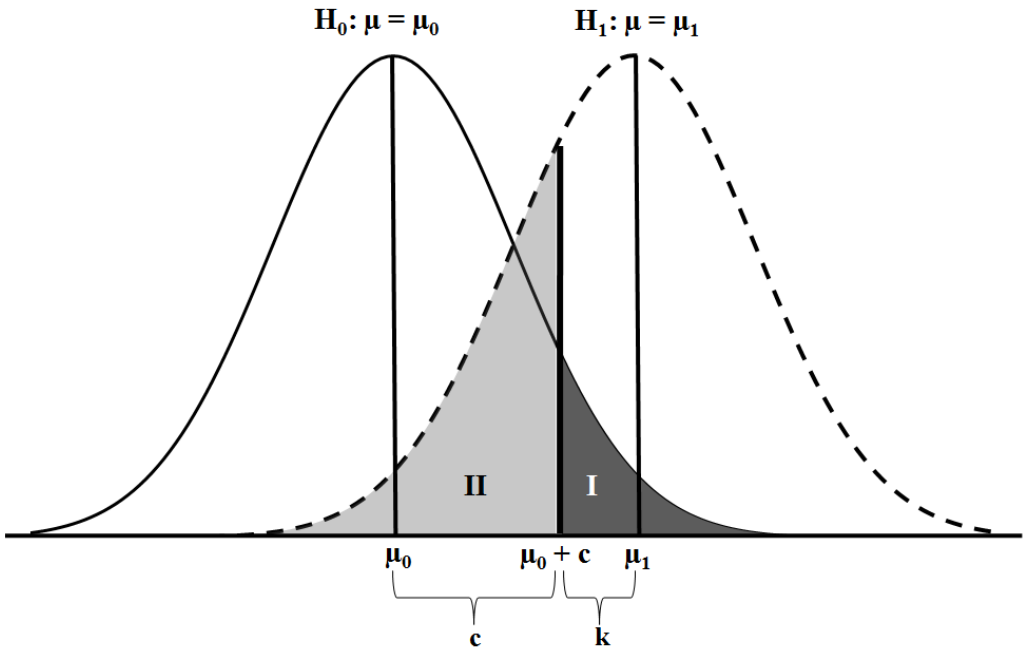
$$d(\bar{z}) = \begin{cases} \mu_0 \text{ for } \bar{z} \leq \mu_0 + c \\ \mu_1 \text{ for } \bar{z} > \mu_0 + c \end{cases} \tag{1}$$

We are looking for a general sample size expression for a family of density pairs $\hat{f}_0$ and $\hat{f}_1$, that have identical variance[3] (although still higher order moments associated with skewness and kurtosis, and so on, may be different) and are permissible for our testing procedure associated with $d(\bar{z})$ above.[4] Our test is illustrated in Figure 1 where the dark shaded region is the probability of a Type I error, and the light shaded region is the probability of a Type

[3]The implicit assumption here being that the effect of any experimental treatment is only limited to the first moment of any density and that the higher order moments remain unchanged. However, the derivations we present, more generally account for changes in the functional forms of densities brought about by the treatment, provided that the variance remains the same (that is, we still allow for higher order moment changes in skewness, kurtosis and so on). For example, the introduction of mid-day meal programs in schools are implicitly assumed only to shift students' mean grades upward, and that the dispersion around the higher mean grades remain the same. Additionally see List et al. (2011).

[4]This rules out other densities that may be associated with outcomes of clinical trials that often produce a multitude of results (different effects of a pill etc.), and consequently need alternate descriptions of Type I and Type II errors.

Figure 1. The Type I and Type II Errors



II error.

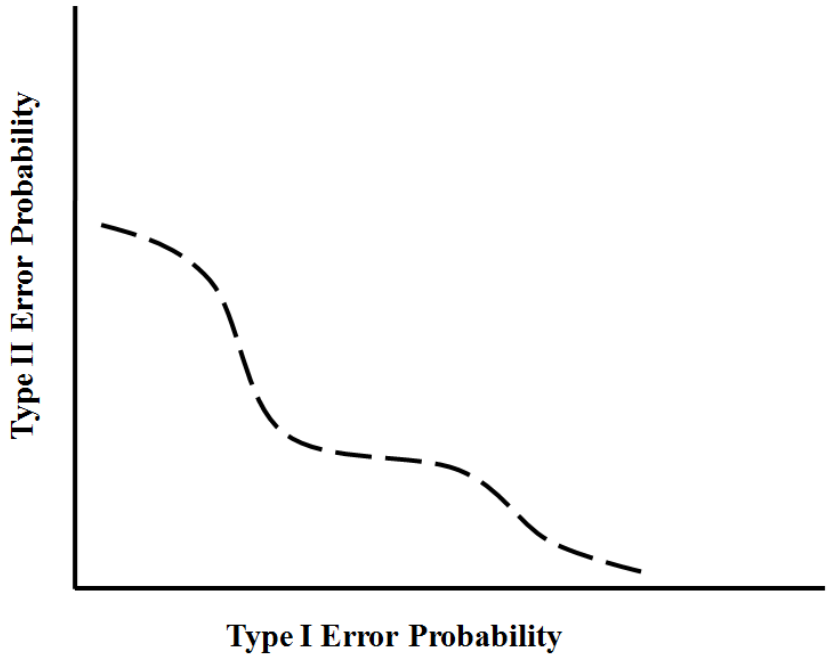
2.1 The Error Tradeoff Frontier

Let p_i be our Type i error ($i \in \{I, II\}$). It is clear that the size of the Type I and II errors (i.e. the values of p_I and p_{II}), would require the calculation of the areas of the shaded regions in Figure 1. Further, the calculation of the shaded regions itself requires the knowledge of the exact functional forms of the densities $\hat{f}_0$ and $\hat{f}_1$ (and therefore the information on $\hat{\mathbf{\Phi}}_0$ and $\hat{\mathbf{\Phi}}_1$). Suppose we know these functional forms, then for each value of c (the critical distance from the mean assumed under the null), it is possible to get a pair of realizations (p_I, p_{II}). Each pair is a unique point on the Type I-Type II error space. Thus, varying c, generates a locus of points (p_I, p_{II}), which we call as the *Error Tradeoff Frontier* as shown in Figure 2.

