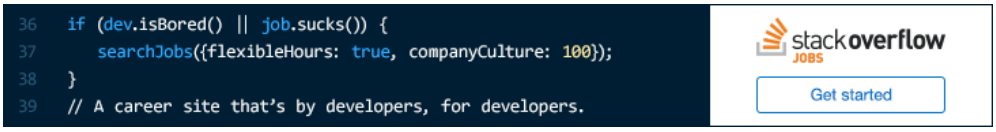
153 5 help

How much do we know about p-hacking “in the wild”?



The phrase *p*-hacking (also: "data dredging", "snooping" or "fishing") refers to various kinds of statistical malpractice in which results become artificially statistically significant. There are many ways to procure a "more significant" result, including but by no means limited to:

- only **analysing an "interesting" subset of the data**, in which a pattern was found;
- **failing to adjust properly for multiple testing**, particularly post-hoc testing and failing to report tests carried out that were not significant;
- **trying different tests of the same hypothesis**, e.g. both a parametric and a non-parametric test (there's some discussion of that in this thread), but only reporting the most significant;
- **experimenting with inclusion/exclusion of data points**, until the desired result is obtained. One opportunity comes when "data-cleaning outliers", but also when applying an ambiguous definition (e.g. in an econometric study of "developed countries", different definitions yield different sets of countries), or qualitative inclusion criteria (e.g. in a meta-analysis, it may be a finely balanced argument whether a particular study's methodology is sufficient robust to include);
- the previous example is related to **optional stopping**, i.e., analyzing a dataset and deciding on whether to collect more data or not *depending on the data collected so far* ("this is almost significant, let's measure three more students!") without accounting for this in the analysis;
- **experimentation during model-fitting**, particularly covariates to include, but also regarding data transformations/functional form.

So we know *p*-hacking can be done. It is often listed as one of the "dangers of the *p*-value" and was mentioned in the ASA report on statistical significance, discussed here on Cross Validated, so we also know it's a Bad Thing. Although some dubious motivations and (particularly in the competition for academic publication) counterproductive incentives are obvious, I suspect it's hard to figure out quite *why* it's done, whether deliberate malpractice or simple ignorance. Someone reporting *p*-values from a stepwise regression (because they find stepwise procedures "produce good models", but aren't aware the purported *p*-values are invalidated) is in the latter camp, but the effect is still *p*-hacking under the last of my bullet points above.

There's certainly evidence that *p*-hacking is "out there", e.g. Head et al (2015) looks for tell-tale signs of it infecting the scientific literature, but what is the current state of our evidence base about it? I'm aware that the approach taken by Head et al was not without controversy, so the current state of the literature, or general thinking in the academic community, would be interesting. For instance do we have any idea about:

- Just how prevalent is it, and to what extent can we differentiate its occurrence from *publication bias*? (Is this distinction even meaningful?)
- Is the effect particularly acute at the $p \approx 0.05$ boundary? Are similar effects seen at $p \approx 0.01$, for instance, or do we see whole ranges of *p*-values affected?
- Do patterns in *p*-hacking vary between academic fields?
- Do we have any idea which of the mechanisms of *p*-hacking (some of which are listed in the bullet points above) are most common? Have some forms proven harder to detect than others because they are "better disguised"?

References

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of *p*-hacking in science. *PLoS Biol*, *13*(3), e1002106.

hypothesis-testing | statistical-significance | p-value | model-selection | reproducible-research

edited Apr 13 at 12:44
Community ♦
1

asked Mar 9 '16 at 13:14
Silverfish
11.7k 11 53 103

6 Your last question is a nice idea for a research: give some raw data to a group of researchers form different fields, equip them in SPSS (or whatever they use) and then record what they are doing while competing with each other for more significant results. – Tim Mar 9 '16 at 13:26

One might be able to do it without the subjects knowing it was happening using a history of kaggle submissions. They aren't publishing, but they are trying every way possible to hit the magic number. – EngrStudent Mar 9 '16 at 13:27

