

# Bayesian Unconditional Power Analysis

John S. Uebersax

July 12, 2007

---

## Contents

[1. Introduction](#)

[2. Power Analysis](#)

[3. The Problem](#)

[3.1 Illustration](#)

[3.2 Asymmetry of the Power Function](#)

[4. Solution](#)

[4.1 Formal Solution](#)

[4.2 Magnitude of Bias](#)

[4.3 Approximate Method](#)

[4.4 Effect on Sample Size Estimates](#)

[5. Discussion](#)

[6. Conclusions](#)

---

## Abstract

Traditional power analysis requires the stipulation of an alternative hypothesis,  $H_1$ . People often base clinical trial power calculations on the alternative hypothesis  $H_1 \theta = d$ , where  $d$  is the value of a statistic or effect size observed in a previous study. While superficially plausible, this is actually a bad idea. It ignores the effect of sampling variation associated with  $d$ . While it may be true that  $d$  is an unbiased estimate of the true effect size,  $\delta$ , power estimates based on  $H_1 \theta = d$  are usually biased. Typically this bias is manifest as a substantial underestimate of statistical power. This has been previously noted in the literature, but the full implications, which may include a radical rethinking of how power analysis is performed and, further, how clinical trials are conducted, have been insufficiently emphasized. Statistical power is more accurately estimated by calculating the posterior *pdf* of  $\delta$  given the previously observed value (or values) of  $d$ , assessing statistical power for each possible value of  $\delta$ , and, then, by means of integration, calculating what is in effect a weighted average of statistical power over all likely values. Ordinarily, this value, which we call here *true power*, is less than the classical power estimate. Larger sample sizes are needed to achieve true powers of .90 or .80 than the conventional method implies. The discrepancy between true and conventional power estimates depends on the sample size of the previous study, which determines the degree of uncertainty about  $d$ . This is an

important and fundamental statistical issue and must be understood by any statistician performing power analyses for clinical trials.

## 1. Introduction

For a pharmaceutical company, clinical trials are extremely important and extremely expensive. Sometimes they fail. If they fail due to lack of treatment efficacy, that is, in a sense, a positive outcome, because attention and resources may then be directed to more promising medicines. However a clinical trial of an effective medicine may fail because to insufficient statistical power. Here we examine a common statistical error that produces overestimates of power, resulting in clinical trials that are underpowered and therefore more likely to fail inappropriately. The topic and principles outlined here should be thoroughly understood by any statistician involved with clinical trial design.

Section 2 will briefly review the basic logic of power analysis and define terms. Section 3 will describe the error referred to above. Section 4 will present methods for avoiding the error and correctly estimating statistical power. Section 5 will present generalizations and discussion.

## 2. Power Analysis

Conventional power analysis addresses the question: for a given null hypothesis ( $H_0$ ), alternative hypothesis ( $H_1$ ), sample size, and Type I error rate ( $\alpha$  level), what is the probability of rejecting  $H_0$  when  $H_1$  is true (and  $H_0$  is false). This probability is called *statistical power*.  $1 - \text{power}$  gives the related quantity,  $\beta$ , the probability of rejecting  $H_0$  when  $H_0$  is true (and  $H_1$  is false).

Let the null hypothesis for a statistical test of treatment efficacy be:

$$H_0: \theta = \delta_0, \quad [1]$$

where  $\theta$  denotes some statistic or index of treatment effect and  $\delta_0$  denotes a specific value of this statistic or index, e.g., 0.

Let the alternative hypothesis be:

$$H_1: \theta = \delta_i, \quad [2]$$

for some specific value of  $\delta_i \neq \delta_0$ .

Classical power analysis is a *conditional statistical inference*: the estimate of power applies only conditional on the specific alternative hypothesis considered. Different values selected for  $\delta_i$  will produce different estimates of statistical power. Estimates of statistical power of a clinical trial, then, are only as accurate or relevant as insofar as  $\delta_i$  is an accurate or relevant estimate of true treatment efficacy.

