Contents lists available at SciVerse ScienceDirect

Journal of Mathematical Psychology

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

Commentary

It really just does not follow, comments on Francis (2013)

Uri Simonsohn

The Wharton School, University of Pennsylvania, 3730 Walnut Street, Philadelphia, PA19104, USA

HIGHLIGHTS

- Publication bias does not invalidate data but Francis proposes acting as if it does.
- The critiques suffer from publication bias, Francis proposes acting as if they do not.
- All studies suffer from publication bias.
- Studies that "survive" the excessive significance test are false-negatives.

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Logical inferences

ABSTRACT

I discuss points of agreement and disagreement with Francis (2013), and argue that the main lesson from his numerous one-off publication bias critiques is that developers of new statistical tools ought to anticipate their potential misuses and develop safeguards to prevent them.

1. Introduction

In numerous one-off critique-articles, Francis presents evidence that individual psychology papers suffer from publication bias, and concludes that the results from these papers ought to be fully ignored. I recently argued the critiques themselves suffer from publication bias and, more importantly, that the recommendation to throw out all data *does not follow* from the presence of publication bias (Simonsohn, 2012). Here I discuss Francis' (2013) reaction to these arguments, seeking to identify the issues we agree and disagree on.

2. If it does not follow, it does not follow

Francis (2013) writes "Simonsohn correctly notes that there are cases where [publication] bias is [statistically] significant and small" (p. xx). Francis argues, however, that it is prudent to act as if statistically significant publication bias always implies practical significance, when he continues: "but there are also cases where it is significant and large, and there is no way to distinguish between these. Scientific progress requires establishing a firm foundation of knowledge, thus, the prudent approach [. . .] is to treat experiment sets that appear to [have publication bias] as unscientific".

Francis and I agree, then, on the fact that the mere presence of publication bias *does not* imply it is consequential. Agree that publication bias *does not* warrant fully ignoring the underlying data. We disagree on what to do about this fact. Francis believes it is prudent to act as if it were not a fact. I believe it is imprudent to do so. I believe it is imprudent to toss studies as if they lacked evidential value before assessing whether they have evidential value. The excessive significance test Francis uses does not assess evidential value. It seems imprudent to me, but not to him, to act as if it did.

3. If you cherry pick, you cherry pick

Francis (2013) writes: "Simonsohn is correct that the published reports about publication bias are themselves biased. With one notable exception [. . .], all of the published investigations [. . .] have reported on experiment sets that are inconsistent/biased" (p. xx).

Francis and I also agree, then, on the fact that his critiques suffer from publication bias. We disagree on the consequences of this fact. Francis believes his reported p-values do adequately represent his false-positive rates, I believe his reported p-values dramatically underestimate his false-positive rates.[1]

E-mail address: uws@wharton.upenn.edu.

[1] Francis is also concerned that my treatment of the impact of multiple comparisons on p-values suffers from a "technical mistake" (p. xx) because "Adjustments for multiple testing, such as Bonferroni correction, involve a change in the criterion required to conclude statistical significance but do not alter the p-value." (p. xx). Three quick reactions. First, it is quite common to correct p-values for multiple comparisons (e.g., both SAS –*proc multtest*- and R –*p.adjust*- implement such corrections). Second, it makes sense to do so because multiple comparisons affect the probability of observing a pattern given the null (what the p-value reflects), not the consequences of rejecting such null (what the α-level reflects). Third, one of the most well cited papers on simultaneous inference is titled "Adjusted ***p*-values** for simultaneous inference" (emphasis added, Wright, 1992).

0022-2496/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmp.2013.03.006

Francis argues that publication bias does not inflate *his* false-positive rates because the papers he is critiquing are not related to one another. The intuition behind this argument can be quite compelling. When computing the probability of event X happening, why should we take into account the fact that we also examined if completely unrelated event Y happened? I answer this question drawing a parallel to how people mess-up when thinking about coincidences.

3.1. Francis' tests are like a young Swedish couple

The summer before starting grad-school, I posted an ad on a Seattle hostel bulletin board, offering a car ride to San Francisco. A Swedish couple, complete strangers to me, tagged along for the trip. Not yet in Tacoma we discovered they had just traveled, in Israel, with my sister's Chilean boyfriend. Wow! What in the world are the odds of that? Certainly lower than 10% (i.e., $p < .1$). Facing such odds, it is incredibly tempting to conclude that something other than chance landed those two blonds in my '83 Accord.

Coincidences like this one strike us as unlikely because we misconstrue the sampling behind them (for fun and accessible discussions of this point see Abumrad & Krulwich, 2009; Belkin, 2002; Paulos, 2002). We act as if we only examined whether the coincidence that did happen happened. But that of course is not at all what we are doing. We constantly monitor the world for *any* coincidence, and when they happen we tell our friends (and sometimes *JMP* readers) about it. The probability of *some* coincidence happening to you in your lifetime is about $p = 1$. Importantly for the present discussion of publication bias tests, the coincidences we are on the lookout for are completely unrelated to one another; they do not constitute different instances of the same phenomenon. They constitute unrelated instances of unrelated phenomena. What ties potential coincidences together is that they are part of the same sampling process: looking out for weird stuff. That is the sampling process we ought to consider when asking what are the odds of *any one* weird thing happening.

Because we do not predict "*this* coincidence will happen," then wait to see if it does, and tell everyone the outcome whether it happened or not, we dramatically underestimate how likely we are to observe *a* surprising coincidence.

Because Francis does not predict "*this* paper will show publication bias," then runs the test and tells everyone whether it does or it does not, he dramatically underestimates how likely he is to observe *a* $p < .1$ test.