Does crossvalidated have any collections (e.g. community wikis) of simple simulation examples of p-hacking? I'm imagining toy examples in which the simulated researcher reacts to "marginally significant" results by collecting more data, experiments with regression specifications, etc. – Adrian Mar 9 '16 at 13:28

1 @Adrian CV is just a Q&A site, it does not hold any data, or code, does not have any hidden repository - everything you find in answers is yours under CC license :) **This** question seems to be asking about collecting such examples. – Tim Mar 9 '16 at 13:34

@Tim of course, I wasn't imagining any hidden code repos -- just code snippets included in answers. For example, someone might ask "what is p-hacking?", and someone might include a toy R simulation in their answer. Would it be appropriate to respond to the current question with code examples? "How much do we know" is a very broad question. – Adrian Mar 9 '16 at 14:01

2 Answers

[Answer massively updated in Dec 2016.]

Great question.

Andrew Gelman likes to write about this topic and has been posting *extensively* about it lately on his blog. I don't always agree with him but I like his perspective on p-hacking. Here is an excerpt from the Introduction to his Garden of Forking Paths paper (Gelman & Loken 2013; a version appeared in American Scientist 2014; see also Gelman's brief comment on the ASA's statement), emphasis mine:

> This problem is sometimes called “p-hacking” or “researcher degrees of freedom” (Simmons, Nelson, and Simonsohn, 2011). In a recent article, we spoke of “fishing expeditions [...]”. But we are starting to feel that the term “fishing” was unfortunate, in that it invokes an image of a researcher trying out comparison after comparison, throwing the line into the lake repeatedly until a fish is snagged. We have no reason to think that researchers regularly do that. We think the real story is that researchers can perform a reasonable analysis given their assumptions and their data, but had the data turned out differently, they could have done other analyses that were just as reasonable in those circumstances.
>
> **We regret the spread of the terms “fishing” and “p-hacking”** (and even “researcher degrees of freedom”) for two reasons: first, because when such terms are used to describe a study, there is the misleading implication that researchers were consciously trying out many different analyses on a single data set; and, second, because it can lead researchers who know they did not try out many different analyses to mistakenly think they are not so strongly subject to problems of researcher degrees of freedom. [...] **Our key point here is that it is possible to have multiple potential comparisons, in the sense of a data analysis whose details are highly contingent on data, without the researcher performing any conscious procedure of fishing or examining multiple p-values.**

So: Gelman does not like the term *p-hacking* because it sort of implies that the researches were actively cheating. Whereas the problems can occur simply because the researchers choose what test to perform/report after looking at the data, i.e. after doing some exploratory analysis.

With some experience of working in biology, I can safely say that *everybody* does that. Everybody (myself included) collects some data with only vague a priori hypotheses, does extensive exploratory analysis, runs various significance tests, collects some more data, runs and re-runs the tests, and finally reports some p-values in the final manuscript. All of this is happening without actively cheating, doing dumb xkcd-jelly-beans-style cherry-picking, or consciously hacking anything.

So to the extent that "p-hacking" is understood broadly *a la* Gelman's forking paths, the answer to how prevalent it is, is that it is almost universal.

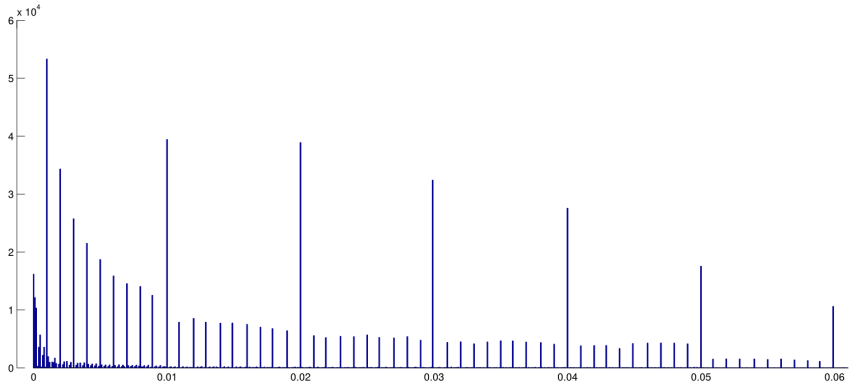
The only exceptions that come to my mind might be pre-registered replications in psychology and perhaps medical trials.

