Don't calculate post-hoc power using observed estimate of effect size[1]

Andrew Gelman[2]

5 Apr 2018

An article recently published in the *Annals of Surgery* states: "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].

This would be a bad idea. The problem is that the (estimated) effect size observed in a study is noisy, especially so in the sorts of studies discussed by the authors. Using estimated effect size can give a terrible estimate of power, and in many cases can lead to drastic overestimates of power (thus, extreme overconfidence of the sort that is rightly deplored in the full article [1]), with the problem becoming even worse for studies that happen to achieve statistical significance.

The problem is well known in the statistical and medical literatures; see, e.g, [2, 3]. For some discussion of the systemic consequences of biased power calculations based on noisy estimates of effect size, see [4], and for an alternative approach to design and power analysis, see [5].

That said, I agree with much of the content of [1]. I agree that the routine assumption of 80% power is a mistake, and that requirements of 80% power encourage researchers to exaggerate effect sizes in their experimental designs, to cheat in their analyses in order to attain the statistical significance that they was supposedly so nearly being assured [6]. More generally, demands for near-certainty, along with the availability of statistical analysis tools that can yield statistical significance even in the absence of real effects, have led to replication crisis and general corruption in many areas of science [7], a problem which I believe is structural and persists even in the presence of honest intentions of many or most participants in the process [8].

I appreciate the concerns of [1] and I agree with their goals and general recommendations, including their conclusion that "we need to begin to convey the uncertainty associated with our studies so that patients and providers can be empowered to make appropriate decisions." There is just a problem with their recommendation to calculate power using observed effect sizes.

---

[1] To be published in *Annals of Surgery*. I thank Aleksi Reito for bringing reference [1] to my attention.

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

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] Goodman, S. N., and Berlin, J. A. (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine* 121, 200-206.

[2] Senn, S. J. (2002). Power is indeed irrelevant in interpreting completed studies. *British Medical Journal* 325, Article 1304.

[4] Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B., Flint, J., Robinson, E. S. J., and Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14, 1-12.

[5] 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.

[6] Gelman, A. (2017). The "80% power" lie. Statistical Modeling, Causal Inference, and Social Science blog, 4 Dec. http://andrewgelman.com/2017/12/04/80-power-lie/

[7] Ioannidis, J. (2016). Evidence-based medicine has been hijacked: a report to David Sackett. *Journal of Clinical Epidemiology* 73, 82-86.

[8] Gelman, A. (2017). Honesty and transparency are not enough. *Chance* 30 (1), 37-39.