Post-hoc power using observed estimate of effect size is too noisy to be useful[1]

Andrew Gelman[2]

25 Sept 2018

An article recently published in the *Annals of Surgery* stated: "as 80% power is difficult to achieve in surgical studies, we argue that the CONSORT and STROBE guidelines should be modified to include the disclosure of power—even if <80%—with the given sample size and effect size observed in that study" [1]. In [2], we argued that this idea, though a well-intentioned response to problematic statistical practices in medical research, would itself be a bad idea.

Two of the authors of the original paper responded, "We respectfully disagree that it is wrong to report post hoc power in the surgical literature. We fully understand that P value and post hoc power based on observed effect size are mathematically redundant; however, we would point out that being redundant is not the same as being incorrect. . . . We also respectfully disagree that knowing the power after the fact is not useful in surgical science." [3]

We again disagree. The problem with the authors' recommended post-hoc power calculations is not that they are "mathematically redundant" but rather that these calculations *will give inaccurate answers* because they are based on extremely noisy estimates of effect size. To put it in statistical terms, their recommended method has *poor frequency properties*.

We are in agreement that "knowing the power after the fact" would be useful, both in designing future studies and in interpreting existing results [4]. But the authors' recommended procedure of taking a noisy estimate and plugging it into a formula does *not* give us "the power"; it gives us a *very noisy estimate* of the power. Not the same thing at all.

Here's an example. Consider an experiment with 200 patients: 100 treated and 100 control, with post-operative survival of 94% for the treated group and 90% for the controls. Then the raw estimated treatment effect is 0.04 with standard error sqrt(0.94*0.06/100 + 0.90*0.10/100) = 0.04. The estimate is just one standard error away from zero, hence not statistically significant. And the crudely estimated post-hoc power, using the normal distribution, is approximately16% (the probability of observing an estimate at least 2 standard errors away from zero, conditional on the true parameter value being 1 standard error away from zero). But this estimate of power is very noisy! Consider that effect sizes consistent with these data could be anywhere from -0.04 to +0.12 (roughly), hence absolute effect sizes could be roughly between 0 and 3 standard

---

[1] To appear in *Annals of Surgery*. I thank Aleksi Reito for bringing references [1] and [3] to my attention.

[2] Department of Statistics, Columbia University, New York. gelman@stat.columbia.edu

errors away from zero, corresponding to power being somewhere between 5% (if the true population effect size happened to be zero) and 97.5% (if the true effect size were three standard errors from zero).

This is the problem with using the raw estimate results from a single, noisy study to estimate power: the point estimate tells us almost nothing, and considering this noisy estimate as "knowing the power after the fact" is an invitation to overconfidence.

Instead we recommend performing design and power analysis using substantively-based effect size estimates [2]. We recognize that this may not be easy. As Bababekov and Chang write in their letter, "it would be difficult to adapt previously reported effect sizes to comparative research involving a surgical innovation that has never been tested." [3]

Fair enough. Our approach requires assumptions. But that's the way it works: if you want to make a statement about power of a study, you need to make some assumption about effect size. Make your assumption clearly, and go from there. Bababekov and Chang write: "As such, if we want to encourage the reporting of power, then we are obliged to use observed effect size in a post hoc fashion." No, no, and no. Researchers are not obliged to use a super-noisy estimate. Researchers are allowed to use scientific judgment when performing power analysis when designing a study, and they are allowed to use scientific judgment when doing design analysis, after doing the study.

I appreciate the authors' general goals in [1] and [3]; there just happens to be a technical problem by which the natural-seeming estimate of power using the point estimate from the study is too noisy to be useful. This is a fundamental problem of limited information which can be resolved only be using external knowledge from the literature, making clear assumptions, or gathering new data.

References

[1] Bababekov, Y. J., Stapleton, S. M., Mueller, J. L., Fong, Z. V., and Chang, D. C. (2018). A proposal to mitigate the consequences of type 2 error in surgical science. *Annals of Surgery* 267, 621-622.

[2] Gelman, A. (2018). Don't calculate post-hoc power using observed estimate of effect size. *Annals of Surgery*.

[3] Bababekov, Y. J., and Chang, D. C. (2018). Post hoc power: A surgeon's first assistant in Interpreting "negative" studies. *Annals of Surgery*.

[4] Gelman, A., and Carlin, J. B. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science* 9, 641-651.