P-value distributions in the literature

Head et al. 2015

I have not heard about Head et al. study before, but have now spent some time looking through the surrounding literature. I have also taken a brief look at their raw data.

Head et al. downloaded all Open Access papers from PubMed and extracted all p-values reported in the text, getting 2.7 mln p-values. Out of these, 1.1 mln was reported as $p = a$ and not as $p < a$. Out of these, Head et al. randomly took one p-value per paper but this does not seem to change the distribution, so here is how the distribution of all 1.1 mln values looks like (between 0 and 0.06):



I used 0.0001 bin width, and one can clearly see a lot of predictable rounding in the reported p-values. Now, Head et al. do the following: they compare the number of p-values in the $(0.045, 0.5)$ interval and in the $(0.04, 0.045)$ interval; the former number turns out to be (significantly) larger and they take it as an evidence of p-hacking. If one squints, one can see it on my figure.

I find this hugely unconvincing for one simple reason. Who wants to report their findings with $p = 0.05$? Actually, many people seem to be doing exactly that, but still it appears natural to try to avoid this unsatisfactory border-line value and rather to report another significant digit, e.g. $p = 0.048$ (unless of course it's $p = 0.052$). So some excess of p-values close but not equal to 0.05 can be explained by researcher's rounding preferences.

And apart from that, the effect is *tiny*.

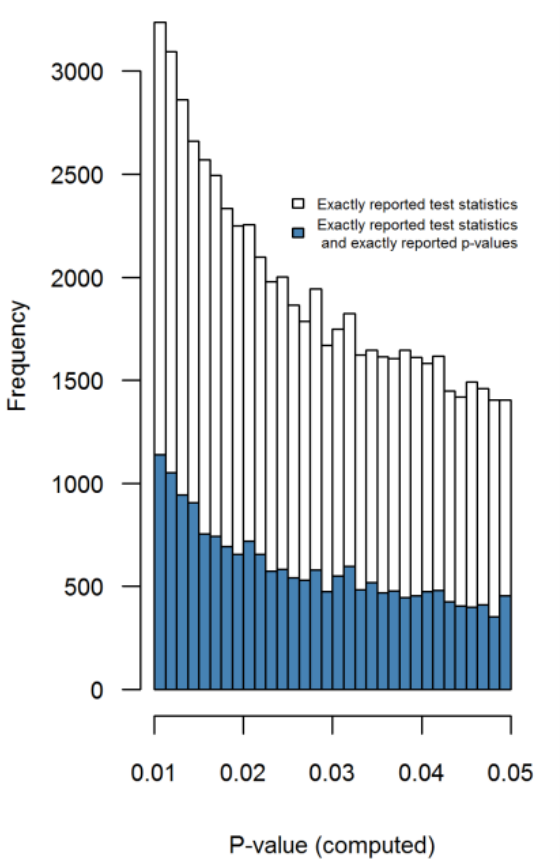
(The only strong effect that I can see on this figure is a pronounced drop of the p-value density right after 0.05. This is clearly due to the publication bias.)

Unless I missed something, Head et al. do not even discuss this potential alternative explanation. They do not present any histogram of the p-values either.

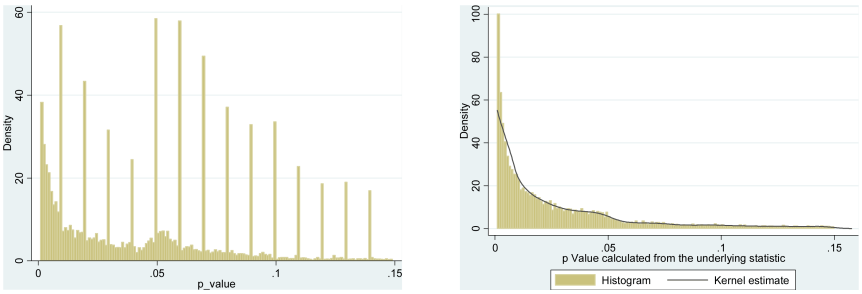
There is a bunch of papers criticizing Head et al. In this unpublished manuscript Hartgerink argues that Head et al. should have included $p = 0.04$ and $p = 0.05$ in their comparison (and if they had, they would not have found their effect). I am not sure about that; it does not sound very convincing. It would be much better if we could somehow inspect the distribution of the "raw" p-values without any rounding.