The *Error Tradeoff Frontier* therefore, is an exact understanding of how one of the errors responds to changes in the other. In particular, how p_{II} responds to p_I above, can be summarized by the function $g_{(n),\hat{f}_0,\hat{f}_1} : (0,1) \longmapsto (0,1)$, where $p_{II} = g_{(n),\hat{f}_0,\hat{f}_1}(p_I)$.[5] It is clear that the function g should be indexed by the density pair $\hat{f}_0$ and $\hat{f}_1$, since the functional

[5]Note that we have used open intervals here, for we are not particularly interested in the end-points.

Figure 2. The Error Tradeoff Frontier



form of g would depend on that of $\hat{f}_0$ and $\hat{f}_1$ (every point on the frontier $p_{II} = g_{(n),\hat{f}_0,\hat{f}_1}(p_I)$, essentially represents a pair of shaded regions obtained from the areas under $\hat{f}_0$ and $\hat{f}_1$ in the relevant intervals). We now establish the result that $g_{(n),\hat{f}_0,\hat{f}_1}$ is non-increasing.

Proposition 1 *The Error Tradeoff Frontier* $g_{(n),\hat{f}_0,\hat{f}_1} : (0,1) \longmapsto (0,1)$ *is always non-increasing for any density pair* $\hat{f}_0$ *and* $\hat{f}_1$ *permissible under* $d(\bar{z})$.

Proof. Trivial. Let $\hat{F}_0$ and $\hat{F}_1$ be the respective cumulative distributions of $\hat{f}_0$ and $\hat{f}_1$. Pick any c' associated with a given (p'_I, p'_{II}). Now choose any $c'' > c'$ so that $p''_I \leq p'_I$ (where p''_I is associated with c''). This follows from the fact that $\hat{F}_0$ is non-decreasing. Similarly, $p''_{II} \geq p'_{II}$ (where p''_{II} is associated with c'') follows from the fact that $\hat{F}_1$ is non-decreasing. Thus if $p''_I \leq p'_I$ then $p''_{II} \geq p'_{II}$ along the *Error Tradeoff Frontier*. This completes the proof. ■

At this stage, we are tempted to believe that methods of sample size determination that weigh Type I errors against Type II errors[6] must necessarily need assumptions on the functional forms of $\hat{f}_0$ and $\hat{f}_1$, or rely on some methods of estimating them (or use Bayesian

[6]See Lipsey (1990). Useful perspectives on this 'act of balancing' can be found in Brown (1983); Cascio and Zedeck (1983); Nagel and Neef (1977); and Schneider and Darcy (1984).

priors). I will demonstrate that with the *Binding Frontier* (the final frontier that we will construct and also the one that we are ultimately interested in), this is not the case.

2.2 Family of Error Tradeoff Frontiers

It is clear that for a given n (since we have a fixed sample), a decrease in one error would not decrease the other. In reality, we can only make conjectures on the specifications of density pairs $\hat{f}_0$ and $\hat{f}_1$ (and do not observe either). In general, the exact shape and position of a given frontier $g_{(n),\hat{f}_0,\hat{f}_1}$, will depend on the functional forms of the density pairs (how one error exactly responds to the other depends on the density specifications under H_0 and H_1). Figure 3 displays this non-increasing relation between p_I and p_{II} for four different (arbitrary) density pairs named A, B, C and D that share the same variance and sample size and are permissible under $d(\bar{z})$. The first challenge is that since there are an infinite number of frontiers for a given sample size, we do not know what the 'outermost' (in the North-Eastern region) frontier looks like, since if we did, then we would want to just fix p_I and p_{II}, and invert that frontier to implicitly solve for n (this will be explained in more detail in a while). The second challenge is that we do not know if the outermost frontier is uniquely determined. As shown in Figure 3, while for some points (p_I, p_{II}), frontier D is the outermost frontier, for others frontier C is the outermost frontier. The next section addresses these challenges and presents the main result of this paper.

3 The Binding Frontier

We ask if it is possible to construct a function $\omega_{(n)} : (0,1) \longmapsto (0,1)$ that binds every frontier for all points (p_I, p_{II}), from the North-East direction for a given sample size n. In other words, we are looking for an upper envelope or a band which is independent of the specifications of $\hat{f}_0$ and $\hat{f}_1$. Figure 4 illustrates this problem.

A first impression is that a frontier should exist, for the function $p_{II} = \omega_{(n)}(p_I) = 1 - p_I$, does work well for a valid frontier. However, such a frontier is not meaningful since points like $(0.5, 0.5)$ on the error space are not very useful. We will thus need additional requirements on $\omega_{(n)}$. The most obvious of these is that it should be dependent on n (what else are we implicitly solving for?). The second additional requirement will be that once we invert $\omega_{(n)}$ to implicitly solve for n, the sample size, it must be non-increasing in both p_I and p_{II} (as is desirable of other known sample size formulae), since the more confident we want to be (the lower the p_i), the higher will be the data requirement. A final requirement is that the expression of n must be scale-invariant (say, a function of $\sigma/(\mu_1 - \mu_0)$).

