Journal of Educational and Behavioral Statistics
Summer 2005, Vol. 30, No.2, pp. 141–167

On the Post Hoc Power in Testing Mean Differences

Ke-Hai Yuan
Scott Maxwell
University of Notre Dame

Retrospective or post hoc power analysis is recommended by reviewers and editors of many journals. Little literature has been found that gave a serious study of the post hoc power. When the sample size is large, the observed effect size is a good estimator of the true effect size. One would hope that the post hoc power is also a good estimator of the true power. This article studies whether such a power estimator provides valuable information about the true power.

Using analytical, numerical, and Monte Carlo approaches, our results show that the estimated power does not provide useful information when the true power is small. It is almost always a biased estimator of the true power. The bias can be negative or positive. Large sample size alone does not guarantee the post hoc power to be a good estimator of the true power. Actually, when the population variance is known, the cumulative distribution function of the post hoc power is solely a function of the population power. This distribution is uniform when the true power equals 0.5 and highly skewed when the true power is near 0 or 1. When the population variance is unknown, the post hoc power behaves essentially the same as when the variance is known.

Keywords: *bias, effect size, observed power*

1. Introduction

Statistical power continues to receive increasing attention. For example, the fifth edition of the APA Publication Manual advises researchers to "Take seriously the statistical power considerations associated with your tests of hypotheses" (2001, p. 24). However, the manual does not distinguish between planned power analysis and post hoc power analysis. Cohen (1988) provided many power tables and suggested small, medium, and large effect sizes. When using these tables, it is ideal to know the true effect size underlying the experiment. In practice, the exact true effect size is generally unknown even after the experiment. But one can estimate the effect size using the standardized mean difference. When the sample size is large, the estimated effect size is near the true effect size.

This research was supported by NSF Grant DMS04-37167. We are thankful to the editor, an associate editor, and a referee for their constructive comments that have improved the article.

One can go one step further using the estimated effect size to construct a power estimate for the current study. This estimator is called the post hoc power (Gillett, 1994a) or the observed power (Hoenig & Heisey, 2001). The question is how much information the observed power provides us about the power underlying the current study.

Hoenig and Heisey (2001) cite 19 journals across a variety of disciplines where recommendations have been published advocating the use of post hoc power to interpret the results of studies with statistically nonsignificant results. The need for further attention to this topic in psychology is exemplified by Finch, Cumming, and Thomason (2001), who found that 37% of studies with statistically nonsignificant results published in *Journal of Applied Psychology* in 1999 interpreted their results as providing evidence that the null hypothesis is true. However, only a total of three articles actually provided power calculations (two a priori and one post hoc). Some journals and textbook writers have addressed this confusion. For example, the instructions to authors for *Animal Behavior* state that "Where a significance test based on a small sample size yields a nonsignificant outcome, the power of the test should normally be quoted in addition to the P value" (2001, vi). Similarly, Pallant (2001) states "Some of the SPSS programs also provide an indication of the power of the test that was conducted, taking into account effect size and sample size. If you obtain a non-significant result and are using quite a small sample size, you need to check these power values" (p. 173). Stevens (1999) provides another example in his statement that "If a post hoc power analysis is done on a study where significance is not found and the effect size is quite small (< .10), then one must decide whether such an effect has any practical significance. On the other hand, when significance is not found and a post hoc power analysis reveals a large or medium effect size, then it is essential to replicate the study with more adequate sample size" (p. 133). Thus, these three references serve as examples of recommendations to perform a power analysis after having obtained statistically nonsignificant results. However, these sources are often unclear about how to perform a power analysis, and in particular how a post hoc power analysis relates to an a priori power analysis. See Hoenig and Heisey (2001) for a recent review in this direction.

Most sources describe post hoc power analyses in either of two ways. First, one commonly recommended method involves finding the power for: (a) a fixed alpha level (typically .05), (b) the sample size used in the study, and (c) a "meaningful" effect size, often expressed in terms of Cohen's conventions for small, medium, and large effect sizes. Ironically, these sources typically do not point out that this so-called post hoc power analysis is, in fact, identical to a prospective power analysis. There is absolutely nothing about the analysis that depends on the data obtained in the actual study. Second, the other frequently mentioned approach involves finding the power for (a) a fixed alpha level (typically .05), (b) the sample size used in the study, and (c) the effect size observed in the study. This is what we refer to as "observed power" or "post hoc power" in this article. There is also a third possibility, which is mentioned much less often in the literature. This approach involves finding the power for (a) a fixed alpha level (typically .05), (b) the sample size used in the study, and (c) a "meaningful" unstandardized effect size, such as a mean difference or an unstandardized regression coefficient, and (d) an estimate of error variance obtained from the data of the current study. This approach is sometimes used in areas of medical research where a meaningful unstandardized effect size can be specified but the magnitude of error variance is unknown before collecting data. However, this approach receives little attention in educational and behavioral research because of the difficulty in specifying meaningful values of unstandardized effect sizes, especially when the measurement scale precludes knowledge of the accompanying error variance. Thus, among the two popular methods, the only approach that is truly "post hoc" is the method we consider in this paper.

There have been several criticisms of the observed power. Assuming the true effect size follows a known prior distribution, Gillett (1994a, 1994b) concluded that the observed power generally underestimates the true power. Of course, different experiments may have different effect sizes even in the same area of study. In our opinion, it is more reasonable to regard the effect size behind the given study as fixed rather than random. With a fixed effect size, the relationship between the observed power and the true population power is not well understood. Hoenig and Heisey (2001) argued that the observed power is a function of the p value, so once the p value is known one should not recalculate the observed power. However, the fact that observed power is a function of the p value does not necessarily imply that they are literally redundant of one another. For example, when sampling from a normal distribution with a known variance, the sample mean is a sufficient statistic for the population mean. Every sensible statistic must be a function of the sample mean. This does not imply that one should not use different estimators for different information. Actually, the sample mean and the z score offer different information about the population quantities in this case. Similarly, the observed power and the p value may both be useful. For example, Greenwald, Gonzalez, Harris, and Guthrie (1996) suggest that a monotonic transformation of the p value may be of interest, even though it is a function of the p value because it estimates the probability that an exact replication will yield a statistically significant result. In fact, Greenwald et al. use this property as partial justification for the continued relevance and importance of reporting p values in behavioral research. However, this probability of statistical significance upon replication is precisely what observed power is intended to estimate. In fact, Posavac (2002) suggests supplementing such indices as confidence intervals with the probability of a statistically significant exact replication (see Macdonald, 2003; Posavac, 2003, for further discussion of this suggestion). Thus, examining the accuracy of observed power is important not only because some journals and authors have recommended that observed power be calculated when results are not statistically significant, but also because other authors have argued that the probability of a statistically significant exact replication is of fundamental interest unto itself whether or not the results obtained in a specific study are themselves statistically significant. We would maintain that the concept of observed power is not only meaningful from a theoretical perspective but may also be of considerable practical interest. Even so, we agree with Hoenig and Heisey (2001) that many researchers have misunderstood the implications of calculating observed power. In particular, a low value of observed power does not necessarily suggest that the original study was underpowered. Instead, it may simply suggest that the underlying effect size is trivially small. We agree with Hoenig and Heisey that confidence intervals and equivalence tests are generally superior to observed power for interpreting the magnitude of effect sizes. Nevertheless, observed power could still be of interest in the sense that it has the potential to reveal the extent to which a replication study is likely to yield a statistically significant result.

The purpose here is to understand what kind of information the observed power will provide. If any, how useful is it? Can one get a better estimator of the power using other statistics from the current study? In section 2 we will study the property of the observed power in detail. This includes the bias, variance, and distributional shape of the observed power as well as the confidence interval for the true population power. We first consider testing the population mean based on one sample whose population variance is known. The main reason for considering such an oversimplified situation is because analytical results can be derived. When population variances are known and equal, the problem of comparing population means of two samples is just a special case of the one sample problem. We will also study the observed power with unknown population variances. Because there is no analytical solution when the variances are unknown, evaluations will be based on numerical integration and simulation. As we shall see, the behavior of the observed power with an unknown variance is very similar to that with a known variance. We present our results in section 2. Examples based on real as well as simulated data are given in section 3. Conclusions are given in section 4. The technical details are provided in the appendix.

2. What the Observed Power Estimates

2.1. One Group With a Known Variance

Consider a sample $x_1, \ldots, x_n$ from a normal distribution $N(\mu_0, \sigma_0^2)$ with an unknown μ_0 but a known σ_0^2. It is well known that the sample mean $\bar{x}$ follows the normal distribution $N(\mu_0, \sigma_0^2/n)$. Consequently, $z = \sqrt{n}\bar{x}/\sigma_0$ follows $N(\sqrt{n}\delta_0, 1)$, where $\delta_0 = \mu_0/\sigma_0$. For the purpose of clarity we mainly consider one-sided test for

$$H_0 : \mu_0 = 0 \text{ vs. } H_1 : \mu_0 > 0. \tag{1}$$

A two-sided test is discussed briefly at the end of this section. When referring the test statistic z to the standard normal distribution, the power or the probability of rejecting the H_0 is

$$\gamma_z(\delta_0) = 1 - \Phi(z_{(1-\alpha)} - \sqrt{n}\delta_0),$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution, and $z_{(1-\alpha)} = \Phi^{-1}(1 - \alpha)$ is the critical value corresponding to probability $1 - \alpha$. When μ_0 is known, it is straightforward to evaluate the true power $\gamma_0 = \gamma_z(\delta_0)$. Unfortunately, μ_0 is always unknown in practice. It is known that $\bar{x}$ is an unbiased estimator of μ_0, and it is a good estimator when sample size n is relatively large. Let $\hat{\delta} = \bar{x}/\sigma_0$. The estimator $\hat{\gamma}_z = \gamma_z(\hat{\delta})$ is then precisely the observed power as defined in the literature. The question is whether $\gamma_z(\hat{\delta})$ provides a good estimate of γ_0 when n is large. From the appendix we have the result

$$E[\gamma_z(\hat{\delta})] = 1 - \Phi[(z_{(1-\alpha)} - \sqrt{n}\delta_0)/\sqrt{2}]. \tag{2}$$