In practice, a variety of strategies are used to select  $\delta_i$ . Common methods include the following:

1. *Guessing*. Sometimes a value is produced by guessing. Although the limitations of this approach are obvious, it is still used fairly often in powering clinical trials. It is ironic that, despite the great attention to rigor and standardization that goes into other aspects of drug development, power analysis of clinical trials is often conducted in a haphazard and extremely subjective way.
2. *Minimum clinically significant difference (MCSD)*. The idea here is to choose for H1 the value of  $\delta$  that corresponds to the smallest clinically meaningful effect. This is a plausible approach, but has the distinct disadvantage of being inherently conservative when the true effect size exceeds the MCSD. This leads to clinical trials that are unnecessarily large and expensive. The excess resources spent on a too-large trial could be better used to develop or evaluate other medicines.
3. *Previously observed effect size*. This approach chooses for H1 a value of  $\delta$  based on previous empirical evidence. The evidence could be the effect size observed in a previous pilot study, a Phase 2 clinical trial, or a pooled effect size observed across several previous studies. Unlike the two approaches above, this one has the merit of formally considering prior information. However, as we shall observe below, **if the previous information is summarized as a point estimate of  $\delta$  only, the resulting power analysis will be biased**, and often strongly so.

The problem associated with 3 applies regardless of whether  $\delta_i$  is based on a single study, an average effect size across several studies, or a pooled result. The important thing is that it is a point estimate, and therefore neglects to consider uncertainty about the true effect size; that is, it fails to consider sampling variation (among other things). We shall demonstrate below that, in practice, this nearly always produces an underestimate of true statistical power. The phenomenon is fairly well documented in the literature (see for example Spiegelhalter, Abrams and Myles, 2004, ch. 6), but people continue to make this mistake, resulting in underpowered, and, consequently, failed clinical trials. Part of the problem is that the discussion is sometimes

embedded in complex Bayesian statistics which obscure its fundamental and intuitively obvious nature. It is also possible that people have wished to avoid confronting this issue, as it implies a general need for larger sample sizes. Here we aim to clear the matter up once and for all: to demonstrate, explain, and quantify the bias of the conventional method, to present simple methods for estimating true statistical power, and to discuss in general terms the implications for clinical trials.

### 3. The Problem

We shall here mainly consider the case of power analysis using for H1 an estimate of treatment effect based on a single previous study. The basic principles, however, apply whenever there is uncertainty concerning true effect size. We will also specifically consider the case of a normally distributed test statistic. Again, the principles apply to many other common statistical tests with unimodal sampling distributions (e.g., tests of proportions, survival analysis, etc.).

The basic problem revolves around what can be thought of as *local asymmetry of the power function*. Suppose a researcher observes an effect size of  $d$  in a previous study. In practice, many people simply use  $\theta = d$  as the alternative hypothesis in power calculations. Clearly this ignores the fact that  $d$  is only a sample value, and only an estimate of true effect size,  $\delta$ .

A naïve rationale for this is as follows: while it is true that  $d$  is only an estimate of true effect size,  $\delta$ , it is our best estimate; it is the expected value of  $\delta$ , or  $E(\delta)$  -- and therefore it makes sense to base power analysis on this value. However this naïve view makes an implicit assumption: that the power function is symmetrical around the value of  $d$ . In reality, except at one point, where  $\theta =$  the critical value associated with the stipulated  $\alpha$  level (which is unlikely to be used as H1 in a power analysis; see below), the power function is asymmetrical. This asymmetry, as we shall see, will often or typically produce a substantial bias in conventional power analysis using H1  $\theta = d$ .

#### 3.1 Illustration

The problem is easily understood with reference to Figure 1 below.

