Life After P-Hacking

Joseph Simmons, University of Pennsylvania, USA
Leif D. Nelson, University of California Berkeley, USA
Uri Simonsohn, University of Pennsylvania USA

We discuss how our commitment to publish replicable results will affect our research lives. We must (1) dramatically increase our sample sizes, (2) follow-up exploratory analyses with confirmatory replications, and, because making replicable discoveries requires significant resources, (3) judge researchers by their best publications rather than by their publication quantity.

Life After P-Hacking

Joseph Simmons, University of Pennsylvania, USA
Leif D. Nelson, University of California, Berkeley, USA
Uri Simonsohn, University of Pennsylvania, USA

EXTENDED ABSTRACT

This paper considers how a commitment to not publishing p-hacked results will change the lives of individual scientists and products of science. We call this "Life After P-Hacking." We discuss four lessons.

Lesson #1

We have to really start caring about statistical power. The freedom to engage in p-hacking created an environment in which researchers were able to get many of their studies to yield statistically significant results despite dramatically underpowering those studies (or studying truly null effects). "Life after p-hacking" means that conducting underpowered studies will now come with significant costs to the individual researcher: (1) by definition, many underpowered studies will yield null effects, and (2) researchers will not know why they did not work – whether the null effects represent a false hypothesis (a truly null effect) or a false-negative. This makes it difficult to learn and difficult to publish, a combination of costs that is unlikely to be sustainable for the typical researcher. The only way forward, then, is to really start caring about statistical power – to make sure our studies are properly powered at, say, 80%.

Of course, ideally we would conduct power analyses in order to properly power our studies, but power analyses require researchers to know what effect sizes they are studying. When effect sizes are unknown, researchers are unlikely to have good intuitions about the size of the effect they are studying, for publication bias and p-hacking means that the literature dramatically overestimates the sizes of most effects. In order to give researchers better intuitions about effect size – and power – we conducted a large MTurk survey (N = 696) that tested a number of "obvious" two-condition between-subjects hypotheses (e.g., men are taller than women). Based on the (unbiased) effect sizes obtained, we then computed how many participants a researcher would need to have in each condition in order to have achieve 80% power. Although some very obvious effects require the kinds of small samples we see in our literature – for example, you need "only" 15 per cell (30 total) to detect that women report owning more pairs of shoes than do men – other obvious effects require much larger samples than we are used to. For example, to detect that men report weighing more than women, a researcher needs 46 participants per tell (94 total). Given that most of the effects that we study are likely to be smaller than "men weigh more than women," we advise researchers investigating unknown effect sizes to have at least 50 participants per cell. Otherwise, their studies are likely to be so underpowered as to prevent them from learning or publishing anything. (We will emphasize that this n>50 heuristic is merely a heuristic, and should not substitute for conducting power calculations when effect sizes are estimable).

Lesson #2

At the exploratory stage, p-hacking is advisable as it helps us learn from the data; but we should replicate any p-hacked result. Some researchers have misinterpreted our crusade against p-hacking as a suggestion that we should never p-hack, no matter what. This is wrong. Conducting lots of unplanned, exploratory analyses on our data is a fruitful way to learn from data. We will describe a real example of this – in which a researcher collects three measures and finds the predicted result for only one of the three (and even then, only after removing outliers). The two measures that did not work were obviously "bad" in retrospect, and the removal of outliers for the third measure is easy to justify. Thus, this is a case in which exploratory p-hacking may have helped identify a result that was not obviously there after the first, planned analysis was conducted. Of course, at this stage one cannot know whether the significant result is due to capitalizing on chance or whether it is real. The only way to tell is to conduct an exact replication, using only the measure that worked and committing in advance to the same outlier rule that worked. Thus, this is what "life after p-hacking" will often involve – p-hack during an exploratory phase but then attempt to directly replicate any result obtained during that phase.

Lesson #3

Fewer papers. A science that requires larger samples (for statistical power) and more direct replications of one's own work (to test whether a p-hacked result is reliable or not) is a science that takes longer. Thus a natural, inevitable consequence of "life after p-hacking" is that we will write fewer papers. This means that our field needs to lower the threshold for what is considered an adequate quantity of published papers, and that published papers should be judged on their merits rather than as lines on one's cv.

Lesson #4

Label your research as not p-hacked. Research that is not p-hacked is more likely to be true and is of therefore higher quality. Until journals start requiring disclosure of all sample size rules, measures, and manipulations upon submission of articles, a reader and reviewer cannot know whether a paper is p-hacked or not. Those who are not p-hacking will benefit from making it know that they are not p-hacking. Thus, we advocate that researchers include the following 21 words in their methods sections: "We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study."