Notice that $E(\hat{\gamma}_z) = \gamma_0 = 0.5$ when $z_{(1-\alpha)} - \sqrt{n}\delta_0 = 0$. It follows from Equation 2 that $\gamma_z(\hat{\delta})$ is: (a) unbiased when true power $\gamma_0 = 0.5$, (b) positively biased when $\gamma_0 < 0.5$, (c) negatively biased when $\gamma_0 > 0.5$. For a given α, the

$$\text{Bias} = \Phi(z_{(1-\alpha)} - \sqrt{n}\delta_0) - \Phi[(z_{(1-\alpha)} - \sqrt{n}\delta_0)/\sqrt{2}]$$

depends on the population effect size and the sample size. When $\alpha = 0.05$, and $\delta_0 = 0.2, 0.5, 0.8$, which are the small, medium, and large effect sizes suggested by Cohen (1988), the respective maximum positive biases are about 0.083, 0.083, 0.076, which happen at sample sizes $n = 5, 1, 1$, respectively. The respective maximum negative biases are about −0.083, −0.083, −0.083, which happens at $n = 199, 32, 12$.

Set $n_0 = (z_{(1-\alpha)}/\delta_0)^2$, which is the sample size that leads to approximately a zero bias. Equation 2 also tells us:

I. The observed power $\gamma_z(\hat{\delta})$ is almost always a biased estimator of the true power γ_0. When $n > (z_{(1-\alpha)}/\delta_0)^2$, the bias is negative. When $n < (z_{(1-\alpha)}/\delta_0)^2$, the bias is positive. Only when $n = n_0$ is the observed power unbiased.

II. The bias with a larger sample size may not necessarily be smaller than that with a smaller sample size.

III. For a fixed $\delta_0 > 0$, the bias becomes trivial when n becomes much greater or much less than $(z_{(1-\alpha)}/\delta_0)^2$.

Observation III is due to the fact that $\Phi(\cdot)$ is insensitive at the tails. For example, let $\alpha = 0.05$ and $\delta_0 = 0.5$, the bias is about −0.083 when $n = 30$, while it is about −0.0084 when $n = 100$.

For Cohen's small, medium, and large effect sizes, we have $n_0 = 68$, 11, and 5, respectively. If the commonly encountered sample sizes range from 15 to 50, then $\gamma_z(\hat{\delta})$ will always contain a positive bias for the small effect size and negative biases for the medium and large effect sizes. Of course, with a similar amount of bias, the small effect size will suffer more than the medium or large effect sizes. For the three effect sizes $\delta_0 = 0.2, 0.5, 0.8$, Figure 1 contains the plots of the true power γ_0, the relative bias

$$\text{Bias}_R = \{\Phi(z_{(1-\alpha)} - \sqrt{n}\delta_0) - \Phi[(z_{(1-\alpha)} - \sqrt{n}\delta_0)/\sqrt{2}]\}/[1 - \Phi(z_{(1-\alpha)} - \sqrt{n}\delta_0)]$$

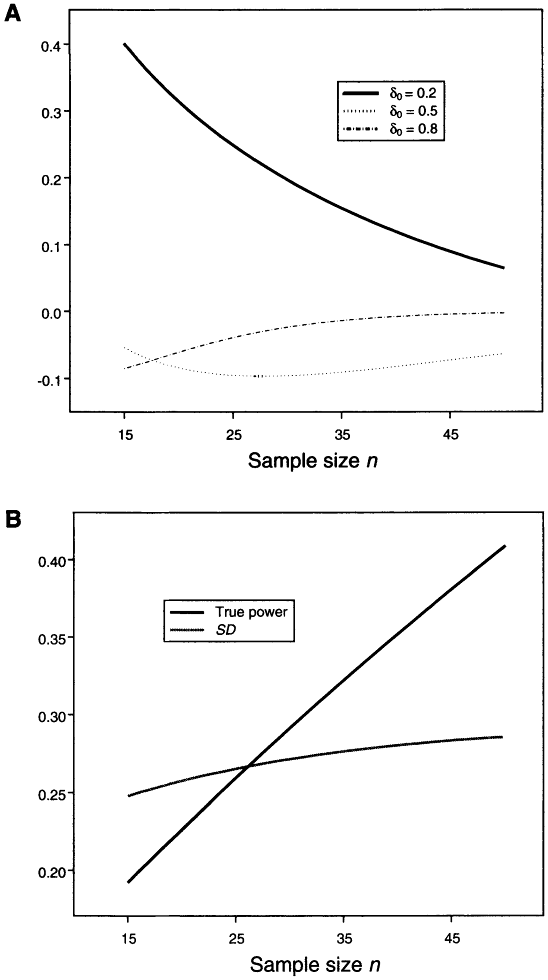


FIGURE 1. *True power, relative bias, and standard deviation of the observed power* $\gamma_z(\hat{\delta})$ *(effect size* $\delta_0 = 0.2, 0.5, 0.8$*; sample size* n = *15–50).* **A**, *relative bias;* **B**, *effect size* $\delta_0 = 0.2$. (*continued*)

and the standard deviation (*SD*) of $\gamma_z(\hat{\delta})$ when *n* changes from 15 to 50. When the true effect size is small, the bias can be 40% of γ_0, which makes the observed power a very poor estimate of the true power, especially for smaller sample sizes. For example, when using the observed effect size to estimate true power, the actual power will be less than suggested by observed power, leading to an underpowered study. When the effect size is medium to large, the relative bias is much smaller. Actually, because the power is under estimated, the predicted sample size leads to

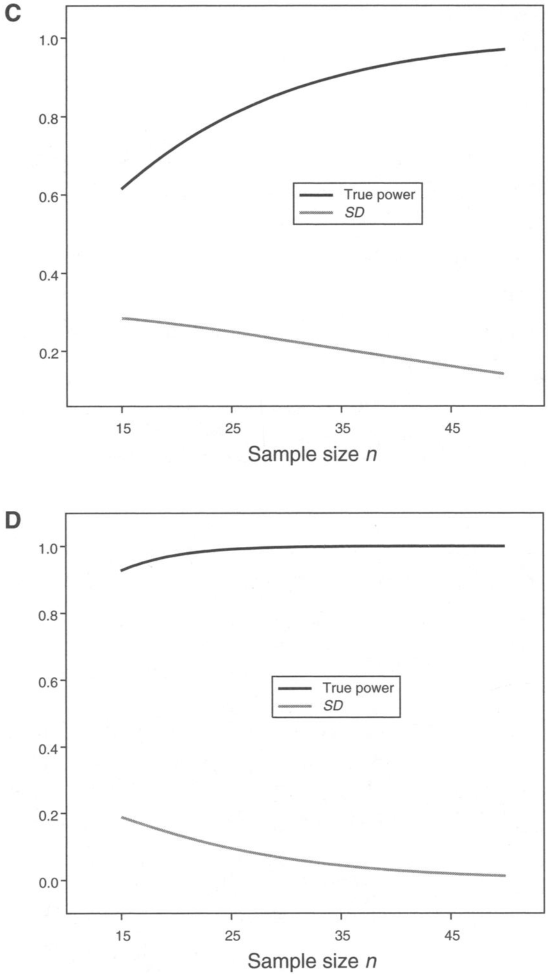


FIGURE 1. (*continued*). **C**, *effect size* $\delta_0 = 0.5$; **D**, *effect size* $\delta_0 = 0.8$.

higher than expected power on average. When large samples do not pose a serious problem or a great cost, negative bias may not pose a big problem.

With the small effect size, the *SD* of $\hat{\gamma}_z$ can be as big as the power γ_0 itself, which makes $\hat{\gamma}_z$ of little value. Similar to the bias, as sample size gets larger, the *SD* of $\hat{\gamma}_z$ does not necessarily get smaller. Actually, the *SD* for the small effect size monotonically increases as γ_0 or n increases. We will explain this further using the CDF of $\gamma_z(\hat{\delta})$ in the later part of this section.

Whether the observed power overestimates the true power depends on whether $z_{(1-\alpha)} - \sqrt{n}\delta_0 > 0$. One cannot use the observed counterpart $z_{(1-\alpha)} - \sqrt{n}\hat{\delta}$

to decide whether it is an over- or underestimation. For example, when $z_{(1-\alpha)} - \sqrt{n}\delta_0 = a > 0$, the observed power will overestimate the true power on average. However, the probability of observing $z_{(1-\alpha)} - \sqrt{n}\hat{\delta} < 0$ is $1 - \Phi(a)$,which can be near 0.5 unless a is fairly large.