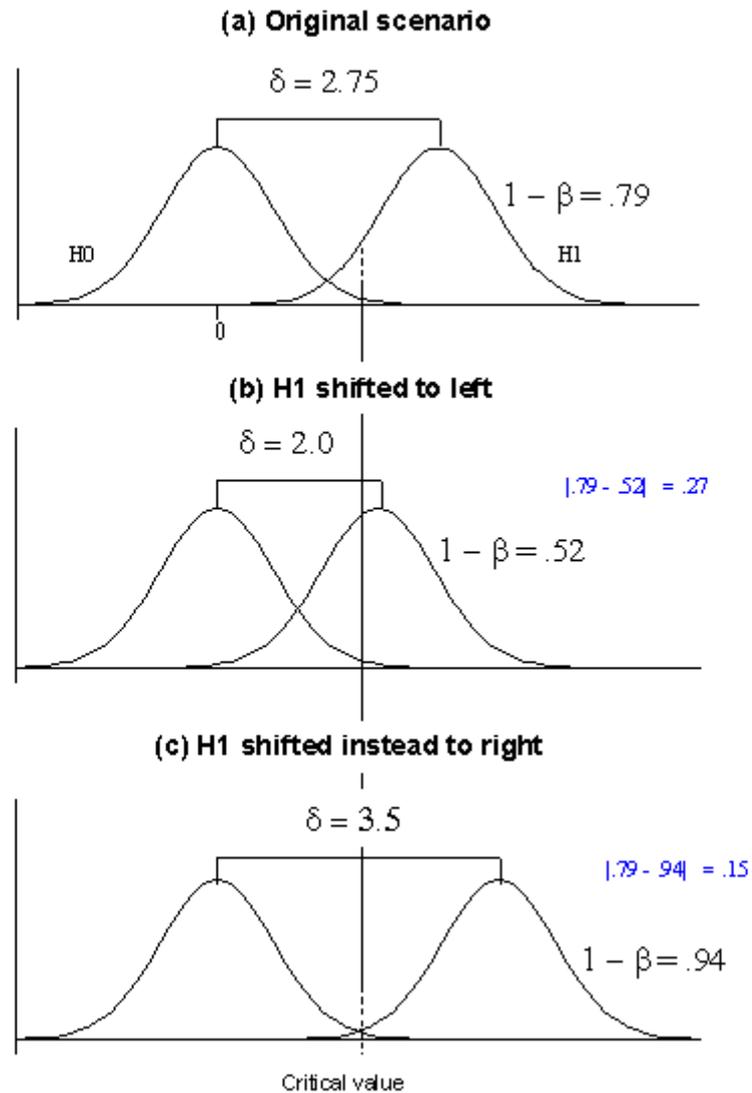


Figure 1. Unequal effect of shifting the H1 sampling distribution to the left (b) versus shifting it to the right (c).

Panel (a) illustrates the standard power analysis model. The distribution on the left is the sampling distribution of the test statistic under the null hypothesis. For example, the test statistic might be  $\theta = (\mu_2 - \mu_1) / \sigma_{\mu_2 - \mu_1}$ , comparing post-treatment means of a medication and placebo group. This distribution is relevant here only insofar as it fixes the critical value(s), or CV. For example, if we stipulate an  $\alpha = .05$  (two-tailed), our upper CV is 1.96.

The distribution on the right shows the sampling distribution of the test statistic under the alternative hypothesis, H1. For convenience, we express H1 in terms of its difference from H0, denoted by the parameter  $\delta$ . In (a),  $\delta$  is 2.75. Statistical power is equal to the proportion of the H1 sampling distribution that falls above the upper CV. In panel (a), the power is .79.

Panel (b) shows what happens if we instead consider an alternative hypothesis corresponding to  $\delta = 2.0$ . This shifts or displaces the H1 sampling distribution .75 to the left relative to its position in panel (a). As we slide this distribution to the left, more of its area falls below the critical value and into the acceptance range of H0. Thus, power (proportion of the distribution above CV) is reduced. In panel (b), only a little more than half the sampling distribution is in the H0 rejection range; power is now .52.

Panel (c) considers an alternative hypothesis corresponding to  $\delta = 3.5$ . Thus, it now shifts the original alternative hypothesis .75 units to the right instead of to the left, relative to the original scenario. Now nearly all of the H1 sampling distribution falls above the critical value; power is .94.

This plainly shows that decreasing the alternative hypothesis effect size by .75 has much more effect on estimated power than increasing it by the same amount. The negative change reduced power from .79 to .52, an absolute difference of .27. The positive change increased power from .79 to .94, and absolute difference of only .15. Relative to the original scenario, reducing  $\delta$  causes a dense region of the sampling distribution to cross the critical value, so that power is strongly affected. However, increasing  $\delta$  only affects the non-dense left tail region, making the effect on power less. The effect of a negative versus positive change of H1, therefore, is asymmetrical.

### **3.2 Asymmetry of the Power Function**

For simplicity, we shall now assume a normally distributed test statistic and a one-tailed test. The power function is then described by the *cdf* of a normal distribution. The function is illustrated in Figure 2 below. It is symmetrical considered relative to the an effect size of  $\delta = CV$  (at which point power is .50, well below the level usually considered for a clinical trial), but not symmetrical relative to any other value of  $\delta$ .