Distributions of p-values without rounding

In this 2016 PeerJ paper (preprint posted in 2015) the same Hartgerink et al. extract p-values from lots of papers in top psychology journals and do exactly that: they recompute exact p-values from the reported t-, F-, χ^2- etc. statistic values; this distribution is free from any rounding artifacts and does not exhibit any increase towards 0.05 whatsoever (Figure 4):



A very similar approach is taken by Krawczyk 2015 in PLoS One, who extracts 135k p-values from the top experimental psychology journals. Here is how the distribution looks for the reported (left) and recomputed (right) p-values:



The difference is striking. The left histogram shows some weird stuff going on around $p = 0.05$, but on the right one it is gone. This means that this weird stuff is due to people's preferences of reporting values around $p \approx 0.05$ and not due to p-hacking.

Mascicampo and Lalande

It seems that the first to observe the alleged excess of p-values just below 0.05 were Masicampo & Lalande 2012, looking at three top journals in psychology:

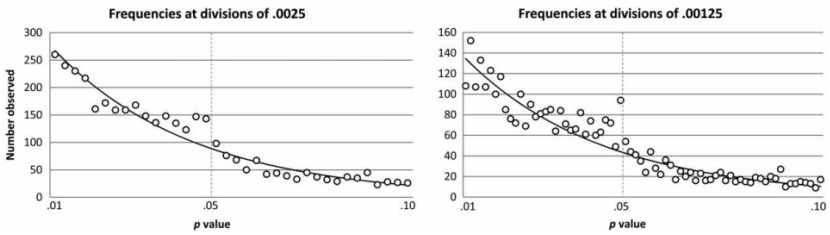
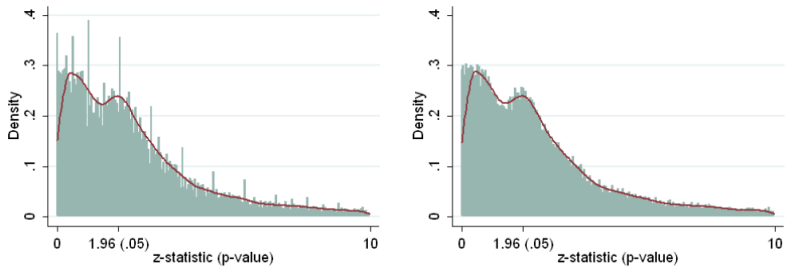


Figure 1. *The graphs show the distribution of 3,627 p values from three major psychology journals.*

This does look impressive, but Lakens 2015 (preprint) in a published Comment argues that this only *appears* impressive thanks to the misleading exponential fit. See also Lakens 2015, On the challenges of drawing conclusions from p-values just below 0.05 and references therein.

Economics

Brodeur et al. 2016 (the link goes to the 2013 preprint) do the same thing for economics literature. The look at the three economics journals, extract 50k test results, convert all of them into z-scores (using reported coefficients and standard errors whenever possible and using p-values if only they were reported), and get the following:



(a) Raw distribution of z-statistics. (b) Unrounded distribution of z-statistics.

This is a bit confusing because small p-values are on the right and large p-values are on the left. As authors write in the abstract, "The distribution of p-values exhibits a camel shape with abundant p-values above .25" and "a valley between .25 and .10". They argue that this valley is a sign of something fishy, but this is only an indirect evidence. I am not sure if this effect is present in biological literature or not because the plots above focus on $p < 0.05$ interval.