Bias is not desired in estimating any parameter. We have the unbiased estimator $\hat{\delta}$ for δ_0. It is interesting to see whether an unbiased estimator of γ_0 can be constructed, based on the unbiased effect size $\hat{\delta}$. Because of the form of $\gamma_z(\delta_0)$, it seems unlikely to have a transformation $g(\cdot)$ such that $g[\gamma_z(\hat{\delta})]$ is unbiased. Notice that $\gamma_z(\hat{\delta})$ is actually a function of $z_{(1-\alpha)} - \sqrt{n}\hat{\delta}$. We would like to explore the possibility of using a linear transformation on $z_{(1-\alpha)} - \sqrt{n}\hat{\delta}$. Specifically, we will seek an unbiased estimator among the form $\gamma_{ab}(\hat{\delta}) = 1 - \Phi\left[a\left(z_{(1-\alpha)} - \sqrt{n}\hat{\delta}\right) + b\right]$. It turns out that

$$E\left[\gamma_{ab}(\hat{\delta})\right] = 1 - \Phi\left[\frac{a\left(z_{(1-\alpha)} - \sqrt{n}\delta_0\right) + b}{(1 + a^2)^{1/2}}\right].$$

Let $E[\gamma_{ab}(\hat{\delta})] = \gamma_z(\delta_0)$ we obtain $\left(1 - a/\sqrt{1 + a^2}\right)\left(z_{(1-\alpha)} - \sqrt{n}\delta_0\right) = b$. Because δ_0 is unknown, the solution to this equation is $a = \infty$ and $b = 0$. For a finite a, the bias in $\gamma_a(\hat{\delta}) = \gamma_{a0}(\hat{\delta})$ is

$$\text{Bias}_a = \Phi\left(z_{(1-\alpha)} - \sqrt{n}\delta_0\right) - \Phi\left[a\left(z_{(1-\alpha)} - \sqrt{n}\delta_0\right)/\sqrt{1 + a^2}\right].$$

When a is large enough, the bias in $\gamma_a(\hat{\delta})$ is much smaller than that in $\gamma_z(\hat{\delta})$. For example, using $a = 10$, for sample sizes $n = 1$ to 400, the maximum bias among the three different effect sizes is only about 0.0012.

It is nice to have a less biased estimator for γ_0. It is also important for the estimator to have a small *SD*. It turns out that the *SD* of $\gamma_a(\hat{\delta})$ depends on the a used in its definition. When a increases, the *SD* of $\gamma_a(\hat{\delta})$ also tends to increase for a wide range of δ_0 and sample sizes. At the medium effect size $\delta_0 = 0.5$, Figure 2 compares the means and *SD*s of $\gamma_z(\hat{\delta})$ and $\gamma_a(\hat{\delta})$ for $n = 5$ to 100, where $a = 20$ is used. The true power γ_0 is also included for reference, but it overlaps with the mean of $\gamma_a(\hat{\delta})$ in the plot. We can see that the *SD*s of both the power estimators are relatively small when γ_0 is large. However, the *SD* of $\gamma_a(\hat{\delta})$ is much greater than that of $\gamma_z(\hat{\delta})$ when γ_0 is small to medium. Especially, for a smaller γ_0, the *SD* of $\gamma_a(\hat{\delta})$ is more than the size of the power itself, which may render $\gamma_a(\hat{\delta})$ of little value. Although the estimator $\gamma_z(\hat{\delta})$ is biased, when power is small to medium, it might be more near γ_0 on average than the less biased estimator $\gamma_a(\hat{\delta})$. Less biased estimators for γ_0 may also be obtained by computer intensive methods such as bootstrap and jackknife. The bias-corrected estimators by these methods may suffer the same problem of increased variances. See Efron and Tibshirani (1993, p. 138) for further discussion on bias correction.

In addition to the bias and *SD*, it is of interest to know the distribution of $\gamma_z(\hat{\delta})$. The appendix provides the CDF of $\gamma_z(\hat{\delta})$ as

$$F_{\gamma_z}(t) = \Phi\left[\left(z_{(1-\alpha)} - \sqrt{n}\delta_0\right) + z_t\right],$$

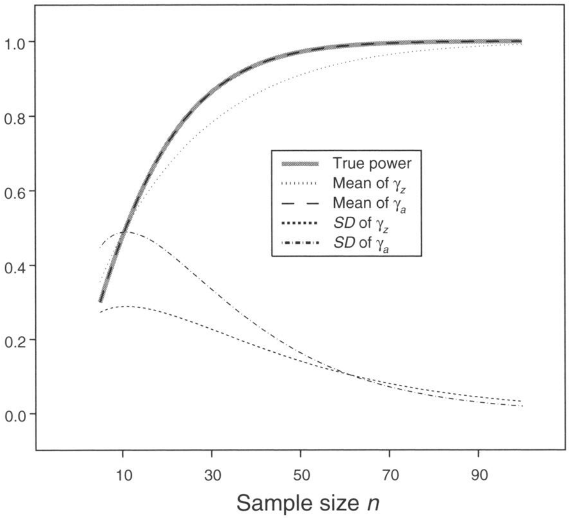


FIGURE 2. *Means and standard deviations of power estimators $\gamma_z(\hat{\delta})$ and $\gamma_a(\hat{\delta})$ at the medium effect size $\delta_0 = 0.5$. (The mean of $\gamma_a(\hat{\delta})$ overlaps with the true power δ_0 in the solid line.)*

where $z_t = \Phi^{-1}(t)$ is the percentile of the standard normal distribution. It is easy to see that, when $z_{(1-\alpha)} - \sqrt{n}\delta_0 = 0$ or $\gamma_0 = 0.5$, $F_{\gamma_z}(t) = t$ so that $\gamma_z(\hat{\delta})$ has a uniform distribution on [0, 1]. Because $z_{(1-\alpha)} - \sqrt{n}\delta_0 = \Phi^{-1}(1 - \gamma_0)$, we can write

$$F_{\gamma_z}(t) = \Phi(z_{(1-\gamma_0)} + z_t). \tag{3}$$

So, in general, the CDF $F_{\gamma_z}(t)$ depends on the γ_0 as well. We may regard Equation 3 as a class of distributions indexed by γ_0. The density of $F_{\gamma_z}(t)$ is given by

$$f_{\gamma_z}(t) = \exp\left\{-\frac{1}{2}z_{(1-\gamma_0)}(z_{(1-\gamma_0)} + 2z_t)\right\}.$$

To get a better picture of the distribution shape of $\gamma_z(\hat{\delta})$, the plots of $f_{\gamma_z}(t)$ against t for $\gamma_0 = 0.1, 0.2, 0.3, 0.4$ are provided in Figure 3. Because $z_{(1-\gamma_0)} = -z_{\gamma_0}$ and $z_t = -z_{(1-t)}$, the plots corresponding to $\gamma_0 = 0.6, 0.7, 0.8$, and 0.9 are just flips of those corresponding to $\gamma_0 = 0.4, 0.3, 0.2$, and 0.1, respectively. It is clear that the distribution of $\gamma_z(\hat{\delta})$ is highly skewed unless $\gamma_0 \approx 0.5$. For $0 < \gamma_0 < 0.5$, the peak of the density function is at the lower end $t = 0$. As can be judged from $f_{\gamma_z}(t)$, this peak is $+\infty$. All the plots in Figure 3 start at $t = 1/1001$. It follows from Equation 3 that γ_0 is actually the median of $F_{\gamma_z}(t)$,that is $P(\hat{\gamma}_z > \gamma_0) = P(\hat{\gamma}_z < \gamma_0) = 0.5$. This implies that, when the median of $\gamma_z(\hat{\delta})$ is available, in the context of meta analysis for example, then it is an unbiased estimate of γ_0. Although $\gamma_z(\hat{\delta})$ tends to overestimate γ_0 when $\gamma_0 < 0.5$, there is also a 50% chance for $\gamma_z(\hat{\delta})$ to be below

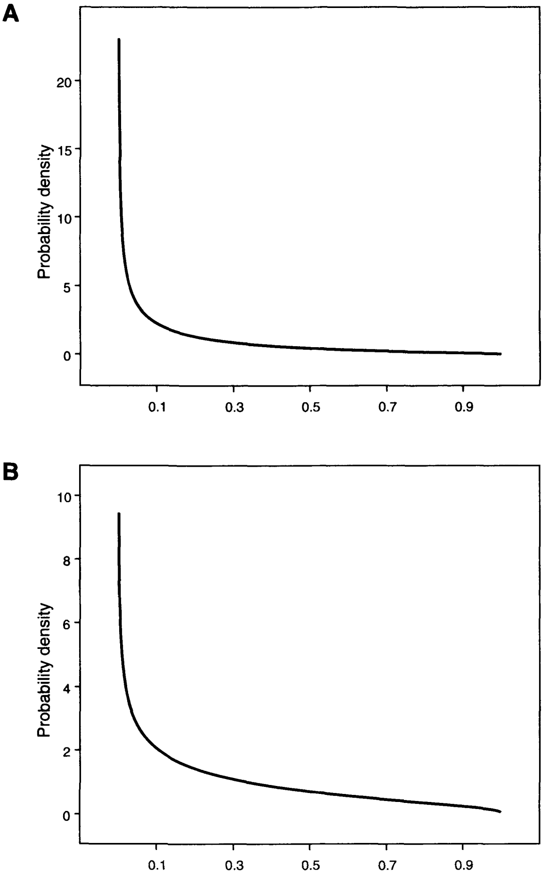


FIGURE 3. *Probability density of* $\gamma_z(\hat{\delta})$ *when true power* $\gamma_0 = 0.1, 0.2, 0.3, 0.4$. **A**, *True power* $\gamma_0 = 0.1$; **B**, *true power* $\gamma_0 = 0.2$. (*continued*)

γ_0. Similarly, when $0.5 < \gamma_0 < 1.0$, $\gamma_z(\hat{\delta})$ tends to underestimate γ_0 on average but there is 50% chance for $\gamma_z(\hat{\delta})$ to be above γ_0 as well. The bias is due to the fact that the magnitude by which $\hat{\gamma}_z > \gamma_0$ is different from the magnitude by which $\hat{\gamma}_z < \gamma_0$ unless $\gamma_0 = 0.5$.