Figure 3. The Error Tradeoff Frontiers

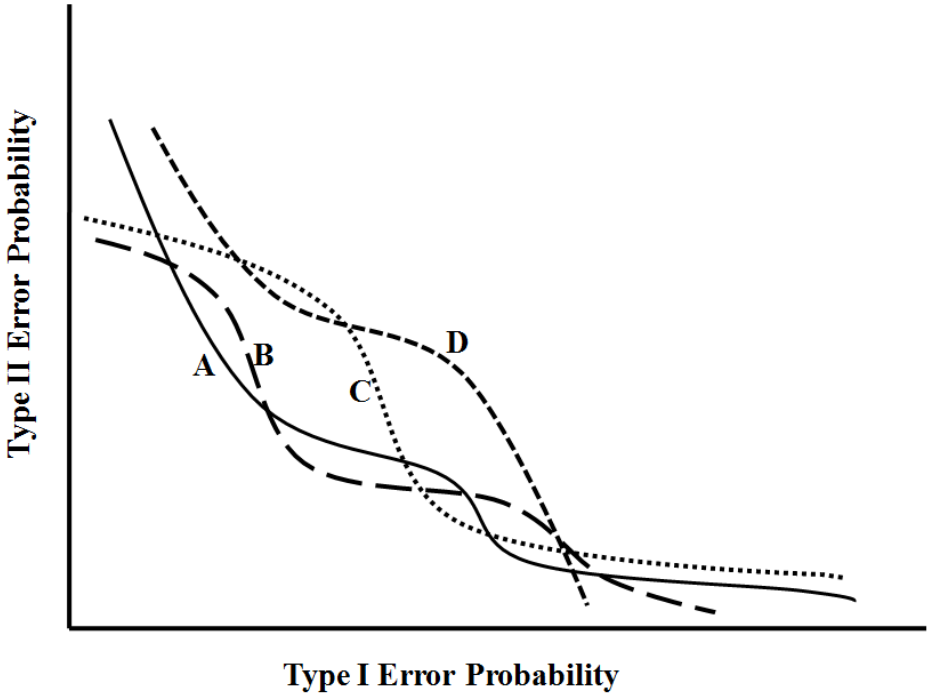
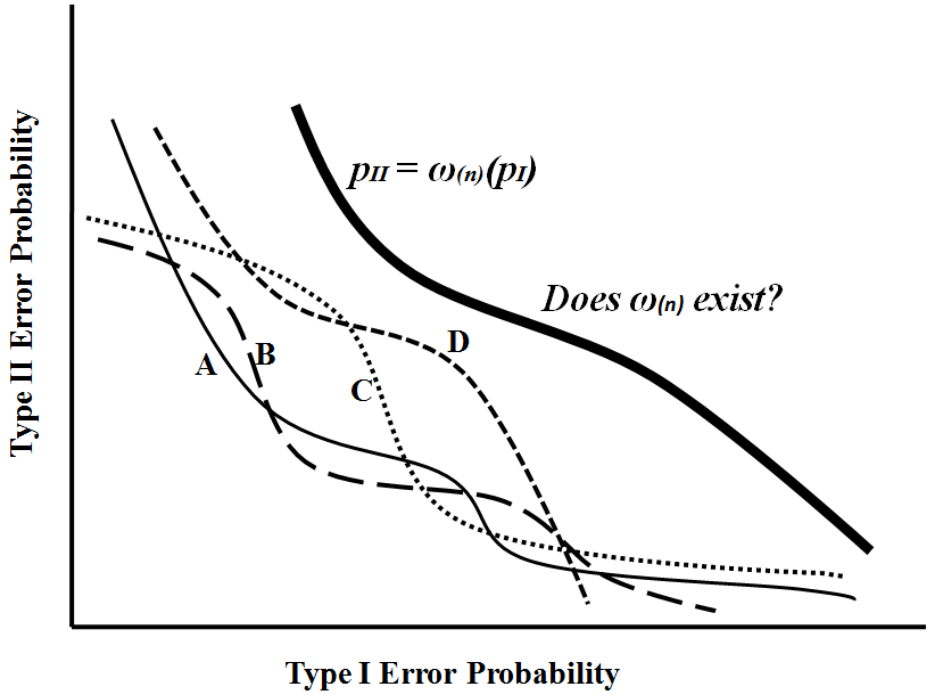


Figure 4. The Binding Frontier



3.1 The Construction

Let α be the (maximum permissible) size of the Type I error. Let c be a non-negative constant such that $p_I = 1 - \hat{F}_0(\mu_0 + c) = P(\bar{Z} - \mu_0 > c|\mu = \mu_0) \leq \alpha$. In other words, according to our decision function, the null is rejected whenever $\bar{Z} > \mu_0 + c$. We now prove the following lemma.

Lemma 1 *The probability of a Type I error does not exceed α when we fix α equal to $\frac{\sigma_Z^2}{nc^2}$.*

Proof. The inequality $P(\bar{Z} \leq \mu_0 + c) \geq P(\mu_0 - c \leq \bar{Z} \leq \mu_0 + c) \geq P(\mu_0 - c < \bar{Z} < \mu_0 + c)$ follows from the fact that the LHS spans more values. Further, the Chebyshev's inequality guarantees that $P(\mu_0 - c < \bar{Z} < \mu_0 + c) = P(|\bar{Z} - \mu_0| < c) \geq 1 - \frac{\sigma_Z^2}{nc^2}$. These two inequalities can be combined to get $P(\bar{Z} \leq \mu_0 + c) \geq 1 - \frac{\sigma_Z^2}{nc^2}$. Finally, subtracting each side of this inequality from unity, gives us $P(\bar{Z} - \mu_0 > c|\mu = \mu_0) \leq \frac{\sigma_Z^2}{nc^2}$. The LHS of this final inequality is in fact the probability of a Type I error. Fixing α equal to $\frac{\sigma_Z^2}{nc^2}$ completes the proof. ■

It is important to note that we made no assumption on the parameter vector $\hat{\mathbf{\Phi}}_0$, or the functional form of the density $\hat{f}_0$ (and hence the distribution function $\hat{F}_0$), in the determination of p_I (the Type I error) above. The value of c above, can be thought of as the critical distance from the mean assumed under the null hypothesis. We similarly define k to be the critical distance from the mean assumed under the alternate hypothesis, and apply the Chebyshev's inequality to prove the following lemma defining β to be the (maximum permissible) size of the Type II error.

