

Post hoc power analysis: is it an informative and meaningful analysis?

Yiran Zhang,¹ Rita Hedo,¹ Anna Rivera,^{2,3} Rudolph Rull,² Sabrina Richardson,^{2,3} Xin M Tu¹

To cite: Zhang Y, Hedo R, Rivera A, *et al.* Post hoc power analysis: is it an informative and meaningful analysis? *General Psychiatry* 2019;**32**:e100069. doi:10.1136/gpsych-2019-100069

Received 19 March 2019
Accepted 26 March 2019

SUMMARY

Power analysis is a key component for planning prospective studies such as clinical trials. However, some journals in biomedical and psychosocial sciences ask for power analysis for data already collected and analysed before accepting manuscripts for publication. In this report, post hoc power analysis for retrospective studies is examined and the informativeness of understanding the power for detecting significant effects of the results analysed, using the same data on which the power analysis is based, is scrutinised. Monte Carlo simulation is used to investigate the performance of posthoc power analysis.

INTRODUCTION

Power analysis plays a key role in designing and planning prospective studies. For clinical trials in biomedical and psychosocial research, power analysis provides critical information about sample sizes needed to detect statistically significant and clinically meaningful differences between different treatment groups. Power analysis also provides critical information for evaluating cost–benefit ratios so that studies can be conducted with minimal resources without compromising on scientific integrity and rigour.

What is interesting is that some journals also ask for power analysis for the study data that were already analysed and reported in a manuscript before considering its publication. Although the exact purposes of such requests are not clearly stated, it seems that this often happens when manuscripts include some non-significant results. As such post hoc power analysis is conceptually flawed, concerns have been raised over the years.^{1–4} Despite these warnings, some journals continue to ask for such information and use it as part of their decision process for manuscript publications.

As most research studies are conducted based on a random sample from a study population of interest, results from power analysis become meaningless, as the random component in the study disappears once data are collected. Power analysis shows the

probability, or likelihood, for a statistical test or model to detect, say, hypothesised differences between two populations, such as the *t* statistic for comparing, say, mean blood pressure level between two groups in a sample of interest in a prospective study. If a sample is selected, outcomes are no longer random and power analysis becomes meaningless for this particular study sample.

Nevertheless, some continue to argue that such power analyses may help provide some indication whether a hypothesis still may be true.^{2–6} For example, if a power analysis based on observed outcomes of interest in a study shows that the sample has low power such as 60% to detect, say, a medium effect size, or Cohen's *d*=0.5 when comparing the means of two group,¹ they argue that this explains why the study fails to find statistically significant results. Therefore, the question is not whether post hoc power analyses makes conceptual sense, but rather if such power estimates can inform power for detecting significant results.

In this article, we focus on comparing the means between two groups on a continuous outcome, and use Monte Carlo simulation to investigate the performance of post hoc power analysis and to see if such power estimates are informative in terms of indicating power to detect statistically significant differences already observed. We begin our discussion with a brief overview of the concept and analytic evaluation of power analysis within the context of two independent samples, or groups.

POWER ANALYSIS FOR COMPARING TWO POPULATION MEANS

We have considered to use two independent samples and let Y_{ik} denote a continuous outcome of interest from subject *i* and group *k* ($1 \leq i \leq n_k$, $k=1,2$). For simplicity and without loss of generality, we assume that for both groups, Y_{ik} follows a normal distribution of



© Author(s) (or their employer(s)) 2019. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹Department of Family Medicine and Public Health, University of California System, Oakland, California, USA

²Naval Health Research Center, San Diego, California, USA

³Leidos, San Diego, California, USA

Correspondence to

Dr Xin M Tu, Family Medicine and Public Health, University of California San Diego, La Jolla, CA 92093, USA; x2tu@ucsd.edu

population mean μ_k and common population variance σ^2 , denoted as

$$N(\mu_k, \sigma^2) \quad (1 \leq i \leq n_k, k = 1, 2).$$

The most common hypothesis in this setting is whether the population means are equal to each other. In statistical lingo, we state the hypothesis as follows:

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs.} \quad H_a : \mu_1 - \mu_2 = \delta \neq 0 \quad (1)$$

where δ is a known (unknown) constant for power (data) analysis, and H_0 and H_a are known as the null and alternative hypotheses, respectively. The above is known as a two-sided hypothesis, as no direction of effect is specified in the alternative hypothesis $H_a : \mu_1 - \mu_2 = \delta \neq 0$. If the directional effect is also indicated, the alternative becomes a one-sided hypothesis. For example, if the alternative is specified as: $H_a : \mu_1 - \mu_2 = \delta > 0$, it is one sided and μ_1 is hypothesised to be larger than μ_2 under H_a . As two-sided alternatives are the most popular in clinical research, we only consider two-sided alternatives in this paper unless stated otherwise. Note also that when testing the hypothesis in equation (1) with data as in data analysis, δ is an unknown constant and p values are calculated based on the null H_0 without any knowledge about δ in the alternative H_a . For power analysis, the difference δ must be specified, since power depends on this parameter. In practice, the normalised difference, $d = \frac{|\mu_1 - \mu_2|}{\sigma}$, is often used, as it is more succinct and generally has a more meaningful interpretation. This normalised difference is known as Cohen's d .¹

The hypothesis in equation (1) is generally tested using the two sample t-test. Let $\bar{Y}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} Y_{ki}$ denote the sample mean of the k th group ($k=1,2$). The difference between the two sample means $\bar{Y}_2 - \bar{Y}_1$ should be close to 0 if H_0 is true. Again, because \bar{Y}_2 and \bar{Y}_1 are random, it is still possible for $\bar{Y}_2 - \bar{Y}_1$ to be very different from 0, although such probabilities are small, especially for large sample size n , a statistical result known as the law of large numbers.⁷ To address this sampling variability in causing wrong conclusions about the hypothesis, type I error, typically denoted by α , is used to indicate the extent to which the difference, $\bar{Y}_2 - \bar{Y}_1$, departs from 0. This error rate is typically set at $\alpha=0.05$ for most studies. For very large sample sizes, α is generally set at a more stringent level, $\alpha=0.01$. Given α , power is the probability that H_0 is rejected when H_a in fact is true.

If $H_0 : \mu_1 - \mu_2 = 0$ is true, the probability of rejecting H_0 , therefore committing type I error α , is the probability:

$$P\left(\left|\frac{\bar{Y}_1 - \bar{Y}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right| \geq z_{\alpha/2}\right), \quad (2)$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution, $P(A)$ denotes the probability of the occurrence of an event A and s is the pooled SD:

$$s = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}, \quad s_k^2 = \frac{1}{n_k-1} \sum_{i=1}^{n_k} (Y_{ki} - \bar{Y}_k)^2, \quad k = 1, 2.$$

The probability in equation (2) is readily evaluated using the t distribution.

Note that like the effect size, the difference between the sample means in equation (2) has also been normalised (by the pooled SE) so that the type I error α does not depend on other artefacts such as different scales that may be used for the outcomes. Note also that we assume a common SD (or variance) between the groups for the t -statistic in equation (2). If this is not the case, we can use the version of the two sample t-test for unequal variances, called the Welch's t-test. For simplicity and without loss of generality, we focus on the equal variance version in equation (2) in what follows. The same conclusions also apply to the Welch's t-test.

