P-Hacking: an Embedded Phenomenon in Academic Research

Héctor Andrés Plata Santos

Abstract

P-Hacking is a term coined for Post Hoc data manipulation. This term implies that a researcher or a group of researchers interfered willingly or not at some point in the data pipeline of a research with the side effect that the resulting p-values are going to be biased because they are not going to take into account the effect of multiple comparisons. In this papers I explore broadly on the introduction, some recent ideas of how p-hacking manifest itself and how hard is to come to a consensus on how to understand it. Finally on the discussion and conclusion section I state why, the only way to overcome this phenomena is to change how we value research.

Introduction

The scientific method is usually considered a reliable method for knowledge discovery. However as we will see there are some incentives on the research cycle that affects the credibility of published studies. One of the main tools used by researchers when they want to confirm that an effect size is different than zero, or in other words, that an effect is statistically significant, is the use of the widely popular p-value. This value corresponds to the probability that the effect under the null hypothesis is just observed by random chance. Researchers usually fix a significance level α at which if the p-value is lower than that they declare that there is evidence to believe that the effect exists.

The powerful idea of the p-value brings with it two main biases into the research cycle (Simonsohn, Nelson, and Simmons 2014). The first one is called "publication bias" or "the file drawer problem". In short, the problem is about the tendency of journals or reviewers to publish only research that contains statistically significant results and not those with inconclusive or non significant results. This behavior and the fact that a researcher performance is measured by a weighted average of the number of papers published means that there is an incentive for them to submit papers that have positive results.

This comes with the second bias on the research cycle and on which this paper is going to focus on, called "p-hacking". Bishop and Thompson (2016) describes the phenomenon as follows:

> ... the practice of reporting only that part of a dataset that yields significant results, making the decision about which part to publish after scrutinizing the data. There are various ways in which this can be done: e.g., deciding which outliers to exclude, when to stop collecting data, or whether to include covariates...

The practice of p-hacking induces the researcher into the multiple comparisons problem, in which a stricter significance level α (like in the Bonferroni correction) is needed in order to make up for the several inferences that are being done at the same time. Unfortunately since there is an incentive to present statistical significant results, the researcher will be prone to avoid correcting the significance level so that it obtains a positive result, giving him / her a better probability of getting published.

However, as Gelman and Eric stated on their paper "The garden of forking paths" (2013), researchers can fall into the trap of p-hacking unwillingly as well. This usually happens when the analysis done is highly contingent on the data. Meaning that there is not a clear set of instructions on how to conduct the analysis previously of the data collection. This lack of a procedure to follow before doing the data analysis is what leads researchers into a multiple comparison problem, since this will incur into a one to many mapping from a scientific question to many statistical hypothesis. This consequence that arises from having an analysis contingent on the data is usually ignored (Gelman and Loken 2013) and thus, not being corrected by the researcher.