The density function together with Figure 3 also show that the mean and variance of $\gamma_z(\hat{\delta})$ is only a function of γ_0. Once γ_0 is given, all the population characteristics of $\gamma_z(\hat{\delta})$ do not depend on n or δ_0. Because of the shape of the distribution, $\mathrm{Var}[\gamma_z(\hat{\delta})\,|\,\gamma_0] = \mathrm{Var}[\gamma_z(\hat{\delta})\,|\,(1 - \gamma_0)]$ is the smallest when γ_0 is near 0 or 1. It has the maximum variance when $\gamma_0 = 0.5$, which is $\sqrt{1/12} \approx 0.289$.

As for the estimator $\gamma_a(\hat{\delta})$, its CDF is given by

$$F_{\gamma_a}(t) = \Phi\left(z_{(1-\gamma_0)} + z_t/a\right). \qquad (4)$$

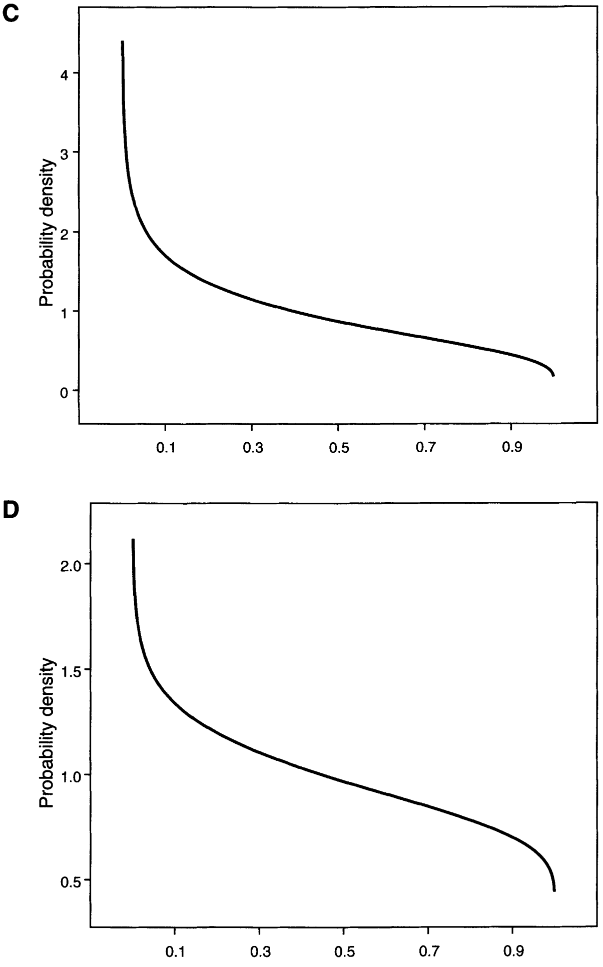


FIGURE 3. (*continued*). **C**, *true power* $\gamma_0 = 0.3$; **D**, *true power* $\gamma_0 = 0.4$.

From the form of $F_{\gamma_a}(t)$ it is easy to see that, for a large a, most of the probability is near 0 and 1, there is little probability in between. This explains why $\gamma_a(\hat{\delta})$ has a greater *SD* than $\gamma_z(\hat{\delta})$.

Using the CDFs in Equations 3 or 4, one can construct a confidence interval for γ_0. Note that, for a general statistic T, its CDF is defined by

$$F(t\,|\gamma) = P(T \leq t\,|\gamma),$$

where γ is the unknown parameter that determines the distribution. When T is a reasonable estimator of γ, the greater the γ the greater the T tends to be. For a given *t*, a greater γ corresponds to a smaller probability $F(t \mid \gamma)$ and vice versa. Let $T = t_0$ be the observed value of T. When $\gamma = \gamma_0$, it is unlikely to have $F(t_0 \mid \gamma_0) > 0.95$ or $F(t_0 \mid \gamma_0) < 0.05$. When $F(t_0 \mid \gamma_0) > 0.95$, we may doubt that γ is smaller than γ_0. Similarly, when $F(t_0 \mid \gamma_0) < 0.05$, we will doubt that γ is greater than γ_0. Consequently, the 90% confidence limits in $[\gamma_L, \gamma_U]$ are numbers satisfying

$$F(t_0 \mid \gamma_L) = 0.95 \quad \text{and} \quad F(t_0 \mid \gamma_U) = 0.05. \tag{5}$$

Equation 5 has been used by several authors to obtain a confidence interval (e.g., Browne & Cudeck, 1993; MacCallum, Browne, & Sugawara, 1996; Steiger & Fouladi, 1997), which is valid as long as $F(t \mid \gamma)$ is monotonically decreasing in γ. Applying Equation 5 with $F = F_{\gamma_z}$, the appendix provides details leading to

$$\gamma_L = 1 - \Phi\left(z_{1-\alpha} + z_{0.95} - \sqrt{n}\hat{\delta}\right) \quad \text{and} \quad \gamma_U = 1 - \Phi\left(z_{1-\alpha} - z_{0.95} - \sqrt{n}\hat{\delta}\right). \tag{6}$$

It is easy to see that the γ_L and γ_U are just the result when applying the power function $\gamma_z(\delta)$ to the confidence limits of $\delta_0 \in \left[\hat{\delta} - z_{0.95}/\sqrt{n}, \hat{\delta} + z_{0.95}/\sqrt{n}\right]$. Such a procedure has been used by Dudewicz (1972) and Taylor and Muller (1995). When solving Equation 5 with $F = F_{\gamma_a}$, we get the same confidence limits as given in Equation 6, which are different from those of applying $\gamma_a(\delta)$ to the confidence limits of δ_0. Actually, the confidence interval of applying $\gamma_a(\delta)$ to the confidence limits of δ_0 does not provide the correct coverage. So the procedure to confidence intervals by solving Equation 5 is more general.

Because the alternative hypothesis is $H_1 : \mu_0 > 0$, one should change $\hat{\delta} - z_{0.95}/\sqrt{n}$ to zero when it is less than zero. Similarly, one should change γ_L to α when it is less than α. The latter is based on the unbiasedness of a reasonable test statistic, that is, under H_1 one should be able to reject H_0 at least with probability α. The expectations of the confidence limits are

$$E(\gamma_L) = 1 - \Phi\left(z_{1-\alpha}/\sqrt{2} + z_{0.95}/\sqrt{2} - \sqrt{n/2}\delta_0\right)$$

and

$$E(\gamma_U) = 1 - \Phi\left(z_{1-\alpha}/\sqrt{2} - z_{0.95}/\sqrt{2} - \sqrt{n/2}\delta_0\right).$$

For the low limits to be useful on average, we need to have $E(\gamma_L) > \alpha$. This implies

$$n > 2\left[z_{1-\alpha}\left(1/\sqrt{2} - 1\right) + z_{0.95}/\sqrt{2}\right]^2/\delta_0^2 .$$

When $\alpha = 0.05$ and $\delta_0 = 0.2$, one needs to have at least $n \geq 24$ for the confidence interval to offer even minimally useful information on average.

The expected length of the interval given by Equation 6 is

$$E(\gamma_U - \gamma_L) = \Phi\left(z_{1-\alpha}/\sqrt{2} + z_{0.95}/\sqrt{2} - \sqrt{n/2}\delta_0\right) - \Phi\left(z_{1-\alpha}/\sqrt{2} - z_{0.95}/\sqrt{2} - \sqrt{n/2}\delta_0\right).$$

It is easy to see that $E(\gamma_U - \gamma_L)$ will approach zero when n gets larger. It is also interesting to note that, because of the bias in $\gamma_z(\hat{\delta})$, the expectation of the upper limit will be smaller than the true power when

$$n > n_1 = [(\sqrt{2} - 1)z_{(1-\alpha)} + z_{0.95}]^2/[\delta_0(\sqrt{2} - 1)]^2.$$

Because at the sample size n_1, the power is essentially 1.0, this will not cause a problem.

2.2. Two Groups With Equal and Known Variances

Consider the two sample problem $x_1, \ldots, x_{n_1} \sim N(\mu_{10}, \sigma_0^2)$ and $y_1, \ldots, y_{n_2} \sim N(\mu_{20}, \sigma_0^2)$, where σ_0^2 is known. The interesting hypothesis here is

$$H_0 : \mu_{20} = \mu_{10} \text{ vs. } H_1 : \mu_{20} > \mu_{10}. \qquad (7)$$

It is well-known that the test statistic for this hypothesis is

$$z = (\bar{y} - \bar{x})\Bigg/\left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{1/2}\sigma_0\right].$$

Let $\delta_0 = (\mu_{20} - \mu_{10})/\sigma_0$. For a given level α, the power for z to reject the H_0 in Equation 7 is

$$\gamma_z(\delta_0) = 1 - \Phi(z_{1-\alpha} - \delta_0[n_1 n_2/(n_1 + n_2)]^{1/2}).$$

Using $\hat{\delta} = (\bar{y} - \bar{x})/\sigma_0$, the observed power is given by

$$\gamma_z(\hat{\delta}) = 1 - \Phi\left(z_{1-\alpha} - \hat{\delta}[n_1 n_2/(n_1 + n_2)]^{1/2}\right). \qquad (8)$$

Similarly, $\gamma_z(\hat{\delta})$ is generally a biased estimator of γ_0 with

$$E[\gamma_z(\hat{\delta})] = 1 - \Phi\{(z_{1-\alpha} - \delta_0[n_1 n_2/(n_1 + n_2)]^{1/2})/\sqrt{2}\}.$$

Compared to the one sample problem, the $\tilde{n} = n_1 n_2/(n_1 + n_2)$ plays the role of sample size. All conclusions in section 2.1 hold for the $\gamma_z(\hat{\delta})$ in Equation 8 when n is replaced by $\tilde{n}$. For example, with the small effect size $\delta_0 = 0.2$, $\gamma_z(\hat{\delta})$ will underestimate γ_0 when $n_1 = n_2$ are below 135.