For power analysis, we want to determine the probability to reject the null H_0 in favour of the alternative H_a . Given the type I error α , sample size n_1 and n_2 , and H_0 and H_a , we can calculate power for reject the null H_0 :

$$\text{power}(n_1, n_2, \alpha, H_0, H_1) = P\left(\left|\frac{\bar{Y}_1 - \bar{Y}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right| \geq z_{\alpha/2} \mid H_a\right). \quad (3)$$

Although similar in appearance, equation (3) is actually quite different from equation (2). equation (2) is generally used to compute p values, in which case $\delta = \bar{Y}_1 - \bar{Y}_2$ and s are all readily computed from the sample means and sample SD from the observed outcomes. As none of these quantities is available when performing power analysis for prospective studies, the power function in equation (3) is evaluated based on sample sizes, n_1 and n_2 , the population mean difference δ and SD σ . Thus, unlike data analysis, all these parameters must be explicitly specified in order to compute power. Although results from similar studies may be used to help suggest and specify δ and σ , they are both conceptually and analytically different from their sample counterparts. Conceptually, δ and σ are determined by the entire study population, while δ and s are specific to a random sample from the study population. Analytically, δ and σ are population-level constants, whereas δ and s are random quantities and vary from sample to sample. The population parameters can be quite different from their sample counterparts.

In practice, we often set power at some prespecified levels and then perform power analysis to determine the minimum sample size to achieve the desired levels of power. We can use the power function equation (3) for this purpose as well. For example, if we want to determine sample sizes n_1 and n_2 to achieve, say, 0.8 power, we can solve for n_1 and n_2 in the following equation:⁸

$$\text{power}(n_1, n_2, \alpha, H_0, H_1) \geq 0.8. \quad (4)$$

Note also that power functions can also be evaluated by replacing the mean difference δ and SD σ with the composite effect size d . Unlike δ , d is unit free and well interpreted, with $d=0.2, 0.5$ and 0.8 representing small, medium and large effect size.¹

POST HOC POWER ANALYSIS FOR COMPARING TWO POPULATION MEANS

In the preceding section, we discussed power analysis for comparing two population means for prospective studies. To evaluate power, we must specify the mean difference between the two population means, regardless of whether we want to estimate power for the given sample size or vice versa. Such a difference is study specific, which, although may be suggested by similar studies, should not be exclusively determined by one single study. This is because unlike the difference between population means δ , difference calculated based on a particular study sample δ is random and can be quite different from δ . As a result, power or sample size determined from sample-based mean difference δ can be quite different from the true power for prospective studies.

The difference between the population and sample-based parameters underscores the problem with post hoc power analysis. Not only is power analysis performed based on the sample-based mean difference, power estimates are also applied back to the same data to indicate power. Post hoc power analysis identifies population-level parameters with sample-specific statistics and makes no conceptual sense. Analytically, such analysis can yield quite different power estimates that are difficult and can be misleading.

To see this, consider again the problem to test the hypothesis in equation (1). Following the discussion in the preceding section, we can use equation (3) to computer power or use equation (4) to determine sample size for a desired level of power. When calculating power or sample size for post hoc analyses for a study with the outcomes already observed, the mean difference δ and SD σ will be set to their sample-based counterparts. Let \bar{X}_1 and \bar{X}_2 denote the sample means, and s denote the pooled sample SD from the study sample. By replacing δ with $\delta = \bar{X}_1 - \bar{X}_2$ and σ with s in equation (3), we obtain the power function for post hoc power analysis.

To help see the difference, we express the two power function side-by-side as follows:

$$\begin{aligned} \text{Prospective Power Analysis: } & \text{power}(n_1, n_2, \alpha, H_0, H_1) = \\ & P \left[\left| \frac{\bar{Y}_1 - \bar{Y}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| \geq Z_{\alpha/2} | \sigma, \delta \right], \\ \text{Post-hoc Power Analysis: } & \text{power}(n_1, n_2, \alpha, H_0, H_1) = \\ & P \left[\left| \frac{\bar{Y}_1 - \bar{Y}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| \geq Z_{\alpha/2} | S, \delta \right], \end{aligned} \quad (5)$$

The prospective power function is determined by the population mean difference δ and SD σ , but the post hoc power function depends on the sample mean difference δ and sample SD s . For large sample sizes, both δ and s will be close to their respective population counterparts, which is guaranteed by the law of large numbers. However, for relatively small sample sizes, δ , s or both can become quite different from their respective population parameters, in which case the two power functions in equation (5) can

yield very different values. Next, we use simulation studies to compare performance of the two power functions.