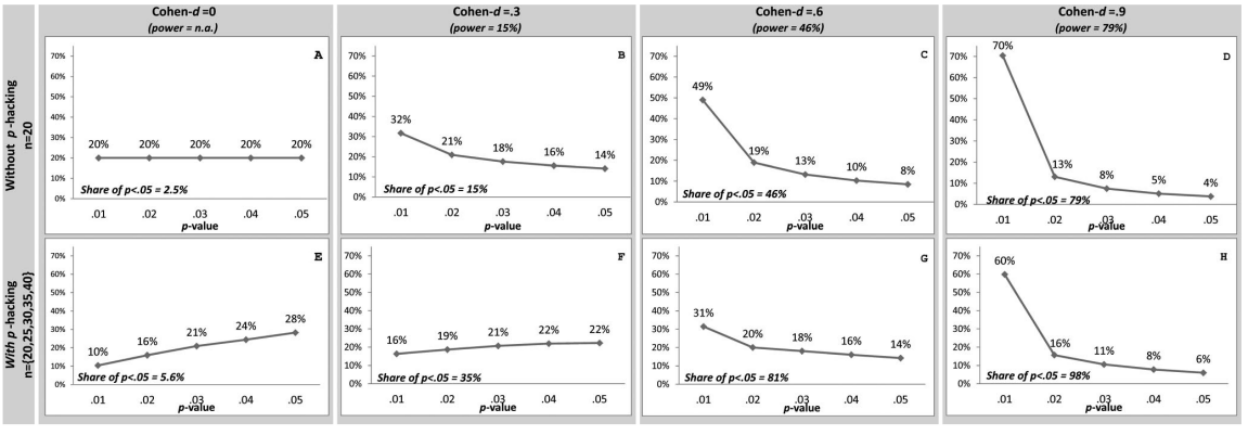

Figure 1: Taken from (Simonsohn, Nelson, and Simmons 2014). P-curves for different true effect sizes in the presence and absence of p-hacking. Graphs depict expected p-curves for difference-of-means t tests for samples from populations with means differing by d standard deviations. A–D: These graphs are products of the central and noncentral t distribution. E–H: These graphs are products of 400,000 simulations of two samples with 20 normally distributed observations. For 1E–1H, if the difference was not significant, five additional, independent obser- vations were added to each sample, up to a maximum of 40 observations. Share of $p < .05$ indicates the share of all studies producing a statistically significant effect using a two-tailed test for a directional prediction (hence 2.5% under the null).

One suggestion that the authors of "The garden of forking paths" (Gelman and Loken 2013) makes, is doing a preregistration of the procedure that is going to follow the research. Nonetheless they also acknowledged that there are some fields where doing analysis contingent to the data is necessary to understand the phenomenons that occurs within the experiment. In those fields, the research or paper should be declared as exploratory or should be followed up with a replication one, where the data analysis and recollection procedure is defined beforehand. Unfortunately, it's difficult to do that in fields where getting new data is expensive or impossible like in economics or in events that only happens once, like an economic crisis or a natural disaster. That's why p-values and results from research should always be analysed with prudence before making any inference about it.

All of the resulting discussion about p-hacking in the scientific community, led to the creation of tools or meta-analysis that allow to incorporate results from different studies of the same topic in order to extract some common knowledge across research. In this case, a methodology called the "P-curve" was defined by Simonsohn, Nelson and Simmons on their paper "P-curve: A Key to the File-Drawer" (2014) that brings some guidance where to suspect that p-hacking has been used to produce false positive results.

The p-curve is no more than a plot of the distribution of the p-values of selected research articles on the same topic. This plot is going to show whether there was some p-hacking involved on the research made. Under the null hypothesis the distribution of the p-value is uniform (Simonsohn, Nelson, and Simmons 2014). Which means that if there is no p-hacking present, one should be able to see a uniform distribution out of the p-curve and suspect that there is no evidential value of the effect size tested to be different than zero. In the case of active p-hacking (a false positive results), the authors of the paper make the assumption that once the significance level is achieved, the Post Hoc selection and data manipulation is going to stop. Thus, giving the p-curve a negative skewness. On the other case, were there is evidential value, the p-curve is going to have a positive skewness and it's going to be magnified by the power of the test. This can be seen on figure 1.

Simonsohn, Nelson and Simmons (2014) also proposes two ways to test if a p-curve is right-skewed. The first one is as follows:

> "A simple method consists of dichotomizing p values as high ($p > .025$) versus low ($p < .025$) and submitting the uniform null (50% high) to a binomial test."

However, since this doesn't account for the variance between the two bins, the method is inefficient. That's why they propose a second method, which is a little more complex and thus, is not going to be discussed throughout this paper (this test can be used on their webpage http://www.p-curve.com/ under "The online app 4.0" tab).