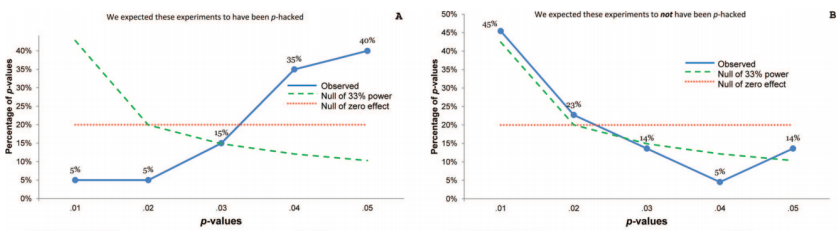
Falsely reassuring?

Based on all of the above, my conclusion is that I don't see any strong evidence of p-hacking in p-value distributions across biological/psychological literature as a whole. There is plenty of evidence of selective reporting, publication bias, rounding p-values *down* to 0.05 and other funny rounding effects, but I disagree with conclusions of Head et al.: there is no suspicious bump below 0.05.

Uri Simonsohn argues that this is "falsely reassuring". Well, actually he cites these papers uncritically but then remarks that "most p-values are way smaller" than 0.05. Then he says: "That's reassuring, but falsely reassuring". And here is why:

> If we want to know if researchers p-hack their results, we need to examine the p-values associated with their results, those they may want to p-hack in the first place. Samples, to be unbiased, must only include observations from the population of interest.
>
> Most p-values reported in most papers are irrelevant for the strategic behavior of interest. Covariates, manipulation checks, main effects in studies testing interactions, etc. Including them we underestimate p-hacking and we overestimate the evidential value of data. Analyzing all p-values asks a different question, a less sensible one. Instead of "Do researchers p-hack what they study?" we ask "Do researchers p-hack everything?"

This makes total sense. Looking at *all* reported p-values is way too noisy. Uri's p-curve paper (Simonsohn et al. 2013) nicely demonstrates what one can see if one looks at carefully selected p-values. They selected 20 psychology papers based on some suspicious keywords (namely, authors of these papers reported tests controlling for a covariate and did not report what happens without controlling for it) and then took only p-values that are testing the main findings. Here is how the distribution looks like (left):

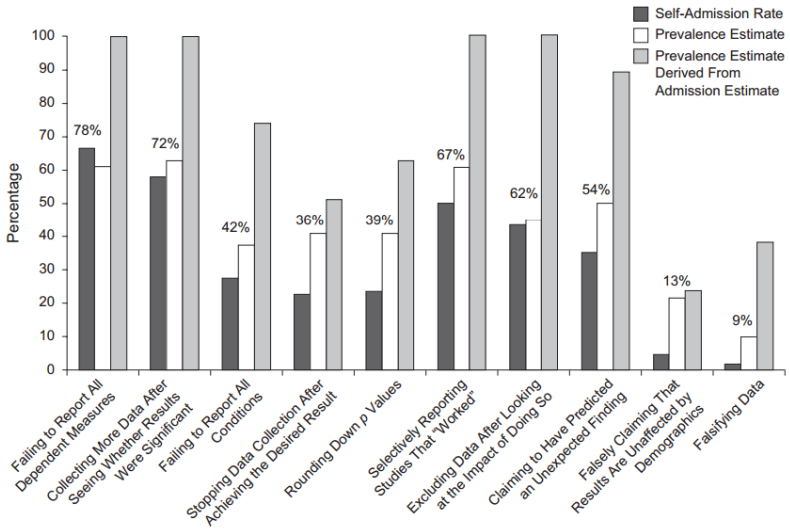


Strong left skew suggests strong p-hacking.

Conclusions

Even though the evidence presented by Head et al. and in the related studies is not particularly convincing, there is hardly any doubt that p-hacking is extremely widespread. As I wrote above, my own anecdotal observations certainly support it. Amusingly, some people even polled researchers to find that many admit doing some sort of hacking (John et al. 2012,

Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling):



And also, everybody heard about the so called "replication crisis" in psychology: more than one half of the recent studies published in the top psychology journals do not replicate (Nosek et al. 2015, Estimating the reproducibility of psychological science). (This study has recently been all over the blogs again, because the March 2016 issue of Science published a Comment attempting to refute Nosek et al. and also a reply by Nosek et al. The discussion continued elsewhere, see post by Andrew Gelman and the RetractionWatch post that he links to. To put it politely, the critique is unconvincing.)