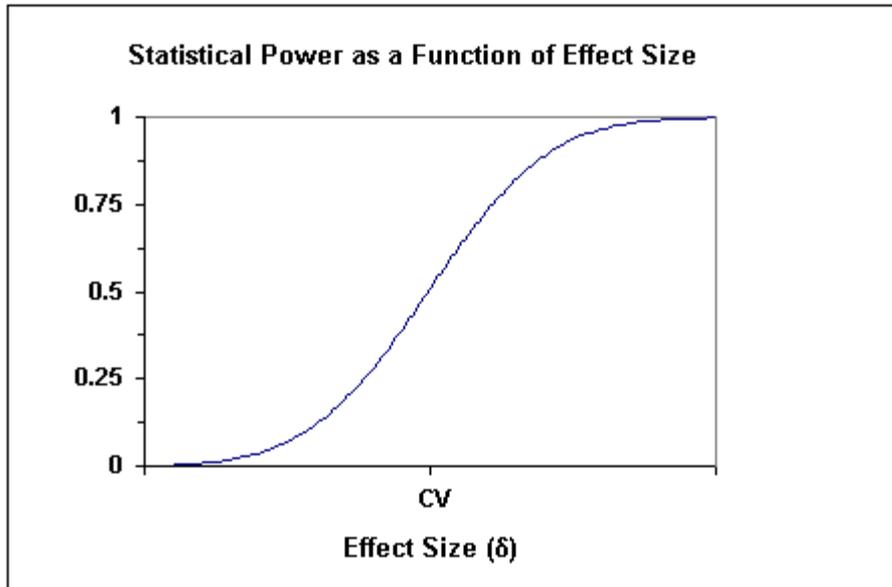


Figure 2. Statistical power as a function of effect size.

Ordinarily we are interested some region in the right half of the overall power function, i.e., that corresponding to H1 values with power  $> .50$ . The entire right half of the normal *cdf* is convex and monotonic increasing.

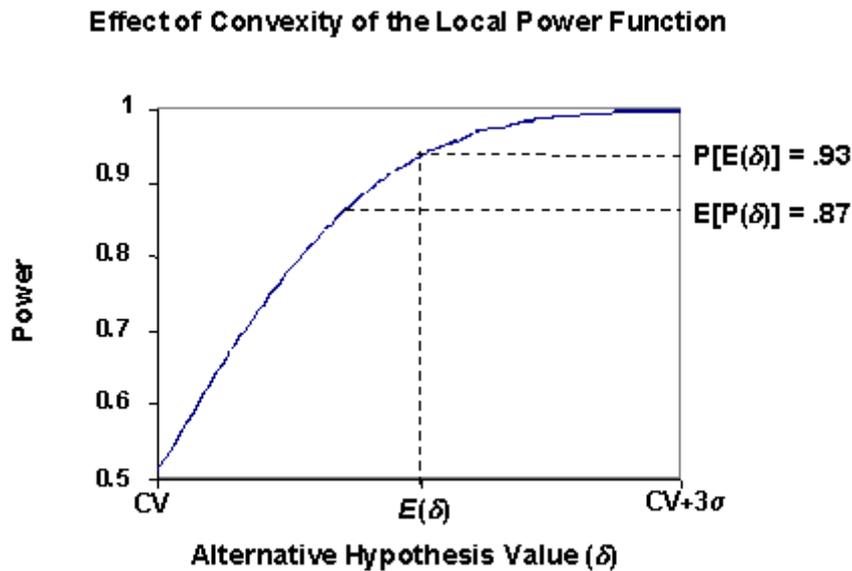


Figure 3.

This leads to an important result. Let  $E(\delta)$  be the expected value of true effect size  $\delta$  considered over some range of the  $x$ -axis. Let  $P[E(\delta)]$  denote the statistical power obtained by using this point-estimate of  $\delta$  as the alternative hypothesis. If, for example, we consider all values of  $\delta$  from  $CV$  to  $CV + 3\sigma_{\epsilon}$ ,

(where  $\sigma_{\delta}$  is the standard error of the test statistic) as equally probable, then  $E(\delta) = CV + 1.5\sigma_{\delta}$ , and  $P[E(\delta)] = .93$ .