Lemma 2 *The probability of a Type II error does not exceed β when we fix β equal to $\frac{\sigma_Z^2}{nk^2}$.*

Proof. Trivial. The steps involved are exactly the same as in the proof of Lemma 1 above. For $\mu_0 = 0$, the steps have been deferred to the Appendix, and make no assumption on the parameter vector $\hat{\mathbf{\Phi}}_1$ or the functional form of density $\hat{f}_1$ in the calculation of $p_{II} = \hat{F}_1(\mu_1 - k)$. ■

Figure 1 illustrates how c and k are related as per the requirements of our decision function, with the regions of Type I and Type II errors labeled respectively as I and II. We are now in a position to prove the central theorem of this paper.

Theorem 1 *Let $t = \frac{(\mu_1 - \mu_0)}{\sigma_Z}$, then $\beta = \omega_{(n)}(\alpha) = \left(\frac{1}{(t\sqrt{n}) - (1/\sqrt{\alpha})}\right)^2$ is a binding frontier.*

Proof. *From Lemma 1, we have $P(\text{Type I error}) \leq \frac{\sigma_Z^2}{nc^2} = \alpha$, or $c = \frac{\sigma_Z}{\sqrt{\alpha n}}$. From Lemma 2, we have $P(\text{Type II error}) \leq \frac{\sigma_Z^2}{nk^2} = \beta$. Putting k equal to $\mu_1 - \mu_0 - c$, where $c = \frac{\sigma_Z}{\sqrt{\alpha n}}$ (from Lemma 1), as per the requirements of our decision function (see Figure 1) completes the proof.* ■

The theorem above, needs an explanation. $t = \frac{|\mu_1 - \mu_0|}{\sigma_Z}$ is the treatment effect.[7] It should be noted that while I have written $g_{(n),\hat{f}_0,\hat{f}_1}$ in terms of p_I and p_{II}, I have intentionally written $\omega_{(n)}$, in terms of α and β, because it helps us immediately distinguish both the errors from their respective upper bounds (α and β are (by construction) respectively the upper bounds on p_I and p_{II}). Formally, what we have shown is that $\int_R \omega_{(n)}(\alpha)d\alpha \geq \int_R g_{(n),\hat{f}_0,\hat{f}_1}(p_I)dp_I$ for any permissible region $R \subset (0,1)$.[8] Now the purpose of our construction is clear, since all the *Error Tradeoff Frontiers* and the *Binding Frontier* are shown for a *given sample size* in Figure 4. One can therefore, choose a point (α', β') on the *Binding Frontier*, and stay assured that p_I and p_{II} are captured under the outermost *Error Tradeoff Frontiers* within the rectangle that (α', β') forms with the axes. Algebraically, this is equivalent to fixing the values of α and β on the curve $\beta = \omega_{(n)}(\alpha)$, and implicitly solving for n to ensure that *'the errors are contained'* for that given sample size. We present this application of Theorem 1 in the following corollary.

Corollary 1 *If sample size is determined according to the rule:*

$$n_\omega = \frac{\sigma_Z^2}{(\mu_1 - \mu_0)^2}\left(\frac{1}{\sqrt{\alpha}} + \frac{1}{\sqrt{\beta}}\right)^2, \tag{2}$$

then the statements $P(\text{Type I error}) \leq \alpha$, and $P(\text{Type II error}) \leq \beta$, are simultaneously true, regardless of the functional forms of the densities $\hat{f}_0$ and $\hat{f}_1$.

Proof. Follows trivially from writing the *Binding Frontier* as $\beta = \omega_{(n)}(\alpha) = \frac{\sigma_Z^2}{n(\mu_1 - \mu_0 - \frac{\sigma_Z}{\sqrt{\alpha n}})^2}$ and solving for n. ■

We solve a quadratic equation in $\sqrt{n}$ above and focus on the greater root (that is $\sqrt{n_\omega}$) since the other root may take negative values, for often there are cases (particularly in the areas of drug-screening and genome analysis) where $\alpha > \beta$ is more desirable than the converse.[9] Now, before we trace out our *Binding Frontier*, it will be helpful to discuss a direct application of the term n_ω in equation (2) above.

---

[7] t here is similar to the expression known as *Cohen's d,* which is known as the 'effect size' (see Cohen (1977)).

[8] This statement only means that since $\alpha \geq p_I$ and $\beta \geq p_{II}$ whenever $\beta = \omega_{(n)}(\alpha)$ and $p_{II} = g_{(n),\hat{f}_0,\hat{f}_1}(p_I)$, the area under the function $\omega_{(n)}$ in the error space should not be less than the area under $g_{(n),\hat{f}_0,\hat{f}_1}$ (as is our requirement). Note that we have implicitly assumed here that $g_{(n),\hat{f}_0,\hat{f}_1}$ is well behaved. When integrating $g_{(n),\hat{f}_0,\hat{f}_1}$ is not straight forward (we only know that $g_{(n),\hat{f}_0,\hat{f}_1}$ is non-increasing and nothing more), one may represent the area with a summation sign (rather than the integral sign). The term $\sum_R g_{(n),\hat{f}_0,\hat{f}_1}(p_{Ij})\delta p_{Ij}$, for example, shows that the summation is done over the index j, in the region R of the integral $\int_R \omega_{(n)}(\alpha)d\alpha$.

