Statistics and Research Integrity[1]

Andrew Gelman[2]

9 Jan 2015

How does statistics fit into questions of research integrity? Traditionally we think of statistics as a check on rash hypothesizing. Or, to put it another way, the p-value is the strength of information against a null hypothesis. And, from this perspective, the role of statistics in research integrity is to protect the integrity of the p-value, by always supplying a standard error for every estimate and a p-value for every positive statement, reporting all analyses that are done, and correcting for sequential testing and multiple comparisons as necessary.

In recent years, however, it has become clear that this classical approach to statistical significance is not enough and that, instead, we must move from a passive to an active approach to ensure research integrity.

The problem is that the conventional system of science publishing is breaking down (or perhaps it was always broken and we did not realize it). I'm not speaking so much of outright fraud or junk journals that will publish anything if you give them the cash, but rather of mainstream researchers and journals, even top outlets such as Science, Nature, Psychological Science, or the Proceedings of the National Academy of Sciences.

The problem comes in two steps, which we can call the *statistical significance filter* and the *publication filter*. The statistical significance filter corresponds to the assumption that when an estimate reaches the "$p < 0.05$" level, that it is treated as true. Or, to take a sampling perspective, that if a comparison in the observed data reaches the $p < .05$ level, that it corresponds to a true pattern in the population. Such an inference is often clearly inappropriate: nearly every data analysis ever published is performed contingent on the data (what is sometimes called "p-hacking" and sometimes called "the garden of forking paths" [1]) and so, even if there is nothing going on, a researcher can have a much greater than 5% chance of getting a "statistically significant, $p<0.05$" result.

In short, statistical significance is not hard to come by, which is one reason why our science communication channels are polluted by published and publicized results such as a claim that female-named hurricanes are more deadly than male-named hurricanes (Proceedings of the National Academy of Sciences) or that women in different phases of their monthly cycle were 20 percentage points more likely to vote for Barack Obama for President (this one appeared in Psychological Science), not to mention the cancer cures and food scares that seem to appear in the newspaper on a regular basis, invariably backed up by publications in top journals.

The statistical significance filter is not just about the ease of obtaining statistical significance. It's also a bias: if an estimate is statistically significant and is measured in a context of high variation (as is usual in the human sciences), it will necessarily be large (more than 2 standard errors from zero, to be precise) and thus will tend to overestimate any true population effect. The smaller estimates are filtered out and only the large ones survive; this is a form of selection bias. So statistical significance can be found from noise, and, if there is an

---

[1] For *European Science Editing,* the journal of the European Association of Scientific Editors.

[2] Department of Statistics and Department of Political Science, Columbia University, New York.

underlying effect, the published estimate is likely to be too high, perhaps by orders of magnitude [2].

The second problem is the publication filter, the attitude that, if something is published, it should have the presumption of being true. This is a big deal. Recently there have been many high-profile cases in psychology research in which published studies have been questioned, sometimes based on first principles and other times based on unsuccessful replications, and the authors of the original studies have responded angrily and defensively. Rather than say, "Hmmm, maybe we made a mistake, maybe our analysis capitalized on chance patterns in our sample, we should think twice about our conclusions," authors commonly respond with defenses of their p-values and attacks on the replications. Defenses can be appropriate, of course—I don't want to imply that questioners of published work are always correct—but I do want to push back against the idea that empirical work that is published deserves some deference.

One superficially appealing argument in favor of the publication filter is that publication, especially in top journals, is difficult: you typically have to get past three or more referees along with a skeptical associate editor. This is fine, but experience tells us that the papers that *do* get through can have serious flaws. This is one reason there is interest in post-publication review, and in more open review. Those three referees put a lot of work into their reports; why not share them with the world? Then any notes of caution can be seen by others. And, if an important problem was *not* noticed by any reviewers, this will be clear as well.

What can be done?

I have addressed two concerns: statistical significance and publication, both of which play valuable roles in screening but which can mislead when they are taken as badges of correctness. Various reforms have been suggested. Statistical significance is close to meaningless in typical research, which is so open-ended that it is essentially impossible to identify what analysis would have been performed had the data been different. One way to make p-values work is to perform *pre-registered replications*, in which all the details of data collection, processing, and analysis are decided ahead of time. Pre-registered replication could work well in fields such as experimental psychology and biology where replications are easy to do, could be more difficult in medicine, and is close to impossible in much of social science. I do think that preregistered replication is a useful ideal, and it is interesting to see some of the opposition to it, which seems to be driven somewhat by fear that prominent results will fail to replicate. Where preregistered replication is not possible, I think we need to move away from p-values and instead perform analyses that perform all possible comparisons of interest. My preferred method here is Bayesian multilevel modeling [3] but other approaches are possible.

When it comes to publication, there have been many reforms proposed, including, as noted above, open review and post-publication review. When considering reforms we should just keep in mind the larger goals of scientific research as a collaborative process, and science communication. Once we accept that certainty is hard to come by in the human sciences, we should be more able to recognize the value in the publication and dissemination of research findings, and to reduce the incentive for sloppy, flashy work.

Here I have discussed issues of statistics and research integrity in fairly general terms. For some specific examples, you can read the references, which point to a recent literature on

research quality in the medical and social sciences. The present article makes no attempt at comprehensiveness; rather, I am raising some issues that are discussed in more detail by myself and others in other places.

Let me conclude by emphasizing that, when I say that the statistical significance filter and the publication filter represent threats to research integrity, this goes beyond concerns about the integrity of individual researchers. Every scientist involved in these disputes could operate at the highest level of personal honesty and integrity, and these problems could still arise. Indeed, some of our difficulties may well arise from the confusion of individual integrity with integrity of the system. Researchers, who themselves have no desire to cheat, can react angrily to skepticism about their p-values and publications. But this is where statistics comes in. The statistical issues are tricky, and unfortunately is it all too possible to follow standard practices and end up in a dead end, producing statistically significant p-values and getting published while discovering nothing. Science, and science communication, are harmed—it's a loss of integrity—even in the absence of any unethical behavior.

References

1. Gelman A, Loken E. The Statistical Crisis in Science. American Scientist 2014; 102:460-5. DOI: 10.1511/2014.111.460. Avaliable at: http://www.stat.columbia.edu/~gelman/research/published/ForkingPaths.pdf

2. Gelman A, Carlin J. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. Perspectives on Psychological Science 2014; 9(6):641-51. DOI: 10.1177/1745691614551642. Avaliable at: http://www.stat.columbia.edu/~gelman/research/published/retropower_final.pdf

3. Gelman A. The Connection Between Varying Treatment Effects and the Crisis of Unreplicable Research: A Bayesian Perspective. Journal of Management 2014; DOI: 10.1177/0149206314525208. Avaliable at: http://www.stat.columbia.edu/~gelman/research/published/bayes_management.pdf