So I would say that we *know* that there *must* be a lot of p-hacking going on, mostly of the Forking-Paths type that Gelman describes; probably to the extent that published p-values cannot really be taken at face value and should be "discounted" by the reader by some substantial fraction. However, this attitude seems to produce much more subtle effects than simply a bump in the *overall* p-values distribution just below 0.05 and cannot really be detected by such a blunt analysis.

edited Dec 2 '16 at 8:49

answered Mar 9 '16 at 13:47

amoeba
38.4k 10 132 193

4 `simply because the researches chose what test to perform/report after looking at the data` Yes; and the problem is anavoidable because double-edged. When a better method is being chosen for the data - is it an overfitting of that specific sample or a meeting of technical calls of that population? Or - removimg outliers - is it faking the population or recovering it? Who will say, ultimately? – ttnphns Mar 9 '16 at 15:18

The kind of answer I was most hoping for was perhaps a brief representation of the current literature, some pointers as to whether the Head et al paper is a fair summary of the latest thinking, etc. I wasn't expecting this answer at all. But I think it's great, and Gelman's thoughts and the practical insights are particularly helpful. When I wrote the question I had similar things in mind to @ttnphns actually (perhaps it shows, I even considered including the word "overfitting".) – Silverfish Mar 9 '16 at 16:04

Nevertheless, aside from the general and inescapable malaise of "how science works in practice" being an imperfect match for the assumptions of statistical testing, I do wonder if this bogeyman "dark art of the malicious p-hackers" is really out there, and if so, just how far it reaches. There are definitely strong (mis)incentives to encourage it. – Silverfish Mar 9 '16 at 16:07

1 You got me curious with this Head et al. paper, @Silverfish, so I must confess that right now, instead of working, I am browsing through some papers criticizing Head et al.'s results and have even already downloaded their raw data... Oh my. – amoeba Mar 9 '16 at 16:08

1 +1. The latest Gelman blog article (andrewgelman.com/2016/03/09/…) covers a lot of ground, and highlights an interesting rejoinder by a group that attempted replications and was then strongly criticized by the original study authors: retractionwatch.com/2016/03/07/… – Wayne Mar 9 '16 at 22:29


```
if (dev.isBored() || job.sucks()) {
    searchJobs({flexibleHours: true, companyCulture: 100});
}
// A career site that's by developers, for developers.
```

stackoverflow JOBS

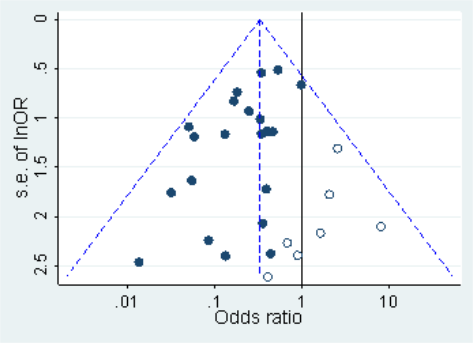
Get started

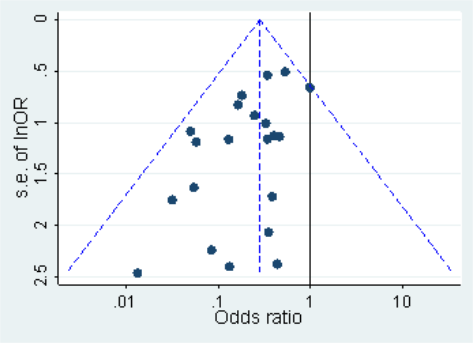
Funnel plots have been a tremendous statistical innovation that turned meta analysis on its head. Basically, a funnel plot shows the clinical and statistical significance on the same plot. Ideally, they would form a funnel shape. However, several meta-analyses have produced funnel plots that show a strong bimodal shape, where investigators (or publishers) selectively