[9] An example is worth mentioning here. What would be more a more harmful outcome for a person interested in buying a car? Predicting that a car was a bad buy when it was not (incorrect rejection of the null)?; or predicting that a car is not a bad buy when it actually is (incorrect failure to reject the null)?. Clearly, one would desire $\beta < \alpha$ here. To provide yet another example of multiple sclerosis (MS) trials in the

3.2 An Application: Two Person Bargaining Games

To see how this works in the context of a two-person bargaining game,[10] in the expression for n_ω above, we fix the (maximum) probabilities of Type I error (α) and Type II error (β) to be 0.05 and 0.10 respectively. For this bargaining protocol involving a pair of subjects, we need a unique random variable that represents the outcome of any given pair. Let x_j^i represent the share of the jth subject in the ith pair ($j \in \{1, 2\}$). So the ith pair of shares is represented by (x_1^i, x_2^i), where $x_1^i + x_2^i = 1$. Since $|x_1^i - 0.5| = |x_2^i - 0.5|$, we can define, without loss of generality $z_i = |x_1^i - 0.5|$. Then let $\bar{Z} = \frac{Z_1 + \ldots + Z_n}{n}$ (where n is the number of observed pairs). $\bar{Z}$ measures the average deviation of the negotiated shares from the equal division solution $(0.5, 0.5)$. Suppose that the population mean of this variable is μ_0. Now, consider the test of the null hypothesis that $\mu_0 = 0$ (i.e. the equal division solution is the population mean). The question is: what would be the minimum sample that is required for such a test to have reasonable power against an alternative hypothesis that the population mean is $\mu_1 > 0$? We will, as an example, consider the alternative hypothesis to be $\mu_1 = 0.02$. It is clear that the sample size that has reasonable power for this alternative hypothesis would also have at least that much power for any $\mu_1 > 0.02$. We have $\alpha = 0.05, \beta = 0.10$, $\mu_0 = 0$ and $\mu_1 = 0.02$. The only limitation is that we do not know the value of σ_Z. This will, in general, involve getting an estimate from a previous (pilot) study under conditions similar to the final experiment. To estimate σ_Z, we use a pilot study that had 14 subjects (7 pairs) in the control group. In this sample, $\hat{\sigma}_Z = 0.0075592$. Using this value gives us $n^* = 8.33$ ≈ 9 pairs (18 subjects). Note that c equals 0.01 for this value of n. In other words, with just 18 subjects, we can be 95% confident that the average outcome is the 50%-50% split (and not a 51%-49% split).

In other words, when we fix α and β equal to 0.05 and 0.10 respectively in the expression for n_ω, we are assured that the actual probabilities of Type I and Type II errors (whatever they equal) do not exceed these respective values, regardless of the densities $\hat{f}_0$ or $\hat{f}_1$, that could have potentially generated the data (in general, we need information on the functional forms of $\hat{f}_0$ and $\hat{f}_1$ to calculate the actual probabilities of Type I and Type II errors).

An important point concerns the estimation of σ_Z. According to Thompson (2012), "a bothersome aspect of sample size formulas such as these is that they depend on the

area of biostatistics, the current standard of care for MS is nearly worthless, so it is reasoned that as long as a new drug is not explicitly harmful, then there is no loss in accepting it to replace the current standard: thus the Type I error can be quite large. Conversely, these patients really need quality treatment, so it is unacceptable to fail a drug that works: hence the Type II error should be small (5% or lower).

[10] These are simultaneous move games where two players engage in face to face negotiation to determine how to split a pie of unit size between themselves. They share the pie as per the negotiation. If they fail to reach a negotiation, each individual gets nothing.

population variance, which generally is unknown. In practice, one may be able to estimate the population variance using a sample variance from past data from the same or a similar population." (*Sampling, Chapter 4, Pages 54-55*).[11] The use of stopping rules and dynamic sampling is often impractical and out of the question for most experimental economists. This is simply because most experiments are such that they need to be conducted before the data analyses. The process of dynamic sampling, on the other hand, requires that the data are analyzed as and when they are generated during the experiment to determine sequential estimates of sample sizes. This undoubtedly adds several challenges to the already involving task of conducting an experiment.

A further important point is to notice that n_ω is a linear transformation of $\hat{\sigma}_Z^2$, and therefore a random variable itself. One can therefore, easily work out 95% confidence intervals for n_ω, directly from the bootstrapped confidence intervals for $\hat{\sigma}_Z^2$ (see Thompson, 2012, and Mooney and Duval, 1993).

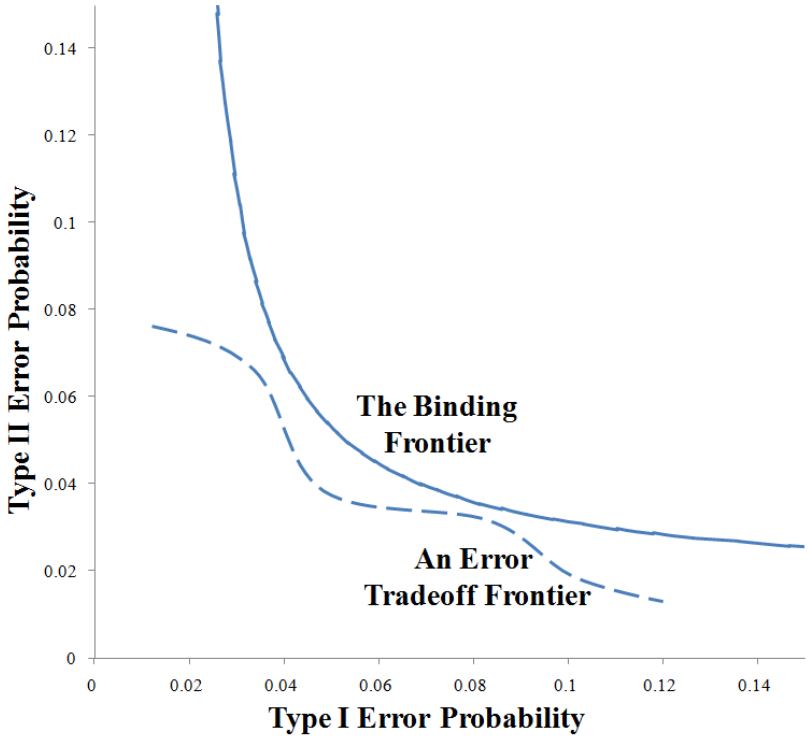
The above calculations can be easily replicated for ultimatum and dictator games, and those involving sequential bargaining (based on the pilot results). For a general J- person bargaining experiment (one where J individuals of a group bargain over a unit pie), one may apply the method above with $z_i = |x_1^i - (1/J)|/J$, or the mean deviation for each group of J individuals. We have explained the calculations in this subsection with $J = 2$ above.

