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

Andrew Gelman[2]

28 Mar 2018

In an article recently published in the *Annals of Surgery*, Bababekov et al. (2018) write: "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."

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 by Bababekov et al. in their article), 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., Lane and Dunlap (1978), Hedges (1984), Goodman and Berlin (1994), Senn (2002), and Lenth (2007). For some discussion of the systemic consequences of biased power calculations based on noisy estimates of effect size, see Button et al. (2013), and for an alternative approach to design and power analysis, see Gelman and Carlin (2014).

That said, I agree with much of what Bababekov et al. (2018) say. 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 (Gelman, 2017b). 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 (Simmons et al., 2011), have led to replication crisis and general corruption in many areas of science (Ioannidis, 2016), a problem which I believe is structural and persists even in the presence of honest intentions of many or most participants in the process (Gelman, 2017a).

I appreciate the concerns of Bababekov et al. (2018) 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 a just a problem with their recommendation to calculate power using observed effect sizes.

---

[1] I thank Aleksi Reito for bringing this article to my attention.

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

References

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.

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.

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

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

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.

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.

Hedges, L. V. (1984). Estimation of effect size under non- random sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics* 9, 61-85.

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

Lane, D. M., and Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology* 31, 107-112.

Lenth, R. V. (2007). Statistical power calculations. *Journal of Animal Science* 85, E24-E29.

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

Simmons, J., Nelson, L., and Simonsohn, U. (2011). False- positive psychology: Undisclosed flexibility in data collection and analysis allow presenting anything as significant. *Psychological Science* 22, 1359–-366.