Statistics and the crisis of scientific replication*

Andrew Gelman†

8 Dec 2014

Top journals in psychology routinely publish ridiculous, scientifically implausible claims, justified based on "$p < 0.05$." Recent examples of such silliness include claimed evidence of extra-sensory perception (published in the *Journal of Personality and Social Psychology*), claims that women at certain stages of their menstrual cycle were three times more likely to wear red or pink clothing and 20 percentage points more likely to vote for the Democratic or Republican candidate for president (published in *Psychological Science*), and a claim that people react much differently to hurricanes with male and female names (published in the *Proceedings of the National Academy of Sciences*). All these studies had serious flaws, to the extent that I (and others) found claims to be completely unconvincing from a statistical standpoint, matching their general implausibility on substantive grounds.

It is easy to dismiss these particular studies, one at a time. But, to the extent that they are being conducted using standard statistical methods, this calls into question all sorts of more plausible, but not necessarily true, claims, that are supported by this same sort of evidence. To put it another way: we can all laugh at studies of ESP, or ovulation and voting, but what about MRI studies of political attitudes, or embodied cognition, or stereotype threat, or, for that matter, the latest potential cancer cure? If we can't trust p-values, does experimental science involving human variation just have to start over? And what to we do in fields such as political science and economics, where preregistered replication can be difficult or impossible? Can Bayesian inference supply a solution? Maybe. These are not easy problems, but they're important problems.

Figure 1 demonstrates what can happen with classical hypothesis testing. A study is performed in which the underlying parameter of interest (typically a causal effect or some other sort of comparison in the general population) is relatively small, and measurements are noisy and biased (not uncommon in a psychology setting in which the underlying constructs are often not clearly defined). In this setting of small effects and large errors, any observed differences that happen to be statistically significant will have a high chance of being in the wrong direction and can drastically overestimate the magnitude of the effect (Gelman and Carlin, 2014).

When applied to the scientific process more generally, the result is a flow of noisy claims which bear only a weak relation to reality, but which attain statistical significance, which is, conventionally, a necessary and sufficient condition for publication, if said result is paired with any story that is qualitatively justified by a substantive theory.

Various researchers in psychology and medicine have made the following linked points: statistical significance cannot generally be taken at face value (Simmons, Nelson, and Simonsohn, 2011); a scientific publication system based on null hypothesis significance tests leads to large-scale errors in reporting; and these problems are particularly severe in the context of low signal and high noise (Button et al., 2013).

Psychology is particularly subject to such problems, for several reasons:

- The objects of study (mental states, personality traits, cognitive and social abilities) are inherently latent and can typically not be precisely defined;
- Theories are correspondingly vague (in comparison with physics or chemistry, say, or even

*For *Significance* magazine.

†Department of Statistics and Department of Political Science, Columbia University, New York

medicine) in that the magnitude and even the direction of effects cannot always be predicted based on theoretical grounds;

- Variation between people is typically large, as is variation across repeated measurements within people; indeed, analysis of this variation is often a central research goal;

- The stakes are low so it is easy to quickly do a small study and write up the conclusions. Unlike in medical research, there is no hurdle to performing a publishable study.

The resulting proliferation of studies with small effect sizes and high noise, along with a willingness of high-profile, prestigious journals such as *Psychological Science* and the *Proceedings of the National Academy of Sciences* to publish surprising, newsworthy findings based on statistically-significant comparisons, has led us to a crisis in scientific replication.

There has been much discussion in psychology and other sciences about the pressure to publish and the pressure to replicate (an issue I have elsewhere; see Gelman, 2014b), but here I want to set aside issues of social interaction and focus on the statistical questions.

In particular: can Bayesian inference solve the replication crisis?

My quick answer: Maybe, but not in any immediate way. Two quick Bayesian alternatives to current statistical practice are to replace p-values by Bayes factors, or to replace hypothesis tests by uncertainty intervals. These approaches may have some advantages but if they are tied to accept/reject (or publish/no-publish) decisions, they are still fundamentally based on null-hypothesis significance testing and so would have the problem that, whatever results happen to reach the hurdle of "strong evidence" (or whatever other threshold would play the role of "statistical significance" in publication decisions) could still be taken as true, and thus would be subject to the same issues of Type S errors and exaggeration factors that are illustrated in Figure 1.

I do see a potential Bayesian solution using informative priors and models of varying treatment effects (Gelman, 2014a), but these steps will not be easy because they move away from the usual statistical paradigm in which each scientific study stands alone. To resolve the replication crisis in science we may need to consider each individual study in the context of an implicit meta-analysis.

References

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. (2014a). The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management*.

Gelman, A. (2014b). Replication controversies. Statistical Modeling, Causal Inference, and Social Science blog, 19 Nov. `http://andrewgelman.com/2014/11/19/24265/`

Gelman, A., and Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science* **9**, 641–651.

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—1366.

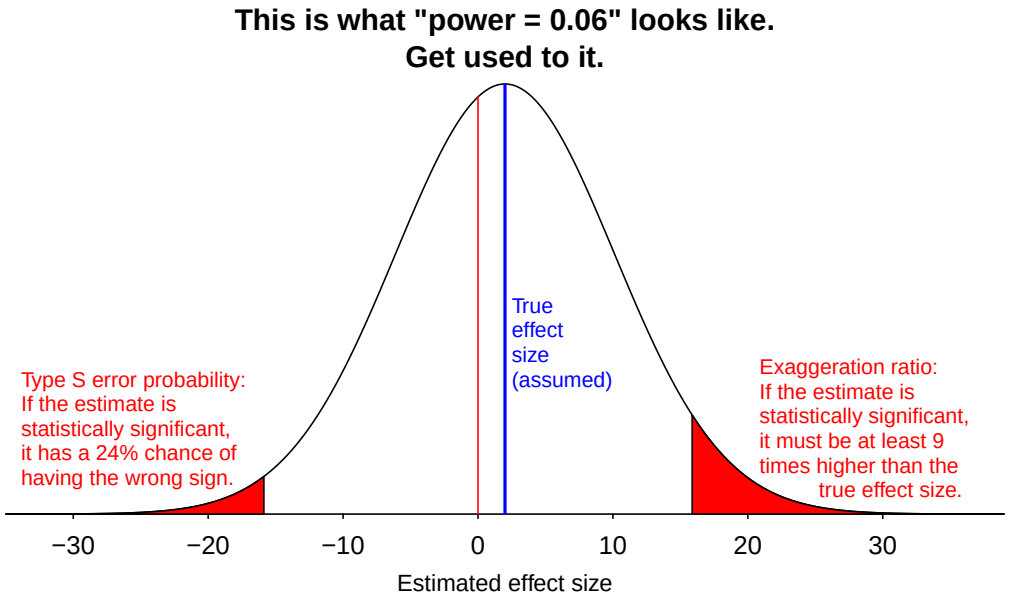

Figure 1: *A study has low power when the population difference or effect size is small, while variation and measurement are small. Such conditions are common (but not limited to) psychology, where theories and hypothesized effects can be speculative, variability between individuals is high, and measurements can be noisy and biased. In low-power studies, the "Type S (sign) error rate"—the probability that the observed difference is in the opposite direction of the true effect or population difference—can be high, even if the estimate is statistically significant. And the "exaggeration ratio"—the factor by which the observed estimate exceeds the true parameter value being estimated—can be huge.*