However, we may also consider the average of  $P(\delta)$  evaluated over the same range of  $\delta$ . Doing this we find that  $E[P(\delta)] = .87$ , where  $E[P(\delta)]$  is the expected value of  $P(\delta)$ .

Due to the convexity of the power function above  $\delta = CV$ , an analogous discrepancy will occur regardless of the range of  $\delta$  considered and the shape of the averaging kernel (here we assumed all values of  $\delta$  as equally likely -- i.e., a rectangular kernel; but we could as easily consider e.g., a normal-shaped kernel), provided the kernel is symmetrical.

This leads to the important result that:

$$E[P(\delta)] \neq P[E(\delta)] \quad [3]$$

Anywhere in the right side of Figure 1, which is the range in which we are usually concerned,  $E[P(\delta)] < P[E(\delta)]$ . This means that, within this range, expected statistical power is always less than the estimate of statistical power calculated solely based on a point estimate of expected treatment effect (i.e., the classical power estimate).

$E[P(\delta)] = P[E(\delta)]$  only when the conventional statistical power estimate is .50, and  $E[P(\delta)] > P[E(\delta)]$  when the conventional estimate is less than .50 -- both unlikely scenarios for a clinical trial. Therefore within the range of power typical for a clinical trial, conventional power estimation that bases  $H_1$  on a point estimate of expected treatment effect overestimates true power. The study will have less power than was estimated, and a greater chance of failure than anticipated.

If the clinical endpoint is such that a negatively valued test statistic  $\theta$  implies treatment efficacy, then the same result occurs; we could simply multiply the test statistic times -1 and apply the same argument as above.

## 4. Solution

### 4.1 Formal Solution

The first step in estimating true statistical power is to stipulate a probability distribution that reflects uncertainty about true effect size,  $\delta$ . In a Bayesian framework, this can be understood as the posterior probability distribution of true effect size given the observed effect in a previous study (and, potentially,

other information, including a subjective or evidence-based prior distribution of treatment effect.) In a frequentist framework, we can derive an analogous *pdf* for true effect size based on consideration of the sampling distribution of the statistic; in this case, we are forming an *unconditional* estimate of statistical power, in the sense that it is not conditional on the single value of  $\delta_i = d$ .

Thus, the same theoretical and practical issue occurs regardless of whether one adopts a Bayesian or a classical (Neyman-Pearson; frequentist) approach. Further, for a normally distributed test statistic, the frequentist approach will lead to the same estimate of true statistical power as a Bayesian model with a rectangular prior. Hereinafter we shall follow a single perspective -- the Bayesian -- for convenience. The approach described below is sometimes called *hybrid-Bayes* power analysis, since it applies Bayesian models for power estimation but assumes that data will be analyzed using conventional methods (assessment of p-values, etc.). This is in contrast with a fully Bayesian approach, which applies Bayesian methods to power estimation with the assumption that data will also be analyzed and evaluated with Bayesian methods.

After specifying the probability distribution for true effect size (based on results of a pilot study and, optionally, other information), one may estimate true statistical power as the value of the following integral:

$$p(\delta) P(\delta) d\delta \quad [4]$$

where:

- $\delta$  = true (population) effect size
- $p(\delta)$  = the posterior pdf of true effect size, given an observed effect size of  $d$  in an earlier study
- $P(\delta)$  = statistical power of present study given alternative hypothesis  $H1: \theta = \delta$ .

When (a) the test statistic sampling distribution is gaussian-shaped and of constant variance and (b) the Bayesian prior is rectangular, then  $p(\delta)$  is equal to the sampling distribution.

When  $p(\delta)$  is normally distributed and the sampling distributions of the test statistic under  $H0$  and  $H1$  are normally distributed, Eq. 4 corresponds to a

convolution of normal distributions and has a closed-form solution (see Spiegelhalter, Abrams & Myles, eq. 6.4). More generally, simple numerical methods can be used to evaluate the integral.

## 4.2 Magnitude of Bias

While other sources have mentioned the bias associated with conventional power analysis based on a point estimate of treatment efficacy, few [*if any?*] have actually demonstrated and drawn attention to the extent of it. Here we show that the bias is quite appreciable.