4 Properties of the Sample Size Expression

In this section, we discuss some of the properties satisfied by the expression n_ω, that are commonly known for other (standard) sample size estimates (See Gore (1981); Kirby et al. (2002); Cochran (2009); and Thompson (2012)). n_ω is clearly decreasing in α and β (a property that the other (smaller) root did not satisfy), meaning that more confidence will require an increase in the sample size. It is also clear that the higher the variance observed in the pilot study, the greater will be the required sample size in the final experiment for given α and β. The fact that n_ω is also decreasing in the *'mean gap'* (a term with which

[11] To address another (and a fortunately a less problematic) issue, one might also argue that the given formula above, also requires fixing a given value of μ_1 and consequently needs justifying a well-identified value of the same. Fortunately, this is not a serious problem, for it is always feasible to fix μ_1 arbitrarily close enough to μ_0, and work out the sample size accordingly. The chosen process will accordingly have at least that much power for any value of μ_1, greater than that selected by us. For instance, the choice of $\mu_0 = 0$, in our bargaining example above, can be interpreted as 'absolutely fair'. $\mu_1 = 0.50$ can be interpreted as 'absolutely unfair'. A choice of $\mu_1 = 0.05$, can be interpreted as 'mostly fair'. We immediately know then that any $\bar{z}$, significantly different from $\mu_1 = 0.05$ will also be significantly different from $\mu_1 = 0.50$. In other words, a population known not to compromise on any element/aspect of fairness will surely not settle on dictatorial outcomes. Such a problem is thus, frequently circumvented by experimenters involved even in other fields of drug testing and clinical trials (see Thompson (2012) and Chow et al (2008)).

Figure 5. The Binding Frontier



we refer to the difference $|\mu_1 - \mu_0|$) implies that the narrower the difference between the means assumed under the null and the alternate hypotheses, the greater will be the sample size requirement.[12] Lastly, the control group size is scale-invariant: since n_ω is a function of $\frac{\sigma_Z}{|\mu_1 - \mu_0|}$.

Finally, in Figure 5, I show one possible *Error Tradeoff Frontier,* alongside our *Binding Frontier* associated with the application discussed in the previous section. The equation $\beta = \omega_{(n)}(\alpha)$ can be written as follows

$$\left(\frac{1}{\sqrt{\alpha}} + \frac{1}{\sqrt{\beta}}\right) = t\sqrt{n_\omega} = \frac{|\mu_1 - \mu_0|\sqrt{n_\omega}}{\sigma_Z} \tag{3}$$

We fix $t\sqrt{n_\omega}$ above by fixing the values of μ_0, μ_1, n_ω, and σ_Z at 0, 0.05, 2 and 0.0075593 respectively (as in our example), and let α and β vary according to the above rule. This generates a locus of points represented by the solid line in Figure 5. This is the *Binding Frontier.* The dashed-line represents an *Error Tradeoff Frontier*, which we would have (say) observed had we made assumptions on the actual distributions of $\bar{Z}$ under H_0 and H_1. It is clear that, for a given sample size, since α is an upper bound to the probability of Type I

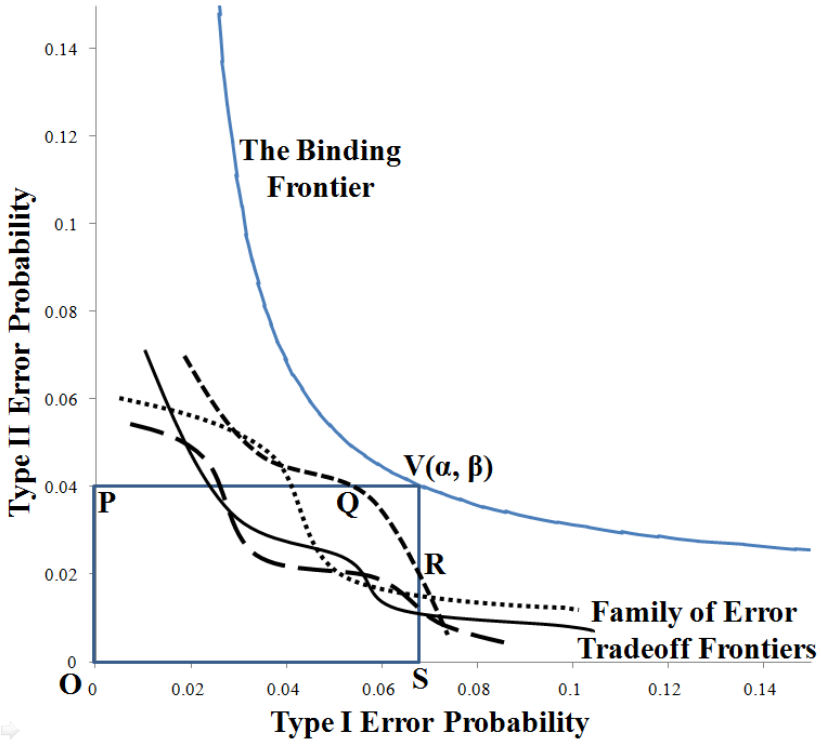
[12]The denominator in the expression for $\sqrt{n_\omega}$ involves $\mu_1 - \mu_0$, whenever $\mu_1 > \mu_0$ and involves $\mu_0 - \mu_1$, whenever $\mu_0 > \mu_1$. Both the cases can be combined to write $\sqrt{n}$ in terms of $|\mu_1 - \mu_0|$.