2.3. One Group With an Unknown Variance

Consider the random sample $x_1, \ldots, x_n$ from $N(\mu_0, \sigma_0^2)$, where both μ_0 and σ_0^2 are unknown. The statistic for testing the hypothesis in Equation 1 is now

$$t = \sqrt{n}\,\bar{x}/s,$$

which follows the well-known noncentral t-distribution with degrees of freedom $n - 1$ and noncentrality parameter $\sqrt{n}\,\delta_0$. Denote the CDF of the noncentral t as

$T_{n-1}(t|\sqrt{n}\delta_0)$ and let $t_{n-1}^{(1-\alpha)} = T_{n-1}^{-1}[(1-\alpha)\,|\,0]$ be the critical value corresponding to probability $1-\alpha$ of the central t-distribution. Then the power of this test is given by

$$\gamma_t(\delta_0) = 1 - T_{n-1}(t_{n-1}^{(1-\alpha)}|\sqrt{n}\delta_0). \tag{9}$$

When replacing the δ_0 by its estimator we get an observed power. In this case, it is easy to show that $\bar{x}/s$ is not an unbiased estimator for δ_0. Let

$$c_n = \Gamma[(n-1)/2]/\{\sqrt{(n-1)/2}\Gamma[(n-2)/2]\}.$$

Then an unbiased estimator of δ_0 is given by $\hat{\delta} = c_n\bar{x}/s$. Replace the δ_0 in Equation 9 by its unbiased estimator we have the observed power $\gamma_t(\hat{\delta})$. Based on the results in section 2.1 we would expect a bias in $\gamma_t(\hat{\delta})$, that is $E[\gamma_t(\hat{\delta})] \neq \gamma_t(\delta_0)$. Note that $\sqrt{n}\hat{\delta}/c_n$ follows $T_{n-1}(t|\sqrt{n}\delta_0)$, the expectation of $\gamma_t(\hat{\delta})$ can be expressed as

$$E[\gamma_t(\hat{\delta})] = 1 - \int_{-\infty}^{\infty} T_{n-1}(t_{n-1}^{(1-\alpha)}\,|\,c_n u)dT_{n-1}(u|\sqrt{n}\delta_0). \tag{10}$$

Because the form of $T_{n-1}(u|\sqrt{n}\delta_0)$ is not so simple, a clean formula like the one in Equation 2 is not available. We need to use a numerical procedure to evaluate Equation 10. Details of this are provided in the appendix.

For $\delta_0 = 0.2$, 0.5 and 0.8, the plots of the true power $\gamma_0 = \gamma_t(\delta_0)$, the relative bias in $\gamma_t(\hat{\delta})$ and its *SD* against the sample size n are in Figure 4, where we choose $n = 15$ to 50 to contrast with Figure 1. Similar to the findings when σ_0 is known, the bias is positive for the small effect size and negative for the medium and large effect sizes. For the small effect size, the bias can be substantial, which renders the observed power of little value. For the small effect size, the *SD* of $\gamma_t(\hat{\delta})$ does not necessarily become smaller as $\gamma_t(\delta_0)$ or n increases. Actually, Figure 4 and Figure 1 contain essentially the same picture. For the same sample size, the powers in Figure 4 are a little smaller, and the variances are slightly larger than those in Figure 1. Most of the biases in Figure 4 are slightly larger, but when $\delta_0 = 0.2$ some of these in Figure 1 are slightly larger.

It is also interesting to get the overall picture of the distribution of $\gamma_t(\hat{\delta})$. Because the random variable $\hat{\delta}$ is in the noncentrality parameter of the noncentral t-distribution, we could not figure out a way to evaluate the CDF

$$F_{\gamma_t}(u) = P\{\gamma_t(\hat{\delta}) \leq u\}$$

exactly or even numerically. We have to use simulations instead. In studying the distribution of $\gamma_t(\hat{\delta})$, we want to know whether the density for F_{γ_t} can be approximately described by these in Figure 3. Unlike F_{γ_t}, the analytical form of F_{γ_t} is not available, and it generally depends on both n and δ_0. It is not clear whether $F_{\gamma t}(u)$ for $\gamma_0 = 0.6$ is a flip of that when $\gamma_0 = 0.4$ or whether F_{γ_t} is uniformly distributed when $\gamma_0 = 0.5$. For the purpose of comparing F_{γ_t} with F_{γ_t}, we approximate the population power $\gamma_0 = \gamma_t(\delta_0) = 0.1, 0.2, \ldots, 0.9$ in the simulation condition by adjusting δ_0 and n. Specifically, for $\gamma_0 = 0.1$ to 0.5, we fixed $\delta_0 = 0.2$ and found the n so that the resulting $\gamma_t(\delta_0)$ in Equation 9 is closest to the given γ_0. Fixing the n at the obtained value, we then adjust the δ_0 for $\gamma_t(\delta_0)$ to approximately equal the γ_0. For $\gamma_0 = 0.6$ to 0.9, we performed the same procedure by starting with $\delta_0 = 0.5$. The combination of n, δ_0 and the corresponding γ_0 used in the simulation are reported in Figure 5, where 20,000 replications are used to evaluate $\gamma_t(\hat{\delta})$ with the unbiased $\hat{\delta}$. Each of the nine histograms is based on 50 equally divided spaces.

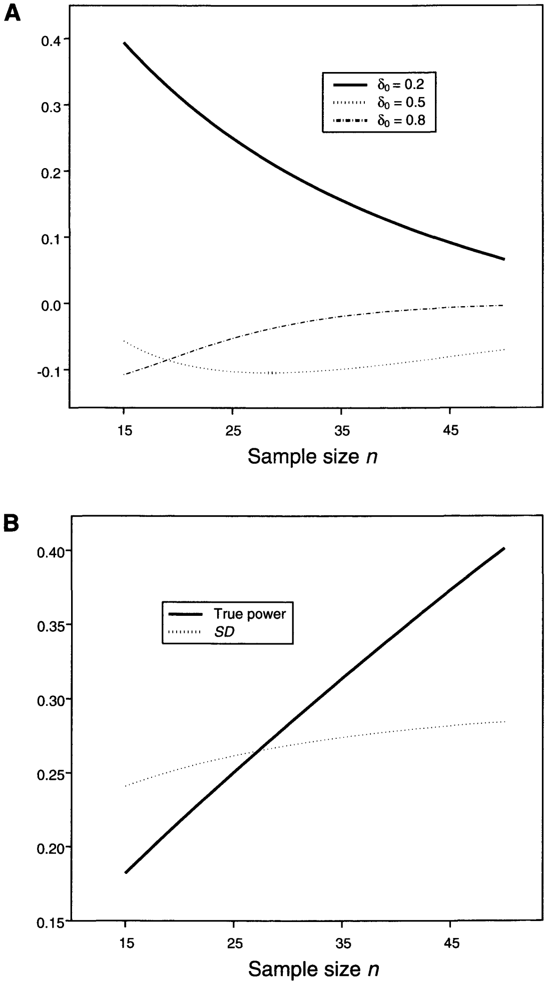


FIGURE 4. *True power, relative bias, and standard deviation of the observed power* $\gamma_t(\hat{\delta})$ *(effect size* $\delta_0 = 0.2, 0.5, 0.8$*; sample size* n *= 15–50).* **A**, *relative bias;* **B**, *effect size* $\delta_0 = 0.2$. *(continued)*

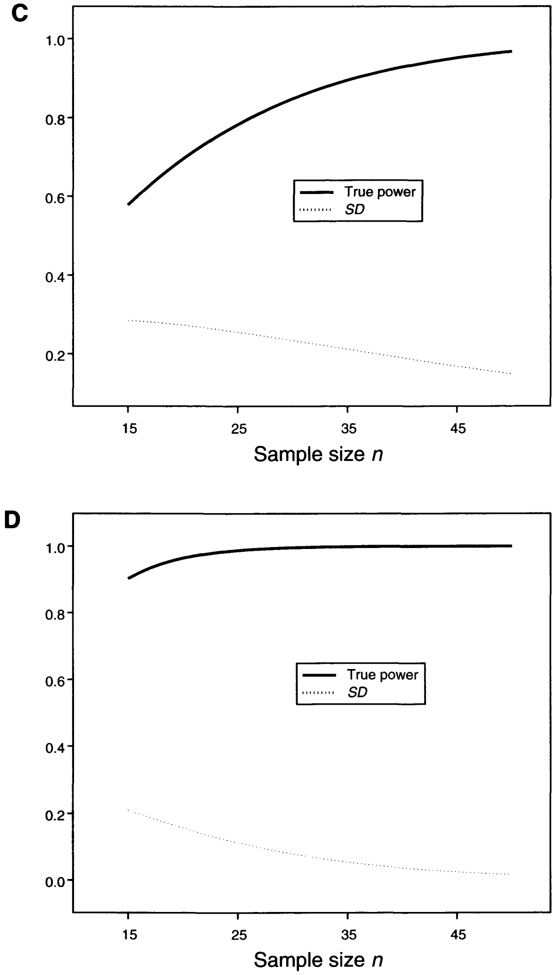


FIGURE 4. (*continued*). **C**, *effect size* $\delta_0 = 0.5$; **D**, *effect size* $\delta_0 = 0.8$.

When $\gamma_0 \approx 0.1$, the distribution of $\gamma_t(\hat{\delta})$ has a big mass at the low end with peak occurring at 0. As γ_0 increases from 0.1 to 0.4, the peak is still at 0, but the distribution spreads to the other part of the range. Comparing the first four histograms to the density plots in Figure 3, we notice that the shapes of the histograms are very much like those of the density plots. When $\gamma_0 = 0.5$, the histogram tells us that $\gamma_t(\hat{\delta})$ approximately follows a uniform distribution. The histograms corresponding to $\gamma_0 = 0.6$ to 0.9 are approximately flips of those corresponding to $\gamma_0 = 0.4$ to 0.1, respectively. This is expected because the behavior of a *t*-statistic can be approximately described by a normal distribution.