An interesting feature is that the amount of bias depends only upon the power associated with the point estimate of efficacy and the sample size of the previous study (the latter, of course, determining the degree of uncertainty about true treatment effect). Therefore with a single table we can provide quite general results.

Table 1 shows levels of true statistical power under various scenarios. We assume a clinical trial that is to be evaluated by a test of post-treatment difference between a medication and placebo group. We further assume that a conventional power analysis is performed, using as H1 the observed difference between medication and placebo groups in a previous study. Various values of  $N$  are considered, where  $N$  is the sample size of each group in the pilot study. Finally, we assume the conventional power analysis has led to power estimates of .80, .90, and .95.

**Table 1. True Power and Conventional Power Estimates Under Various Scenarios**

N	Classical Power Estimate		
	0.80	0.90	0.95
25	0.65	0.72	0.77
50	0.69	0.77	0.83
75	0.71	0.80	0.86
100	0.73	0.82	0.88
150	0.74	0.84	0.90
250	0.76	0.86	0.92
500	0.78	0.88	0.93
1000	0.79	0.89	0.94

Note: Tabled values are true statistical power for various combinations of (a) power estimated by the conventional method

(.80, .90, .95) and (b) pilot study sample size per group ( $N$ ).

Thus, for example, with  $N=100$ , when the conventional power estimate is .90, the true power is .82.

The discrepancies between true and conventionally-estimated power are substantial. For example, even if the  $N$  of each group in the pilot study is 100, when the conventional method estimates statistical power as .90, the true power is actually .82.

### 4.3 Approximate Method

In the absence of another means to estimate the exact value of  $E[P(\delta)]$  as described above, there is a simple and convenient alternative. One may approximate the integral to any required degree of accuracy using simple spreadsheet calculations based on the output of standard power-estimation software, such as PASS 2000 (Hintze, 2001). In essence, this method approximates the integral of Eq. 4 as a weighted average of power for various values of effect size  $\delta$ , where weights are defined by the densities of the *pdf*,  $p(\delta)$  at representative points.

The calculations are quite simple. An example is shown in Table 2. Here we consider again a test of post-treatment mean differences between a medication group and placebo group. We assume that a pilot study ( $N_1 = N_2 = 100$ ) showed means of 122.9 (medication) and 100 (placebo), where higher values imply better outcomes, and  $\sigma_1 = \sigma_2 = 50$ .

**Table 2. Estimating True (Unconditional) Power Using a Spreadsheet**

A	B	C	D	E	F	G	H	I	J
z	$\phi(z)$	w	$\mu_2 - \mu_1$	N1	N2	$\sigma_1$	$\sigma_2$	Power	$w \times$ Power
-3	0.0044	0.0044	1.7	100	100	50	50	<b>0.0568</b>	0.0002
-2	0.0540	0.0540	8.8	100	100	50	50	<b>0.2372</b>	0.0128
-1	0.2420	0.2421	15.9	100	100	50	50	<b>0.6111</b>	0.1479
0	0.3989	0.3990	22.9	100	100	50	50	<b>0.9001</b>	0.3592
1	0.2420	0.2421	30.0	100	100	50	50	<b>0.9888</b>	0.2393
2	0.0540	0.0540	37.1	100	100	50	50	<b>0.9995</b>	0.0540
3	0.0044	0.0044	44.1	100	100	50	50	<b>1.0000</b>	0.0044
Sums:	0.9997	1.0000							<b>0.8179</b>

We begin by approximating  $p(\delta)$  in the first four columns. From a Bayesian viewpoint, we assume a rectangular prior for treatment effect; from a

frequentist viewpoint, we base  $p(\delta)$  on the sampling distribution of the previous treatment effect. Column A contains a series of z-values over which the distribution is evaluated. Typically these are evenly spaced and centered at 0; more accurate results are obtained when there is odd number (5, 7, 9, 11) of them. Column B contains the ordinate of the standard normal curve associated with these z values, or  $\phi(z)$ . These are rescaled in Column C by dividing each value by their sum, producing weights (w) for the weighted average. Next we calculate the various values of true effect size ( $\delta$ , or, here,  $\mu_2 - \mu_1$ ) corresponding to the selected z values.

Columns E, F, G, and H contain the sample sizes and assumed standard deviations of the two groups for the planned clinical trial. Column I contains the estimated statistical power given the information in columns D through H. This is obtained easily by supplying the values in columns D through H to a computer program like PASS 2000.