3.2. Replications differentiate coincidences from facts

Francis (2013) rightly notes, replying to my first attempt to make the previous point, that we do not take into account all possible hypotheses when we report results from lab experiments. We just report the p-value we get without worrying about the ("unrelated") study that failed last month. This is true. As I previously wrote: "Although we should, we never do disclose – let alone correct for – the size of our file-drawers. Instead we address this problem with replications. Even if we got a study to work only after 44 attempts, there is still just a 5% chance of it working again under the null: replication p values are kosher. Without replications, however, p-values are meaningless if we do not take into account the size of the file-drawer behind them. Francis cherry picks but neither replicates nor shows his drawer" (Simonsohn, 2012, pp. 597).

Another way to think of this issue is through the exploratory vs. confirmatory research distinction. P-values assume one is engaging in confirmatory research. I argue that Francis research is exploratory and ought to be reported as such, he also argues it is exploratory but that it need not be reported as such because the exploration involves unrelated hypotheses.

3.3. Francis' publication bias tests are not unrelated to each other

Let us leave aside the issue that cherry picked p-values from unrelated tests inflate false-positive rates as much as those from related ones, and focus on whether Francis' critiques can indeed be considered as unrelated to one another.

The critiques seem anything but unrelated. They constitute interchangeable pieces of evidence that could be swapped, excluded or combined across any of the ten or so excessive-significance papers Francis has written.

The critiqued literatures are indeed typically unrelated from the perspective of the original authors. For example, two of the critiqued literatures include verbal overshadowing and precognition — they have nothing in common with each other phenomenon-wise. Nevertheless, they constitute Study 1 and Study 2 in a single critique by Francis (2012). Notably, had only one of these two demonstrations worked, Francis would consider it acceptable for him to drop the other. They are related enough to be reported in the same paper, his argument goes, but unrelated enough to be dropped without affecting false-positive rates.

The verbal overshadowing literature critiqued by Francis (2012), moreover, aggregates across different authors, different years, different manipulations, and different dependent variables. The critiques by Francis, in contrast, are by the same author, published the same year, conducting the same statistical test, to examine the exact same question (do social psychologists report all their failed studies?). The critiques seem at least as related to each other as the verbal overshadowing experiments do.

If, on the other hand, we were to consider the critiques as unrelated because that is how Francis sees them, then the critiqued authors themselves could argue just as reasonably that *their* studies are unrelated, perhaps because they test a phenomenon in different domains or with different moderators. This is a problem with conceptual replications, those that succeed are considered as related to previous demonstrations, those than fail are not (Pashler & Harris, 2012, p. 533). The critiqued authors, in short, could use Francis' argument of unrelatedness to fend off Francis' criticism of publication bias in their experiments.

If we were to consider as cherry picking only the selective reporting of studies that are related to each other, and we were to let authors conducting the studies determine if they are related, then the set of cherry-pickers would be an empty one.

3.4. Francis' drawer: filled with false-negatives

I have just argued that because he cherry picks, Francis' false-positive rates are higher than his reported p-values are. This is not to the same as saying his published results are likely to be false-positive. Prob(Data|Hypothesis)≠Prob(Hypothesis).

In fact, I would assume all his critiques are true-positives, all studies he has critiqued do suffer from publication bias. The problem is that all studies he has *not* (yet?) critiqued do also. Francis has zero false-positives among his critiques, because when "negative" means publishing in an environment without publication bias, the only two possible outcomes of a publication bias test are: true-positive and false-negative.

Francis tests answer this question: "has a large enough set of published studies been compiled to reject the obviously false null that all studies, regardless of outcome, would be reported?" When the answer is no, Francis file-drawers the false-negative result.

4. What are we learning from the (8 so far) Francis critiques?

We are not learning that publication bias happens, we already knew that. We are not learning that the critiqued studies ought to be ignored, because that just does not logically follow from them containing publication bias. We are not learning that the critiqued studies have more severe publication bias than others, because Francis' selective reporting of results, and non-representative selection of studies to analyze in the first place, prevents us from making such inference.

What do we learn then?

We learn that new statistical tools, perhaps especially those that provide potential critics with access to easy publications, can be misused. We learn, then, that developers of new tools ought to include in their papers safeguards to prevent their misuse. These safeguards may involve recommended disclosure guidelines (Simmons, Nelson, & Simonsohn, 2011, 2012). For instance, one could require critics to disclose how they selected the target of their critique: exploratorily or confirmatorily. Safeguards may also involve providing detailed discussions of likely inappropriate uses of the new tool. Safeguards may also involve providing guidance to peer-review teams for assessing work employing the new tool. Creating these safeguards is not only in the general interest of the field, but in the interest of the tool-creators themselves.

References

Abumrad, J., & Krulwich, R. (Producer) (2009). Stochasticity. *RadioLab*, Retrieved from http://www.radiolab.org/2009/jun/15/.

Belkin, L. (August 11 2002). The odds of that. *The New York Times*.

Francis, G. (2012). Too good to be true: publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, 1–6.

Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology*,.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? three arguments examined. *Perspectives on Psychological Science*, *7*(6), 531–536.

Paulos, J. A. (2002). The 9–11 lottery coincidence. *ABC News*.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. *Dialogue, The Official Newsletter of the Society for Personality and Social Psychology*, *26*(2), 4–7.

Simonsohn, U. (2012). It does not follow evaluating the one-off publication bias critiques by Francis (2012a, 2012b, 2012c, 2012d, 2012e, in Press). *Perspectives on Psychological Science*, *7*(6), 597–599.

Wright, S. P. (1992). Adjusted *p*-values for simultaneous inference. *Biometrics*, 1005–1013.

Please cite this article in press as: Simonsohn, U. It really just does not follow, comments on Francis (2013). Journal of Mathematical Psychology (2013), http://dx.doi.org/10.1016/j.jmp.2013.03.006