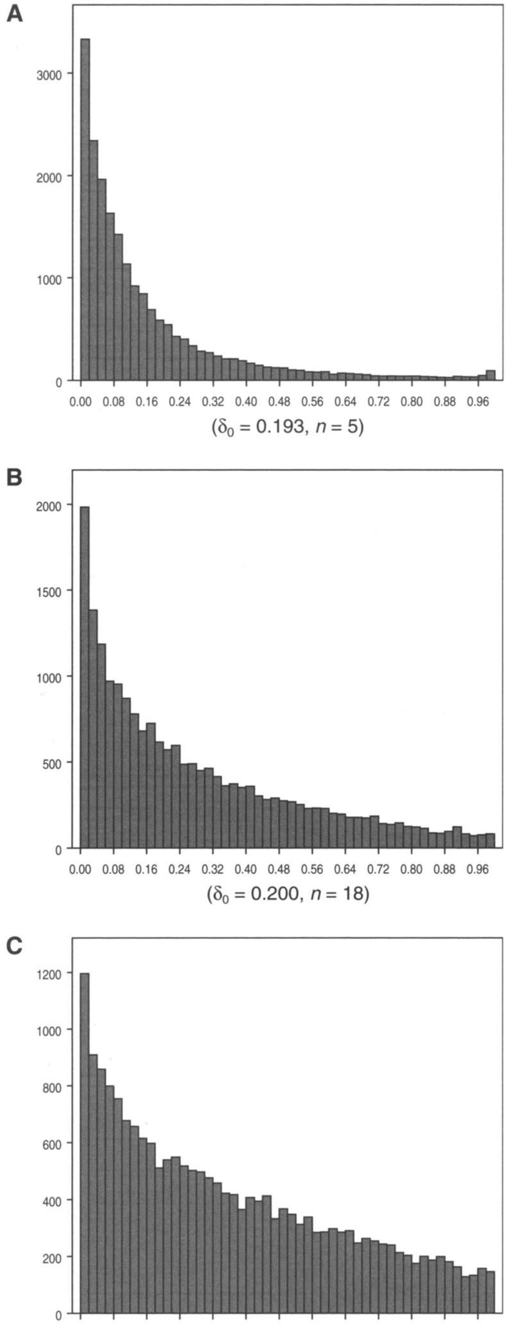
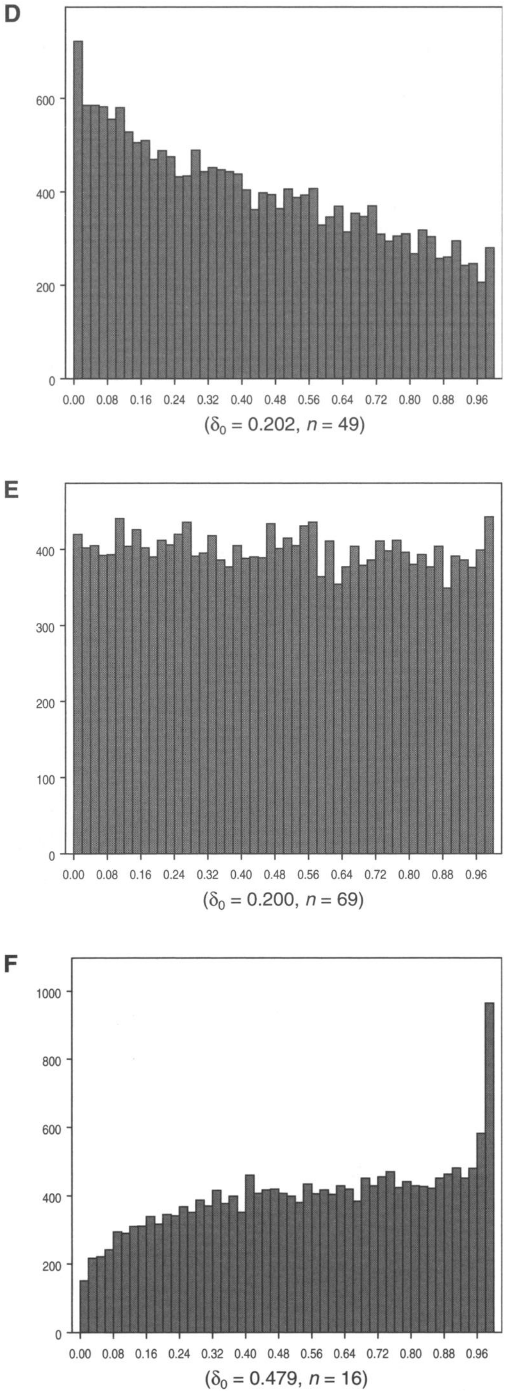


FIGURE 5. *Histogram (density) of $\gamma_t(\hat{\delta})$ when true power γ_0 = 0.1–0.9.* **A**, *true power γ_0 = 0.100;* **B**, *true power γ_0 = 0.203;* **C**, *true power γ_0 = 0.301.* (*continued*)



(δ_0 = 0.202, n = 49)

(δ_0 = 0.200, n = 69)

(δ_0 = 0.479, n = 16)

FIGURE 5. **D**, *true power* γ_0 *= 0.401;* **E**, *true power* γ_0 *= 0.500;* **F**, *true power* γ_0 *= 0.600.* (*continued*)

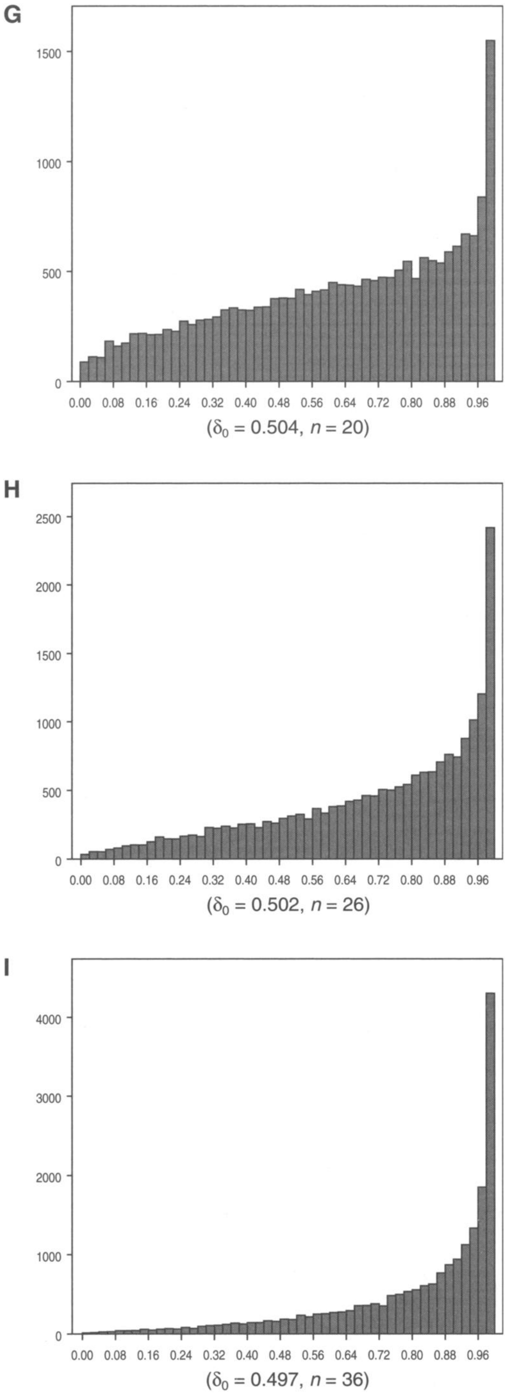


FIGURE 5. (*continued*). **G**, *true power* γ_0 = *0.701;* **H**, *true power* γ_0 = *0.801;* **I**, *true power* γ_0 =*0.900.*

In summary, the distribution of $\gamma_t(\hat{\delta})$ is highly skewed unless $\gamma_0 = 0.5$. We also predict that the median of $\gamma_t(\hat{\delta})$ is γ_0. Further study in this direction is needed. It is also interesting to know, after γ_0 being given, whether the distribution of $\gamma_t(\hat{\delta})$ is mathematically independent of n and δ_0.

2.4. Two Groups With Equal but Unknown Variance

Consider $x_1, \ldots, x_{n_1} \sim N(\mu_{10}, \sigma_0^2)$ and $y_1, \ldots, y_{n_2} \sim N(\mu_{20}, \sigma_0^2)$, where μ_{10}, μ_{20} and σ_0^2 are all unknown. The statistic for testing the hypothesis in Equation 7 becomes

$$t = \sqrt{\tilde{n}}\left(\bar{y} - \bar{x}\right)/s, \tag{11}$$

where s is the pooled sample standard deviation and $\tilde{n} = n_1 n_2/(n_1 + n_2)$. It is well-known that the t in Equation 11 follows a noncentral t-distribution with degrees of freedom $n_1 + n_2 - 2$ and noncentrality parameter $\sqrt{\tilde{n}}\delta_0$. Let $n_3 = n_1 + n_2 - 2$ and

$$c_{n_{12}} = \Gamma[(n_1 + n_2 - 2)/2]/\{\sqrt{(n_1 + n_2 - 2)/2}\Gamma[(n_1 + n_2 - 3)/2]\},$$

the unbiased estimator of δ_0 is given by $\hat{\delta} = c_{n_{12}}\left(\bar{y} - \bar{x}\right)/s$. The observed power is

$$\gamma_t(\hat{\delta}) = 1 - T_{n_3}\left(t_{n_3}^{(1-\alpha)} \middle| \sqrt{\tilde{n}}\hat{\delta}\right). \tag{12}$$

The properties of the $\gamma_t(\hat{\delta})$ in Equation 12 will be much the same as the one of the one sample problem considered in section 2.3. That is, it overestimates the true power when $\tilde{n}$ is below 50 for the small effect size, and underestimates the true power when $\tilde{n} > 15$ for the medium and large effect sizes. Its distribution is also highly skewed when γ_0 is away from 0.5, and approximately uniformly distributed when $\gamma_0 = 0.5$.