Column J contains the product of the cells of Column C and Column I. The sum of these products is a weighted average, here, .8179, which is the estimate of true statistical power. This is the same as the exact estimate to better than three decimal places.

Basically, by means of this simple spreadsheet we have evaluated the integral of Eq. 4 with a primitive 7-point quadrature and obtained very accurate results. Using these same z values as our basis, we would expect similarly accurate results whenever  $p(\delta)$  is gaussian. If  $p(\delta)$  is not gaussian (e.g., beta-shaped), more intervals may be used.

#### **4.4 Effect on Sample Size Estimates**

Using this same basic model one can determine the sample size required to achieve a stipulated level of true statistical power. For example, using the same scenario considered in Table 1 (test of difference between means), we may ask how many subjects are needed in each group to achieve a true power of .90, given that (a) the H1 effect size leads to a classical power estimate of .90, and (b) in the pilot study,  $N_1 = N_2 = N = 50, 100, \text{ or } 200$ . The results are summarized in Figure 4.

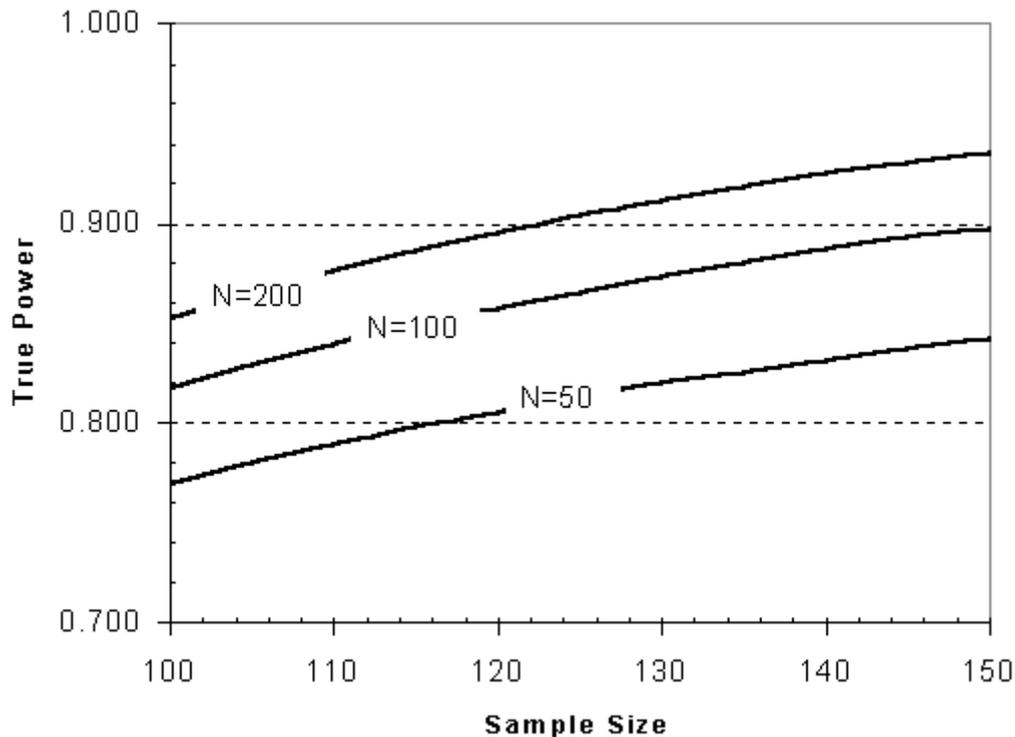


Figure 4. True statistical power (y-axis) as a function of planned sample size of a new study (x-axis), given three N's (200, 100, and 50) of the pilot study and a classical power estimate of .90.

When  $N = 200$  for the pilot study, 123 subjects are required in each group of the new study to reach a true power of .90. Thus, 23% more subjects are needed to achieve a power of .90 than the conventional method suggests. When  $N = 100$  for the pilot study, the new study will need 153 subjects in each group to have a true power of .90. When  $N = 50$  for the pilot study, 246 subjects per group are required in the new study to reach .90 power.