error, and β, an upper bound to the probability of Type II error, the *Error Tradeoff Frontier* will always be bounded by the *Binding Frontier* (guaranteed by Lemmas 1 and 2). We make do with the latter, simply because the former is not observed without assumptions on the distribution. Note that an increase in n, will shift both the frontiers inward (and a decrease, outward). An increase in σ_Z, on the other hand, will shift both the frontiers outward (and a decrease, inward). An increase or decrease in the *mean gap* will have a similar effect as that of the sample size n. An intersection between the two frontiers (at a fixed pair of Type I and Type II error levels) can only be observed, if the *Binding Frontier* is associated with a sample size greater than that of the *Error Tradeoff Frontier.* This difference between the sample sizes can be thought of as a 'cost' of not knowing the actual distribution of Z (or the cost of not making an assumption on the same). Since this cost will always remain, we refer to our sample size as the *satisficing sample size* since it displays a question of feasibility of arriving at a sample size given limited or no information on the distribution of Z.

5 A Comparison With Existing Approaches and the CLT

In what follows, we keep up with our practice of indexing sample size expressions for n with the functional forms of the associated *Tradeoff Frontiers.* Sample size formulae of the type $n_g = h_{\hat{f}_0,\hat{f}_1}(\alpha, \beta; \hat{\mathbf{\Phi}}_0, \hat{\mathbf{\Phi}}_1)$, are frequently dependent on α and β via terms like $z_{\alpha/2}$, and z_β (critical values of the standard normal variate), or $t_{\alpha/2,k}$ (critical value of the t-distribution with k degrees of freedom). Thus, the specifications of the underlying distributions become important inputs in the determination of sample sizes (and the shapes as well as the positions of the *Error Tradeoff Frontiers*). These specifications are either consequences of distributional assumptions, or the reliance on asymptotic normality. In both the cases, however, the choices of α or β that translate to critical values like $z_{\alpha/2}$ and $t_{\alpha/2,k}$, involve a process of inversion of cumulative distribution functions that often yield complex expressions (and therefore add to the complexity of the functional forms of $g_{(n),\hat{f}_0,\hat{f}_1}$, the *Error Tradeoff Frontiers*). For example, for the cumulative distribution function F of the standard normal variate, a choice of $\alpha = 0.05$ translates to $F^{-1}(1-(\alpha/2)) = F^{-1}(0.975) = 1.96 = z_{\alpha/2}$, which in turn, helps us determine n_g. Although, such processes are often aided by statistical tables and computational software, there are often cases, where distributional assumptions are questionable. Both α and β appear directly in the specification of n_ω without being affected by assumptions on functional forms (of underlying distributions). The sample size expression for n_ω of this paper is fundamentally different from all other expressions known

Figure 6. The Error Tradeoff Frontiers



for n_g so far in the literature, since α and β represent *upper bounds* on Type I and Type II errors, rather than their actual (desired) realizations.

The decision of avoiding distributional assumptions, however, has its costs. To give an idea of the magnitude of the same, let us fix α and β respectively to 0.05 and 0.20.[13] The assumption that Z is distributed normally, leads us to a sample size of $15.68\sigma_Z^2/(\mu_1 - \mu_0)^2$, as against $45\sigma_Z^2/(\mu_1 - \mu_0)^2$, suggested by n_ω.[14] The latter is almost thrice that of the former that rests on normality assumptions (the *Error Tradeoff Frontier* meets the *Binding Frontier* at $(\alpha, \beta) = (0.05, 0.20)$). Using the former sample size expression always entails a risk that Z actually belongs to a different (from normal) population, and may hence require a greater sample size for $(\alpha, \beta) = (0.05, 0.20)$. The advantage of using the expression for n_ω, is that it is a natural upper bound on *all sample-size expressions* that emanate from the knowledge of distributions relevant to context - the sample size we choose will always exceed the minimum size required by *any* density pair $\hat{f}_0$ and $\hat{f}_1$.

[13]The choice of $(\alpha, \beta) = (0.05, 0.20)$, was suggested by Cohen (1977, 1988). Related discussions are found in Ray and Vermeulen (1999); van-Belle (2008); Freiman et al. (1986); Desu and Raghavarao (1990); and Lwanga and Lemshow (1991).

[14]See Lehr (1992) and van-Belle (2008) for why the sample size expression under the normality assumption is called a 'thumb-rule'.

6 Conclusion

This paper is about a method aimed at containing Type I and Type II errors without making assumptions on the underlying densities under H_0 and H_1 when techniques of dynamic sampling are not feasible. In short, we are not weighing one error against the other, but simply *taming both together*. Figure 6 effectively summarizes the motivation behind this paper, by putting together an arbitrary family of *Error Tradeoff Frontiers* for a given sample size, and the respective *Binding Frontier* for the same sample size. Choosing the point V on the *Binding Frontier* at $(\alpha, \beta) = (0.041, 0.068)$ fixes the maximum permissible values of p_I and p_{II}. Inverting $\omega_{(n)}$ with $(\alpha, \beta) = (0.041, 0.068)$ to implicitly work out n_ω ensures that the pairs (p_I, p_{II}) are contained in the region *OPQRS*. Areas of drug testing and drug screening involving clinical trials may benefit too from such an approach provided that the associated outcomes are systematically aligned with our decision function. This paper discusses a new approach in the determination of sample sizes for experiments in economics, psychology and other social experiments, addressing the issue that the requirements of these experiments are significantly different from those in clinical labs and scientific surveys.