We have only studied the observed powers for one-sided tests. The discussion also implies the difficulty with estimating the true power for a two-sided test. Considering the two-sided test for $H_0 : \mu_0 = 0$ vs. $H_1 : \mu_0 \neq 0$ and $\delta_0 > 0$. When the variance σ_0^2 is known, the power function is

$$\gamma_z(\delta_0) = 1 - \Phi\left(z_{(1-\alpha/2)} - \sqrt{n}\delta_0\right) + \Phi\left(-z_{(1-\alpha/2)} - \sqrt{n}\delta_0\right)$$

and

$$E\left[\gamma_z(\hat{\delta})\right] = 1 - \Phi\left[\left(z_{(1-\alpha/2)} - \sqrt{n}\delta_0\right)/\sqrt{2}\right] + \Phi\left[\left(-z_{(1-\alpha/2)} - \sqrt{n}\delta_0\right)/\sqrt{2}\right].$$

It is obvious that $E\left[\gamma_z(\hat{\delta})\right] \neq \gamma_z(\delta_0)$. Actually, the difference between $\Phi\left(-z_{(1-\alpha/2)} - \sqrt{n}\delta_0\right)$ and $\Phi\left[\left(-z_{(1-\alpha/2)} - \sqrt{n}\delta_0\right)/\sqrt{2}\right]$ is tiny because $\Phi(\cdot)$ is not sensitive at its tails. The bias in $\gamma_z(\hat{\delta})$ is mainly due to the difference between $\Phi\left[\left(z_{(1-\alpha/2)} - \sqrt{n}\delta_0\right)/\sqrt{2}\right]$ and $\Phi\left(z_{(1-\alpha/2)} - \sqrt{n}\delta_0\right)$ which has been discussed in detail.

3. Examples

In this section, we present three examples to illustrate empirically the analytical property of the post hoc power obtained in the previous section. The first two are based on real data, the third one is based on simulated data.

Example 1

Table 2.1 of Efron and Tibshirani (1993, p. 11) contains the survival times of 16 mice randomly assigned to two conditions. Seven received medical treatment, nine are under control. The interest is whether the mean survival time μ_1 of the experimental group is greater than that of the control group μ_2. Assuming a common variance, the problem falls into the context of section 2.4. With a t score of 1.121, there is not enough evidence against the null hypothesis $H_0 : \mu_1 = \mu_2$. With an unbiased estimate of $\hat{\delta} = 0.534$, the post hoc power for this study is $\hat{\gamma}_t = 0.262$. However, we cannot conclude that the power for this study is approximately 0.262 because of the bias and variance of $\hat{\gamma}_t$. Similarly, we cannot use $\delta = 0.534$ in determining the required sample size for a future study to achieve a power of, say, $\gamma = 0.90$.

Example 2

To compare different methods of teaching arithmetic, 45 students were randomly divided into five groups, each with nine students. Table 3.1 of Everitt (1996, p. 50) contains test scores of these students after being taught by each of the five methods. The interest here is in the difference between different teaching methods. We only use the data for the first two methods to illustrate the problem with the post hoc power. The t-score for the first two methods is 0.724, which is not significant. The unbiased estimator for $\delta = (\mu_1 - \mu_2)/\sigma$ is $\hat{\delta} = 0.325$, which is between small and medium. If the test is two-sided $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$, then the post hoc power is only $\hat{\gamma}_t = 0.099$. If the test is one-sided $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 > \mu_2$, the post hoc power is $\hat{\gamma}_t = 0.163$. According to the results in the previous section, these power estimates do not provide valuable information about the power underlying the current study. The estimated effect size cannot be used to guide the data collection in a future study for comparing the two teaching methods.

There is a common limitation in the previous two examples because we do not know the true effect size and thus cannot compare the post hoc power with the true power. The next example is based on simulated data for which the true power is known.

Example 3

To contrast the post hoc power with the true power, three samples are generated. The first one $x \sim N(\mu_0, 1)$ is to illustrate the post hoc power with one group. The second sample $x \sim N(0, 1)$ and the third one $y \sim N(\mu_0, 1)$ are to illustrate the post hoc power with two groups. For simplicity, all the three samples have size 15. Because the population variances of the samples are known, we can set σ at

TABLE 1
Post Hoc Power vs. Population Power

	(a) $\delta_0 = 0.5$			
	One-group z test	One-group t test	Two-group z test	Two-group t test
γ_0	0.61	0.58	0.39	0.38
$\hat{\gamma}$	0.93	0.92	0.10	0.09
	0.88	0.72	0.74	0.69
	0.19	0.18	0.21	0.20
	0.31	0.25	0.12	0.11
	0.46	0.49	0.64	0.52

	(b) $\delta_0 = 0.8$			
	One-group z test	One-group t test	Two-group z test	Two-group t test
γ_0	0.93	0.90	0.71	0.69
$\hat{\gamma}$	1.00	0.99	0.32	0.24
	0.99	0.93	0.93	0.90
	0.61	0.59	0.50	0.48
	0.75	0.63	0.36	0.34
	0.85	0.89	0.88	0.77

their population value 1.0 so that a z test can be used in testing the mean. We can also treat σ as unknown so that a t test is needed. Because the population effect size $\delta_0 = \mu_0$ is known, we can obtain the population powers for the one-group z test, one-group t test, two-group z test, and two-group t test. Setting $\mu_0 = 0.5$, the population powers (γ_0) are in the first row of Table 1(a). The next five rows of numbers ($\hat{\gamma}$) in Table 1(a) are the post hoc power based on estimated effect sizes with five replications. It is clear that none of the post hoc powers is near the corresponding true power. For example, with the one group z test, the true population power is 0.61, the post hoc power ranges from 0.19 to 0.93. Using any of the five post hoc powers in the role of the true power can be misleading. Table 1(b) contains the parallel result when $\delta_0 = \mu_0 = 0.8$. Results in the previous section suggest that the post hoc power is more stable when the population power γ_0 is bigger. This is empirically verified by the results in Table 1(b), especially for the one group case when γ_0 is above 0.9.

4. Conclusion

Retrospective power analysis has sometimes been required by reviewers or journal editors. We give a systematic study of the commonly used power estimator in a retrospective analysis. We found that the observed power is almost always a biased estimator of the true power. Its distribution is highly skewed when γ_0 is not near 0.5. When $\gamma_0 = 0.5$, the variance of the observed power reaches its maximum. As a point estimator, the observed power cannot offer useful information when γ_0 is small. This is because of not only its variance but also its substantial relative bias. When $\gamma_0 > 0.37$, the relative bias is roughly around or below 10%, which may be acceptable. When

$\gamma_0 > 0.78$, the *SD* of the observed power is less than 1/3 of γ_0 so that the observed power may provide some useful information. For a small γ_0, the observed power may not provide any useful information regardless of the sample size!

Power tables in Cohen (1988) are based on fixed-effect size. In practice, few people know their true effect size. The observed effect size and effect size obtained from a previous study are actually random quantities. Effect size obtained from a meta-analysis also contains sampling errors. Actually, we may think of such an effect size as estimated from a study with a large sample size. Our results indicate that the commonly used power estimator does not provide useful information when power is small. However, if the goal is to determine the sample size needed to detect an effect that is considered practically significant, then one can refer to tables in Cohen (1988) for valuable information. In such a context, we can decide the practically significant effect size without consulting the post hoc effect size. Using post hoc power to determine sample size for next step data collection is not recommended.

The confidence interval may provide a better picture of the power than the point estimator. For a small effect size, the sample size has to be relatively large so that the interval will be at least above 0.05.

We have also explored a less biased method of estimating retrospective power. Unfortunately, it is unlikely to be a better estimator than the commonly used formula.

Appendix

This appendix provides the expectation of $\gamma_z(\hat{\delta})$ analytically, and outlines how $\mathrm{Var}[\gamma_z(\hat{\delta})]$, $E[\gamma_t(\hat{\delta})]$ and $\mathrm{Var}[\gamma_t(\hat{\delta})]$ were numerically obtained.

Note that $\gamma_z(\hat{\delta})$ is a special case of $\gamma_{ab}(\hat{\delta})$, we will just obtain the expectation of $\gamma_{ab}(\hat{\delta})$. It is obvious

$$E[\gamma_{ab}(\hat{\delta})] = 1 - E\{\Phi[a(z_{(1-\alpha)} - \sqrt{n}\hat{\delta}) + b]\}.$$

So we only need to obtain $E\{\Phi[a(z_{(1-\alpha)} - \sqrt{n}\hat{\delta}_0) + b]\}$. When σ_0 is known, $\sqrt{n}\hat{\delta} = \sqrt{n\bar{x}}/\sigma_0$ follows the normal distribution $N(\sqrt{n}\delta_0, 1)$. So we have $\sqrt{n}\hat{\delta} = \sqrt{n}\delta_0 + z$, where $z \sim N(0, 1)$.