ILLUSTRATIONS

In this section, we use Monte Carlo simulation to compare the prospective and post hoc power functions. In all cases, we set a two-sided alpha at $\alpha=0.05$ and Monte Carlo sample size at 1000.

We again assume a normal distribution $N(\mu_k, \sigma^2)$ for the outcomes Y_{ik} , with μ_k denoting the (population) mean of group k and σ^2 the common (population) variance. We set the population-level parameters as follows:

$$\mu_1 = 0, \quad \mu_2 = \mu_1 + \delta, \quad \sigma = 1.$$

For convenience, we assume a common sample size for both groups, that is, $n_1 = n_2 = n$. We set δ and n to different values so we can see how the two power functions change for different effect size and sample size.

Given all these parameters, we can readily evaluate the prospective power function in equation (5). For post hoc power analysis, we simulate a sample from the normal distributions, compute the sample mean difference δ and sample SD s based on the simulated outcomes and evaluate the post hoc power function in equation (5). Unlike its prospective counterpart, this power function depends on the particular sample simulated. If δ and s are close to δ and σ , the two power functions will be close to each other. As indicated earlier, this will be the case for large sample sizes thanks to the law of large numbers. For relatively small samples, δ , s or both can be quite different from their population counterparts, in which case using the post hoc power function to informative power can be misleading.

With Monte Carlo simulation, we can readily examine the difference between the two power functions. By repeatedly simulating samples from the population distributions, we can look at the variability of the post hoc power function and see how it performs with respect to predicting true power.

Shown in figure 1 are the histograms of the post hoc power function based on 1000 Monte Carlo sample sizes with the mean difference $\delta=0.5$ (figure 1A), $\delta=1$ (figure 1B) and $\delta=2$ (figure 1C) and $n=50$, along with the power from the prospective power function (the vertical line segment). As expected, the true power increases as the mean difference δ becomes large. For $\delta=2$, true power is close to 1 and there is not much variability in the post hoc power function. For the other two values of δ , there is quite a large amount of variability in the post hoc power, covering the entire power function range between 0 and 1. The direction of skewness of the histogram changes from right skewed for $\delta=0.5$ to left skewed for $\delta=1$ to more left skewed at $\delta=2$. Unless the true power is at its upper bound 1, post hoc power is too variable to be informative for the true. For $\delta=2$, although there is not much variability, post hoc power is not really informative for any piratical purposes either, since the true power is close to 1.

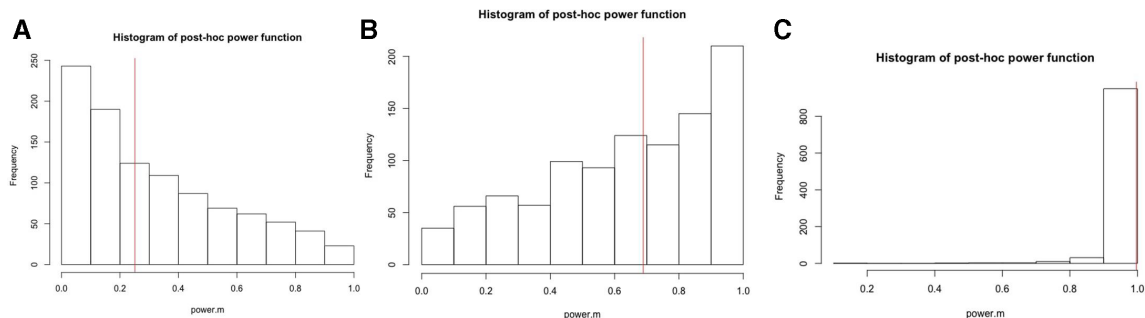


Figure 1 Histograms of post hoc power, along with true power, based on 1000 Monte Carlo sample sizes with the mean difference: (A) $\delta=0.5$; (B) $\delta=1$ and (C) $\delta=2$ and a sample size $n=50$.