This shows that the discrepancy between required sample sizes estimated by the conventional and by the hybrid Bayes method vary from modest to pronounced, depending on the degree of uncertainty associated with the estimate of effect size in the earlier study. It is rather surprising that this has not attracted more interest among applied researchers to date.

## 5. Discussion

### 5.1 Other Sources of Uncertainty and Information

Here we considered H1 based on a single previous study and uncertainty about true effect size mainly as a function of the sample size of that study. However, the same principles apply more generally. The degree to which the

conventional method will underestimate power is a function of uncertainty about true effect size, regardless of the source of uncertainty.

If multiple previous studies have been performed, then, assuming equivalent populations, a pooled effect size estimate can be constructed using the total N across all previous studies. If the total N is sufficiently large (e.g., 1000 or more), then true power will approach classical power estimates.

However it is also possible that populations, protocols, or medications of the previous studies are not consistent. In this case one may need to consider that true effect size is itself a random variable or *random effect*. This introduces a new source of uncertainty about true effect size as inferred from previous data -- so that conventional power analysis will more severely underestimate true power. In this case, one may estimate  $p(\delta)$  using statistical models that consider both a random effect across studies and sampling error of each study, and then proceed as described here to estimate true power. Methods developed in the area of Bayesian meta-analysis (Sutton & Higgins, in press) may be adapted to this purpose.

The issue of a random treatment effect is especially problematic if there has been only a single previous study, as then there is no empirical basis for estimating the variance or distribution of the random effect. Here a Bayesian framework, which permits consideration of non-empirical (e.g., based on scientific models, expert opinion, etc.) information in estimating  $p(\delta)$  seems almost unavoidable.

## 6. Conclusions

We have shown how conventional methods underestimate the true statistical power of a clinical trial when H1 is formed without taking due account of uncertainty about effect size; and have also shown that the amount of bias depends on the degree of uncertainty. The results here are not universally applicable. For example, sometimes power calculations are performed on some other basis than a previously observed effect size, and sometimes true effect size might be known fairly precisely. However it is also clear that many clinical trials are powered based on very uncertain estimates of effect size, such that the results here apply. For these cases, unless one is prepared to consider across-the-board sample size increases of 25%, 50%, or even 100%, some alternatives are needed.

One possibility is to move to fully Bayesian data analysis in clinical trials. This may involve an ongoing updating of the posterior estimate of treatment effect as each subject completes the protocol; it is generally believed that this may increase statistical power somewhat, but more work is required to quantify this.

A fully Bayesian approach may have a greater positive effect on statistical power by enabling the fuller consideration of prior information on treatment efficacy. Relevant prior information may come from several sources, including meta-analysis, biological or pharmacological models, and expert opinion. A Bayesian framework enables one to formally estimate the value of prior information. A small investment in obtaining a precise prior estimate may easily produce considerable savings in terms of a smaller required study sample size. Regulatory agencies show signs of increased receptivity to a fully Bayesian approach (CDRH, 2006).

One very promising solution is to apply a Bayesian statistical approach within a broader decision-theoretic framework (e.g., Patel & Ankolekar, in press.). This weighs the incremental value of recruiting additional subjects (to increase statistical power) against (1) the incremental costs of a larger clinical trial, (2) revenue loss associated with a delay in bringing a new drug to market, and (3) consideration that there are often other candidate drugs in which a company may invest time and resources.

## 7. References

Berger JO, Wolpert RL. *The Likelihood Principle*, 2nd ed. Hayward, CA: IMS, 1988.

Berry DA, Stangl DK (eds.) *Bayesian Biostatistics*. New York: Marcel Dekker, 1996.

CDRH. Draft Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials. FDA. May 23, 2006 .

<http://www.fda.gov/cdrh/osb/guidance/1601.html>

Hintze, J. PASS 2000 User Guide. NCSS. Kaysville, Utah., 2001.

Joseph L, Wolfson DB, du Berger, R. Some comments on Bayesian sample size determination. *The Statistician: Journal of the Institute of Statisticians*, 44, 167-171, 1995.

Patel NR, Ankolekar S. A Bayesian approach for incorporating economic factors in sample size design for clinical trials of individual drugs and portfolios of drugs. *Statistics in Medicine* (in press).

Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. New York: Wiley, 2004.

Sutton SJ, Higgins JPT. Recent developments in meta-analysis. *Statistics in Medicine* (in press).