7 Appendix: Proof of Lemma 2

Proof. For $\mu_0 = 0$, we substitute for c from Lemma 1 for $P(\bar{Z} < \frac{\sigma_Z}{\sqrt{\alpha n}} | \mu = \mu_1)$. Also, for any k, we know from Chebyshev's inequality that $P(\mu_1 - k < \bar{Z} < \mu_1 + k | \mu = \mu_1) \geq 1 - \frac{\sigma_Z^2}{nk^2}$. Let $k = \mu_1 - \frac{\sigma_Z}{\sqrt{\alpha n}}$, so that $P(\underbrace{\frac{\sigma_Z}{\sqrt{\alpha n}}}_{\mu_1 - k} < \bar{Z} < \underbrace{2\mu_1 - \frac{\sigma_Z}{\sqrt{\alpha n}}}_{\mu_1 + k} | \mu = \mu_1) \geq 1 - \frac{\sigma_Z^2}{nk^2}$. But we know that $P(\bar{Z} \geq \frac{\sigma_Z}{\sqrt{\alpha n}} | \mu = \mu_1) \geq P(\underbrace{\frac{\sigma_Z}{\sqrt{\alpha n}}}_{\mu_1 - k} < \bar{Z} < \underbrace{2\mu_1 - \frac{\sigma_Z}{\sqrt{\alpha n}}}_{\mu_1 + k} | \mu = \mu_1)$, since LHS spans more values. On combining these inequalities we get $P(\bar{Z} \geq \frac{\sigma_Z}{\sqrt{\alpha n}} | \mu = \mu_1) \geq 1 - \frac{\sigma_Z^2}{nk^2}$. The complement of this event leads us to $P(\bar{Z} < \frac{\sigma_Z}{\sqrt{\alpha n}} | \mu = \mu_1) \leq \frac{\sigma_Z^2}{nk^2}$. The LHS here is in fact the probability of a Type II error. Finally, fixing β equal to $\frac{\sigma_Z^2}{nk^2}$, gives us $P(\text{Type II error}) \leq \frac{\sigma_Z^2}{nk^2} = \beta$. This completes the proof. ■

References

[1] Beal S. 1989. Sample size determination for confidence intervals on the population mean and on the differences between two population means. *Biometrics* **45**: 969-977.

[2] Berger J. 1993. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag: New York, NY.

[3] Brown G. 1983. Errors, type I and II. *American Journal of Disorders in Childhood* **137**: 586-591.

[4] Cascio W, Zedeck S. 1983. Open a new window in rational research planning: Adjust alpha to maximize statistical power. *Personnel Psychology* **36**: 517-526.

[5] Chick S, Frazier P. 2012. Sequential sampling with economics of selection procedures. *Management Science* **58**(3): 550-569.

[6] Chow S, Shao J, Wang H. 2008. *Sample Size Calculations in Clinical Research.* Chapman & Hall/CRC: Boca Raton, FL.

[7] Cochran W. 1977. *Sampling Techniques.* Wiley: UK.

[8] Cochran W. 2009. *Sampling Techniques.* Wiley: New Delhi, India.

[9] Cohen J. 1977. *Statistical Power Analysis for the Behavioral Sciences.* Academic Press: New York, NY.

[10] Cohen J. 1988. *Statistical Power Analysis for the Behavioral Sciences.* Lawrence Erlbaum Associates: Hillsdale, NJ.

[11] Desu M, Raghavarao D. 1990. *Sample Size Methodology.* Academic Press: Boston, MA.

[12] Freiman JA, Chalmers TC, Smith H, Kuebler RR. 1986. The importance of beta, the type II error, and sample size in the design and interpretation of the randomized controlled trial: Survey of 71 "negative trials". In *Medical Uses of Statistics*, Bailar JC, Mosteller F (eds). New England Journal of Medicine Books: Waltham, Massachusetts.

[13] Gore S. 1981. Assessing clinical trials - trial size. *British Medical Journal* **282**: 1687-1689.

[14] Kirby A, Gebski V, Keech A. 2002. Determining the sample size in a clinical trial. *Medical Journal of Australia* **177**(5): 256-257.

[15] Kraemer H, Thiemann S. 1987. *How Many Subjects? Statistical Power Analysis in Research.* Sage Publications: Newbury Park, CA.

[16] Lehr R. 1992. Sixteen S-squared over D-squared: A relation for crude sample size estimates. *Statistics in Medicine* **11**: 1099-1102.

[17] Lipsey M. 1990. *Design Sensitivity: Statistical Power for Experimental Research.* Sage Publications: Newbury Park, CA.

[18] List, J., Sadoff, S., Wagner, M. (2011). So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design. Experimental Economics, 14(4), 439-457.

[19] Lwanga S, Lemshow S. 1991. *Sample Size Determination in Health Studies: A Practical Manual.* World Health Organization: Geneva, Switzerland.

[20] Mooney, Christopher Z., Duval, Robert D. (1993). Bootstrapping: A Nonparametric Approach to Statistical Inference (English). Newbury Park, London: Sage.

[21] Nagel SS, Neef M. 1977. Determining an optimum level of statistical significance. In *Evaluation Studies Review Annual,* M Guttentag, S Saar (eds). Sage: Beverly Hills, CA.

[22] Noether G. 1987. Sample size determination for some common nonparametric tests. *Journal of the American Statistical Association* **82**: 645-647.

[23] O'Brien P, Fleming T. 1979. A multiple testing procedure for clinical trials. *Biometrics* **35**: 549-556.

[24] Pocock S. 1977. Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**: 191-199.

[25] Ray J, Vermeulen M. 1999. Sample size estimation for the sorcerers apprentice. *Canadian Family Physician* **45**: 1732-1739.

[26] Schneider A, Darcy R. 1984. Policy implications of using significance tests in evaluation research. *Evaluation Review* **8**: 573-582.

[27] Thompson S. 2012. *Sampling.* Wiley: Hoboken, NJ.

[28] van-Belle G. 2008. *Statistical Rules of Thumb.* Wiley: Hoboken, NJ.

[29] Wald A. 1947. *Sequential Analysis.* Wiley: New York, NY.