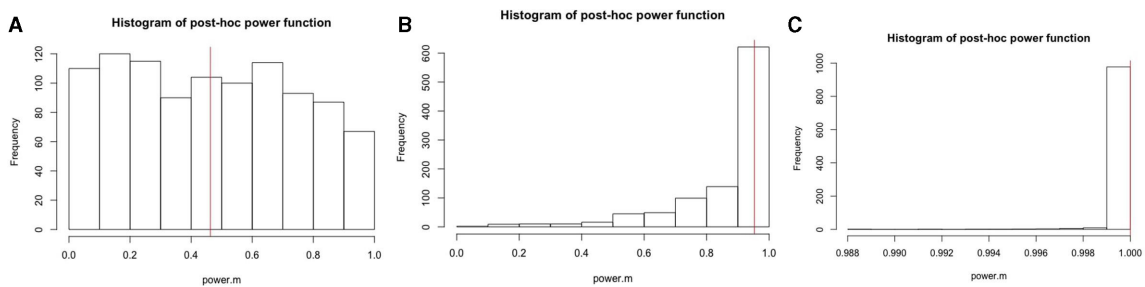


Figure 2 Histograms of post hoc power, along with true power, based on 1000 Monte Carlo sample sizes with the mean difference: (A) $\delta=0.5$, (B) $\delta=1$ and (C) $\delta=2$ and a sample size $n=100$.

Shown in [figure 2](#) are three histograms of the post hoc power function based on the same parameters, but with a large sample size $n=100$. Both the prospective and post hoc power functions show the same patterns as observed in [figure 1](#), as δ increases. Even with the true power close to 1 in the case of $\delta=1$, there is still variability in the post hoc function as shown in [figure 2B](#).

DISCUSSION

Power analysis is an indispensable component of planning clinical research studies. However, when used to indicate power for outcomes already observed, it is not only conceptually flawed but also analytically misleading. Our simulation results show that such power analyses do not indicate true power for detecting statistical significance, since post hoc power estimates are generally variable in the range of practical interest and can be very different from the true power.

In this report, we focus on the relatively simple statistical model for comparing two population means of continuous outcomes. The same considerations and conclusions also apply to non-continuous outcomes and more complex models such as regression. In general, post hoc power analyses do not provide sensible results.

Contributors All authors participated in the discussion of the statistical issues and worked together to develop this report. AR, RR, SR and XT discussed the problems with the justification of post hoc power analysis and interpretation of such power analysis results within the contexts of their studies and how to approach and clarify the issues in clinical research. YZ, RH and XT worked together to develop

the formulas for the power functions and the R codes, and performed simulation studies to understand the performance of post hoc power analysis. XT drafted the manuscript and all authors helped revise the paper.

Funding This study was supported by National Institutes of Health, Navy Bureau of Medicine and Surgery Grant UL1TR001442 of CTSA N1240.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Commissioned; internally peer reviewed.

Data availability statement No additional data available.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

REFERENCES

1. Cohen J. *Statistical power analysis for the behavioral sciences*. 2nd ed. Mahwah, NJ: LEA, 1988.
2. Heonig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician* 2001;55:19–24.
3. Kraemer HC, Mintz J, Noda A, et al. Caution regarding the use of pilot studies to guide power calculations for study proposals. *Arch Gen Psychiatry* 2006;63:484–9.
4. Levine M, Ensom MHH. Post hoc power analysis: an idea whose time has passed? *Pharmacotherapy* 2001;21:405–9.
5. Korn EL. Projection from previous studies. *Control Clin Trials* 1990;11:67–9.
6. Korn EL. Projecting power from a previous study: maximum likelihood estimation. *Am Stat* 1990;44:290–2.
7. Kowalski J, Tu XM. *Modern applied U statistics*. New York: Wiley, 2007.
8. Tu XM, Kowalski J, Zhang J, et al. Power analyses for longitudinal trials and other clustered designs. *Stat Med* 2004;23:2799–815.