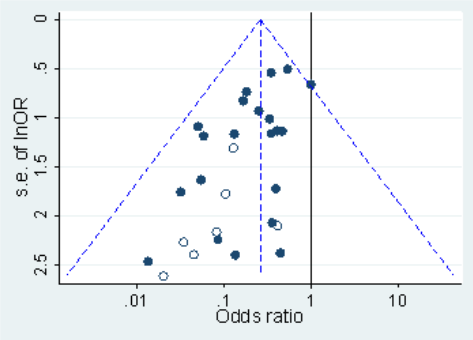
withheld results that were null. The result is that the triangle becomes wider, because smaller, less powered studies used more drastic methods to "encourage" results to reach statistical significance. The Cochrane Report team has this to say about them.

> If there is bias, for example because smaller studies without statistically significant effects (shown as open circles in Figure 10.4.a, Panel A) remain unpublished, this will lead to an asymmetrical appearance of the funnel plot with a gap in a bottom corner of the graph (Panel B). In this situation the effect calculated in a meta-analysis will tend to overestimate the intervention effect (Egger 1997a, Villar 1997). The more pronounced the asymmetry, the more likely it is that the amount of bias will be substantial.

The first plot shows a symmetrical plot in the absence of bias. The second shows an asymmetrical plot in the presence of reporting bias. The third shows an asymmetrical plot in the presence of bias because some smaller studies (open circles) are of lower methodological quality and therefore produce exaggerated intervention effect estimates.







I suspect most authors are unaware of the methods they use to p-hack. They don't keep track of the overall number of models they fit, applying different exclusion criteria or opting for different adjustment variables each time. However, if I had to mandate a simple process, I would love to see the total number of models fit. That's not to say there might be legitimate reasons to rerun models, for instance we just ran through a Alzheimer's analysis not knowing ApoE had been collected in the sample. Egg on my face, we reran the models.

edited Mar 9 '16 at 19:02

answered Mar 9 '16 at 18:45

AdamO
21.3k 36 76

2 I like that you emphasize "investigators **(or publishers)** selectively withheld results that were null". Given that fail to reject the null ≈ no publication, the fault is not necessarily squarely on the investigators. – Cliff AB Mar 9 '16 at 18:57

One aspect of my question was the distinction between "p-hacking" and "publication bias" - this answer in some

ways conflates the two. Would I be right to interpret what you're saying in that way, ie "publication bias is in essence a form of p-hacking, but by the publisher"? – Silverfish Mar 9 '16 at 21:07

@Silverfish Publication bias, per the earlier comment, can be driven by either the authors or the publishers. But yes, it is most definitely p-hacking. Funnel plots may have been applied to published research, but they are applicable in any setting where "scientific replication" starts to show discrepancies. Confirmatory trials for drugs, or implementations of business policies across a number of centers or wholesalers, no matter. Whenever you are dealing with replications, a funnel plot can provide some evidence of p-hacking by showing gaps where null results should have fallen. – AdamO Mar 10 '16 at 0:06

1 Hmm. First I wanted to protest and claim that publication bias is different from p-hacking (similarly, I think, to how @Silverfish framed his Q too), but then I realized that it is more tricky to draw the boundary than I originally thought. Performing jelly-beans-style multiple comparisons and only reporting significant ones (p-hacking?) is not very different from performing multiple studies and only reporting significant ones (which is publication bias by definition). Still, p-hacking in the sense of massaging the data until they yield p<0.05 does feel sufficiently different to me. – amoeba Mar 10 '16 at 10:12

1 @amoeba I had the same concern, but after reading OP's question, I realized it concerned the consequences of p-hacking on the "sausage end of things". Most p-hacking methods are usually omitted from reporting. So having been blinded to what the statistician does, how then do we reconcile the differences? Well, we need independent attempts to replicate and confirm findings. – AdamO Mar 10 '16 at 17:40