Journal of Mathematical Psychology 57 (2013) 188–189

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Commentary

Interrogating *p*-values☆

Andrew Gelman

Columbia University, Statistics, New York, United States

ARTICLE INFO

Article history:
Available online 19 April 2013

Keywords:
p-values
Statistics
Replicability

ABSTRACT

This article is a discussion of a paper by Greg Francis for a special issue edited by E.J. Wagenmakers.

Much of statistical practice is an effort to reduce or deny variation and uncertainty. The reduction is done through standardization, replication, and other practices of experimental design, with the idea being to isolate and stabilize the quantity being estimated and then average over many cases. Even so, however, uncertainty persists, and statistical hypothesis testing is in many ways an endeavor to deny this, by reporting binary accept/reject decisions.

Classical statistical methods produce binary statements, but there is no reason to assume that the world works that way. Expressions such as Type 1 error, Type 2 error, false positive, and so on, are based on a model in which the world is divided into real and non-real effects. To put it another way, I understand the general scientific distinction of real vs. non-real effects but I do not think this maps well into the mathematical distinction of $\theta = 0$ vs. $\theta \neq 0$. Yes, there are some unambiguously true effects and some that are arguably zero, but I would guess that the challenge in most current research in psychology is not that effects are zero but that they vary from person to person and in different contexts.

But if we do not want to characterize science as the search for true positives, how *should* we statistically model the process of scientific publication and discovery? An empirical approach is to identify scientific truth with *replicability*; hence, the goal of an experimental or observational scientist is to discover effects that replicate in future studies.

The replicability standard seems to be reasonable. Unfortunately, as Francis (2013) and Simmons, Nelson, and Simonsohn (2011) have pointed out, researchers in psychology (and, presumably, in other fields as well) seem to have no problem replicating and getting statistical significance, over and over again, even in the absence of any real effects of the size claimed by the researchers.

The problem is that the purported replications are not pure replications: they are new studies, not quite what came before and each with its own choice of data to study and analyses to pursue.

As a student many years ago, I heard about opportunistic stopping rules, the file drawer problem, and other reasons why nominal *p*-values do not actually represent the true probability that observed data are more extreme than what would be expected by chance. My impression was that these problems represented a minor adjustment and not a major reappraisal of the scientific process. After all, given what we know about scientists' desire to communicate their efforts, it was hard to imagine that there were file drawers bulging with unpublished results.

More recently, though, there has been a growing sense that psychology, biomedicine, and other fields are being overwhelmed with errors (consider, for example, the generally positive reaction to the paper of Ioannidis, 2005). In two recent series of papers, Gregory Francis and Uri Simonsohn and collaborators have demonstrated too-good-to-be-true patterns of *p*-values in published papers, indicating that these results should not be taken at face value. And this is happening not just in areas such as ESP studies which are generally considered pathological science (in the sense of Langmuir, 1953) but also in subfields of psychology where there is essentially no doubt that many of the effects being studied are real (notwithstanding rhetorical deconstructions such as that of Mitchell, 2000, 2008; see Borsboom & Mellenbergh, 2004).

All this is in addition to the well-known difficulties of interpretation of *p*-values (e.g., Krantz (1999), Gelman (2012)), and to the problem that, even when all comparisons have been openly reported and thus *p*-values are mathematically correct, the "statistical significance filter" ensures that estimated effects will be in general larger than true effects, with this discrepancy being well over an order of magnitude in settings where the true effects are small (a scenario that we would expect to see often, given the oft-noted problem that the low-hanging fruit of science have already been picked), as discussed by Gelman and Weakliem (2009).

☆ For *Journal of Mathematical Psychology*, in response to the article, "Replication, statistical consistency, and publication bias", by Gregory Francis. We thank E. J. Wagenmakers for organizing the discussion, Uri Simonsohn for helpful comments, and the Institute for Education Sciences and the National Science Foundation for partial support of this work.

E-mail address: gelman@stat.columbia.edu.

0022-2496/$ – see front matter
http://dx.doi.org/10.1016/j.jmp.2013.03.005

For all these reasons, I applaud the work of Gregory Francis in pointing out incoherence in sets of published *p*-values. At the same time, I do not know exactly what to do with the *p*-values that he himself reports. I am not so worried about Francis's own file drawers—I take his results as summaries of the papers that he reports on, just as, more generally, I take *p*-values as data summaries and do not see them, on their own, as guides to decisions, whether or not they have been purposefully selected from a larger population of tests. In these examples, though, the evidence from Francis's *p*-value analysis seems to be weak compared to other available information (as in the notorious Bem paper, where Francis's test gives equivocal results, while information internal to the paper makes it clear that Bem chose among many specifications before reporting his results). Although I do not know how useful Francis's particular method is, overall I am supportive of his work as it draws attention to a serious problem in published research.

Finally, this is not the main point of the present discussion but I think that my anti-hypothesis-testing stance is stronger than that of Francis (2013). I disagree with the following statement from that article:

> For both confirmatory and exploratory research, a hypothesis test is appropriate if the outcome drives a specific course of action. Hypothesis tests provide a way to make a decision based on data, and such decisions are useful for choosing an action. If a doctor has to determine whether to treat a patient with drugs or surgery, a hypothesis test might provide useful information to guide the action. Likewise, if an interface designer has to decide whether to replace a blue notification light with a green notification light in a cockpit, a hypothesis test can provide guidance on whether an observed difference in reaction time is different from chance and thereby influence the designer's choice.

I have no expertise on drugs, surgery, or human factors design and so cannot address these particular examples—but, speaking in general terms, I think Francis is getting things backward here. When making a decision, I think it is necessary to consider effect sizes (not merely the possible existence of a nonzero effect) as well as costs. Here I speak not of the cost of hypothetical false positives or false negatives but of the direct costs and benefits of the decision. An observed difference can be relevant to a decision whether or not that difference is statistically significant.

References

Borsboom, D., & Mellenbergh, G. J. (2004). Why psychometrics is not pathological: a comment on Mitchell. *Theory and Psychology*, *14*, 105–120.

Francis, G. (2013). Replication, statistical consistency, and publication bias (with discussion). *Journal of Mathematical Psychology*, *57*(5), 153–169.

Gelman, A. (2012). *P*-values and statistical practice. *Epidemiology*,.

Gelman, A., & Weakliem, D. (2009). Of beauty, sex, and power: statistical challenges in estimating small effects. *American Scientist*, *97*, 310–316.

Ioannidis, J. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), e124.

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association*, *44*, 1372–1381.

Langmuir, I. (1989). Pathological science. Transcribed and edited by R.N. Hall. *Physics Today*, *42*(10), 36–48.

Mitchell, J. (2000). Normal science, pathological science and psychometrics. *Theory and Psychology*, *10*, 639–667.

Mitchell, J. (2008). Is psychometrics pathological science? *Measurement: Interdisciplinary Research and Perspectives*, *6*, 7–24.

Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allow presenting anything as significant. *Psychological Science*, *22*, 1359–1366.