After the release of this new tool, an article called "The Extent and Consequences of P-Hacking in Science" (Head et al. 2015) came to the surface, in which a combination of text mining and p-curve testing was used to demonstrate that p-hacking is commonly used in research to obtain false positive results. Unfortunately, it didn't take long for some papers to appear ((Ulrich and Miller 2015) and (Bishop and Thompson 2016)) that demonstrated the difficulty of detecting p-hacking just by using the p-curve test for right-skewness of the p-curve. Ulrich and Miller (2015) stated that:

> "... p-curve fails to detect false positive results when researchers conduct multiple statistical tests. In these cases, p-curves are right skewed, thereby mimicking the existence of real effects even if no effect is actually present..."

This result combined with the idea of a one to many mappings from "The garden of forking paths" undermines the effectiveness of the p-curve as a tool for detecting p-hacking in a major way. In the general case a p-curve at most can tell whether there was no p-hacking involved in some research topic if the null hypothesis is true. But it can't tell apart p-hacking due to multiple comparisons or because there exists a real effect.

Discussion and Conclusion

P-hacking is a sensible topic among the research community, because it affects the credibility of research by publishing false positive results. This means that the whole evolution of science is on a constant danger, since the advancement of any field is made by incremental steps. This means that there could be research that is based on p-hacking that might be used for policy or decision making. This could have dire consequences on extreme cases.

The fact that there exists an interest on understanding and undermining when Post Hoc selection is being done and when is not, shows the relevance of this topic for the process of knowledge generation. Unfortunately, this task is easier said than done. P-hacking can easily be a by product of bad analysis methodologies (Gelman and Loken 2013) or an action made by a researcher in order to improve it's performance score. Either way, it's easy to see how fragile the p-value is as an objective way of showing evidential value of a phenomenon in science.

Even though, there are tools being created to better understand when p-hacking is present, like the p-curve, and there are recommendations being made on how to minimize the effects of multiple comparisons on research (Gelman and Loken 2013) (Simonsohn, Nelson, and Simmons 2014) (Head et al. 2015), there is still going to be an incentive for researchers to ignore these recommendations and try to get positive results. This incentive emerges, as stated in the introduction, from the fact that there exists a publication bias and that researchers performance is measured by number of publications. As long as the structure of the research cycle stays the same, no quantity of statistical rigor or new methodologies being created to identify p-hacking is going to eradicate the problem. The incentive will still be there.

This is the reason why I think a transition from pay-walled journals, to open sourced one's that are peer reviewed and that incentivises replication articles and a preregistration of the data analysis being done, is the way to eliminate the negative incentive of p-hacking. This would erase the need for researchers to force positive results in order to get published. It would also accelerate the knowledge generating process in science. One of the examples that I known of, is on the field of machine learning, specifically the subset corresponding to deep learning. Where most of the research done is open sourced and preprints are widely available through web pages such as https://arxiv.org/ or https://distill.pub where interactive research is being incentivized.

In conclusion, the phenomenon of p-hacking arises from the structure of the current research cycle in science. The effects and magnitude of it are still hard to estimate, since there isn't a reliable tool that can tell with

precision when it happens. And the only way to overcome this drawback is to change not how research is done, but how research is valued.

References

Bishop, Dorothy V.M., and Paul A. Thompson. 2016. "Problems in Using P-Curve Analysis and Text-Mining to Detect Rate of P-Hacking and Evidential Value." *PeerJ*, February, 1–16.

Gelman, Andrew, and Eric Loken. 2013. "The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No 'Fishing Expedition' or 'P-Hacking' and the Research Hypothesis Was Posited Ahead of Time." Columbia University; Penn State University.

Head, Megan L., Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. 2015. "The Extent and Consequences of P-Hacking in Science." *PLOS Biology* 13 (3): 1–15.

Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons. 2014. "P-Curve: A Key to the File-Drawer." *Journal of Experimental Psychology: General* 143 (2): 534–47.

Ulrich, Rolf, and Jeff Miller. 2015. "P-Hacking by Post Hoc Selection with Multiple Opportunities: Detectability by Skewness Test?: Comment on Simonsohn, Nelson, and Simmons (2014)." *Journal of Experimental Psychology: General* 144 (6): 1137–45.