Consequently,

$$E\{\Phi[a(z_{(1-\alpha)} - \sqrt{n}\hat{\delta}) + b]\} = E\{\Phi[a(z_{(1-\alpha)} - \sqrt{n}\delta_0) + b + az]\}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \Phi[a(z_{(1-\alpha)} - \sqrt{n}\delta_0) + b + at] e^{-t^2/2} dt. \tag{Al}$$

Let $h = a(z_{(1-\alpha)} - \sqrt{n}\delta_0) + b$. Then

$$\Phi[a(z_{(1-\alpha)} - \sqrt{n}\delta_0) + b + at] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{h+at} e^{-u^2/2} du. \tag{A2}$$

Putting Equation A2 into Equation A1,

$$E\{\Phi(h + az)\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left[e^{-t^2/2} \int_{-\infty}^{h+at} e^{-u^2/2} du \right] dt. \tag{A3}$$

We should realize that the double integrals in Equation A3 are with respect to the density functions of two independent standard normal variables T and U, which are actually the probability

$$P\{aU + T \le h\} = P\left\{(aU + T)/\sqrt{a^2 + 1} \le h/\sqrt{a^2 + 1}\right\} \tag{A4}$$

Note that $(aU + T)/\sqrt{a^2 + 1} \sim N(0, 1)$, the probability in Equation A4 is just $\Phi\left(h/\sqrt{a^2 + 1}\right)$. This gives the expectations of $\gamma_z(\hat{\delta})$ and $\gamma_a(\hat{\delta})$ in section 2.1.

Now we turn to the variance of $\gamma_z(\hat{\delta})$. It is easy to see

$$\mathrm{Var}[\gamma_a(\hat{\delta})] = \mathrm{Var}\left\{\Phi\left[a\left(z_{(1-\alpha)} - \sqrt{n}\hat{\delta}\right) + b\right]\right\} = \mathrm{Var}\{\Phi(h + az)\}.$$

Because $\mathrm{Var}[\Phi(h + az)] = E[\Phi(h + az)]^2 - \{E[\Phi(h + az)]\}^2$ and $E[\Phi(h + az)]$ is known, we only need to outline the procedure for obtaining $E[\Phi(h + az)]^2$. Because $z \sim N(0, 1)$, we have

$$E[\Phi(h + az)]^2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} [\Phi(h + at)]^2 e^{-t^2/2} dt.$$

Note that $0 < \Phi(h + at) < 1.0$, we have

$$0 < \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} [\Phi(h + at)]^2 e^{-t^2/2} dt - \frac{1}{\sqrt{2\pi}} \int_{-c}^{c} [\Phi(h + at)]^2 e^{-t^2/2} dt \le 2\Phi(-c).$$

When choosing $c = 6.0$, $E[\Phi(h + az)]^2$ can be approximated by

$$\frac{1}{\sqrt{2\pi}} \int_{-c}^{c} [\Phi(h + at)]^2 e^{-t^2/2} dt \tag{A5}$$

with an error of less than 10^{-8}. The integral in Equation A5 can be evaluated by the well-known Simpson rule of numerical integration (see Etter, 1992). We used $m = 50{,}000$ points in the Simpson rule when calculating the numbers for plots in Figures 1 and 2. Double the number of points gave the same value of the integral to the 8th decimal place. Note that we have used the SAS internal function *probnorm*$(h + at)$ when evaluating $\Phi(h + at)$.

The CDF F_{γ_z} in section 2.1 was obtained using the following steps:

$$\begin{aligned} F_{\gamma_z}(t) &= P\left\{\gamma_z(\hat{\delta}) \le t\right\} \\ &= P\left\{1 - \Phi\left(z_{(1-\alpha)} - \sqrt{n}\hat{\delta}\right) \le t\right\} \end{aligned}$$

$$
\begin{aligned}
&= P\{\Phi(z_{(1-\alpha)} - \sqrt{n}\hat{\delta}) \geq 1 - t\} \\
&= P\{z_{(1-\alpha)} - \sqrt{n}\hat{\delta} \geq \Phi^{-1}(1 - t)\} \\
&= P\left\{\sqrt{n}(\hat{\delta} - \delta_0) \leq z_{(1-\alpha)} - \sqrt{n}\delta_0 - \Phi^{-1}(1 - t)\right\} \\
&= \Phi[(z_{(1-\alpha)} - \sqrt{n}\delta_0) - \Phi^{-1}(1 - t)].
\end{aligned}
$$

The density function f_{γ_z} follows by taking the derivative of F_{γ_z}.

Because the γ_L and γ_U in Equation 6 can be obtained in essentially the same way, we only give the steps to obtaining γ_L. Notice that $F_{\gamma_z}(t) = \Phi(z_{(1-\gamma_0)} + z_t)$, and the observed value of t is $t_0 = \hat{\gamma}_z$. Setting $F_{\gamma_z}(\hat{\gamma}_z) = 0.95$ and using $z_x = \Phi^{-1}(x)$, we have

$$\Phi^{-1}(1 - \gamma_L) + \Phi^{-1}(\hat{\gamma}_z) = z_{0.95}. \tag{A6}$$

Moving $\Phi^{-1}(\hat{\gamma}_z)$ in Equation A6 to the right side and applying the transformation $\Phi(\cdot)$ give

$$1 - \gamma_L = \Phi[z_{0.95} - \Phi^{-1}(\hat{\gamma}_z)]. \tag{A7}$$

Because $\hat{\gamma}_z = 1 - \Phi(z_{(1-\alpha)} - \sqrt{n}\hat{\delta})$ and $\Phi^{-1}(1 - x) = -\Phi^{-1}(x)$,

$$
\begin{aligned}
\Phi^{-1}(\hat{\gamma}_z) &= \Phi^{-1}[1 - \Phi(z_{(1-\alpha)} - \sqrt{n}\hat{\delta})] \\
&= - \Phi^{-1}[\Phi(z_{(1-\alpha)} - \sqrt{n}\hat{\delta})] \\
&= \sqrt{n}\hat{\delta} - z_{(1-\alpha)}.
\end{aligned} \tag{A8}
$$

Combining Equations A7 and A8 gives

$$\gamma_L = 1 - \Phi(z_{0.95} + z_{(1-\alpha)} - \sqrt{n}\hat{\delta}).$$

The integral in Equation 10 of section 2.3 was evaluated similarly as in evaluating Equation A5. First, finite numbers c_l and c_u were found to replace the infinity limits of the integral. The proper c_l and c_u depend on the noncentrality parameters and the degrees of freedom. Large enough numbers are chosen so that the integral outside $[c_l, c_u]$ is less than 10^{-8} in calculating all the numbers used for plots in Figure 4. For the finite c_l and c_u, the integral in Equation 10 was evaluated by the well-known trapezoidal rule for numerical integration (see Etter, 1992). We used $m = 50{,}000$ trapezoids so that the first eight digits of the integral value are the same even when double the number of trapezoids. Specifically, the SAS internal function *probt(t, df,* λ) provides the value of $T_{df}(t\,|\,\lambda)$. The integral on the interval $[a, b]$ is given by

$$\begin{aligned} I_{ab} &= \left[probt(b, n-1, \sqrt{n}\delta_0) - probt(a, n-1, \sqrt{n}\delta_0)\right] \\ &\quad \times \left[probt(t_{n-1}^{(1-\alpha)}, n-1, c_n a) + probt(t_{n-1}^{(1-\alpha)}, n-1, c_n b)\right]/2\,. \end{aligned}$$

The variance of $\gamma_t(\hat{\delta})$ was evaluated similarly.

References

American Psychological Association (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Cohen, J. J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Dudewicz, E. J. (1972). Confidence intervals for power, with special reference to medical trials. *Australian Journal of Statistics, 14,* 211–216.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* New York: Chapman & Hall.

Etter, D. M. (1992). *Fortran 77 with numerical methods for engineers and scientists.* Redwood City, CA: Benjamin/Cummings.

Everitt, B. S. (1996). *Making sense of statistics in psychology.* Oxford: Oxford University Press.

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the Journal of Applied Psychology: Little evidence of reform. *Educational and Psychological Measurement, 61,* 181–210.

Gillett, R. (1994a). The average power criterion for sample size estimation. *Statistician, 43,* 389–394.

Gillett, R. (1994b). Post hoc power analysis. *Journal of Applied Psychology, 79,* 783–785.

Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and *p* values: What should be reported and what should be replicated? *Psychophysiology, 33,* 175–183.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculation for data analysis. *American Statistician, 55,* 19–24.

Instructions to authors (2001). *Animal Behavior, 62,* i–viii.

MacCallum, R. C., Browne, M. W. & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1,* 130–149.

Macdonald, R. R. (2003). On determining replication probabilities: Comments on Posavac (2002). *Understanding Statistics, 2,* 69–70.

Pallant, J. (2001). *SPSS survival manual: A step by step guide to data analysis using SPSS for Windows* (Versions 10 and 11). Philadelphia, PA: Open University Press.

Posavac, E. J. (2002). Using *p* values to estimate the probability of a statistically significant replication. *Understanding Statistics, 1,* 101–112.

Posavac, E. J. (2003). Response to Macdonald. *Understanding Statistics, 2,* 71–72.

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests* (pp. 221–257). Hillsdale, NJ: Erlbaum.

Stevens, J. P. (1999). *Intermediate statistics: A modern approach* (2nd ed.). Mahwah, NJ: Erlbaum.

Taylor, D. J., & Muller, K. E. (1995). Computing confidence bounds for power and sample size of the general linear univariate model. *American Statistician, 49,* 43–47.

Authors

KE-HAI YUAN is Associate Professor, Department of Psychology and Lab for Social Research, University of Notre Dame, Notre Dame, IN 46556; kyuan@nd.edu. His areas of specialization are psychometric theory and applied multivariate statistics.

SCOTT MAXWELL is Professor, Department of Psychology, University of Notre Dame, Notre Dame, IN 46556; smaxwell@nd.edu. His areas of specialization are experimental design and applied behavioral statistics.

Manuscript received March 17, 2003
Revision received October 17, 2003
Accepted November 3, 2003