Selective publication of findings: Why does it matter, and what should we do about it?

Maximilian Kasy

November 6, 2019

Introduction

It has been argued in recent years that a sizable share of published research in the social sciences and the life sciences is not replicable [Camerer et al., 2016, Open Science Collaboration, 2015, Camerer et al., 2018], and much of conventional statistical inference in these fields is distorted [Ioannidis, 2005, Gelman and Loken, 2013]. One possible reason for this lack of replicability is selective publication of findings, whether by researchers (p-hacking, specification searching) or journals (publication bias). Either source of selection implies that the probability that some finding is published depends on the finding itself, and not just on the question, the method, or the sample size of a study.

The debates around this perceived replication crisis have led to the the creation of a number of projects, initiatives, and centers that aim to improve the transparency and reproducibility of research. These initiatives include, among others, the project on *Reproducibility and Replicability in Science* by the National Academy of Science, the *Berkeley Initiative for Transparency in the Social Sciences*, the *Institute for Quantitative Social Science* at Harvard, the *Meta-Research Innovation Center at Stanford*, as well as *Teaching Integrity in Empirical Research*, spanning several institutions. The reforms that have been promoted by these initiatives, and by others, include changes in norms (don't put "stars" based on significance in your tables), changes in journal policies (requiring pre-analysis plans for experimental research, accepting papers based on registered reports), and changes in the institutional infrastructure for academic research (journals for null-results, journals for replication studies).

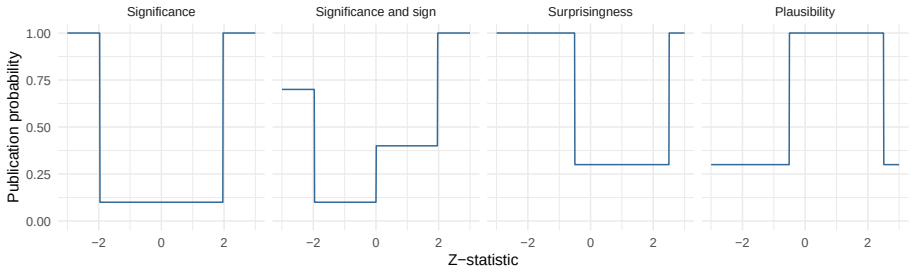
These are worthwhile efforts toward increasing the transparency, replicability, and credibility of published research. A core goal of these efforts is to reduce the degree to which published findings are selected. I believe, however, that there are several conceptual issues that deserve a more careful discussion at this point, in order to guide further reform efforts.

In the following I will argue that different justifiable objectives for scientific institutions lead to contradictory recommendations. We need to be explicit about our objectives in order to discuss the tradeoffs between them. Replicability and the validity of conventional statistical inference constitute one such

objective, and they indeed require that publication decisions do not depend on findings. This is what motivates much of current reform efforts. Validity of inference is presumably not the only objective, however – it could easily be achieved by estimates derived from a random number generator. Relevance of findings might be another objective. If our goal is to inform decision makers or to maximize social learning, there is a strong rationale to selectively publish surprising findings. A third objective could be the plausibility of published findings. If there is some uncertainty about the quality of studies and we want to avoid publishing incorrect results, we might want to selectively publish unsurprising findings. How can we resolve the tension between these contradictory recommendations? I will outline one possibility below, proposing a functionally differentiated publication system, with different outlets focusing on different objectives. Measures that are promoted by current reformers, such as pre-analysis plans and registered reports, would have to play a crucial role in such a system.

Following these policy proposals, I will take a step back and argue that these debates raise some fundamental questions for statistical theory. In order to coherently discuss these issues, statistical theory needs a model of the work of empirical research that goes beyond the *single-agent* model of statistical decision theory. We should understand statistics (quantitative empirical research) as a *social* process of communication and collective learning that involves many different actors with differences in knowledge and expertise, different objectives, and constraints on their attention and time, who engage in strategic behavior.

Figure 1: Some possible forms of selection



Forms of selection To set the stage for the subsequent discussion of alternative objectives, let us first briefly review different forms that selection based on findings might take, and what we know about selection in the current publication system. Findings might be selected by researchers – which specifications are included in a paper, which outcome variables or controls are considered, etc. Findings might also be selected by journals – are null results published, or

results that contradict conventional wisdom, etc.

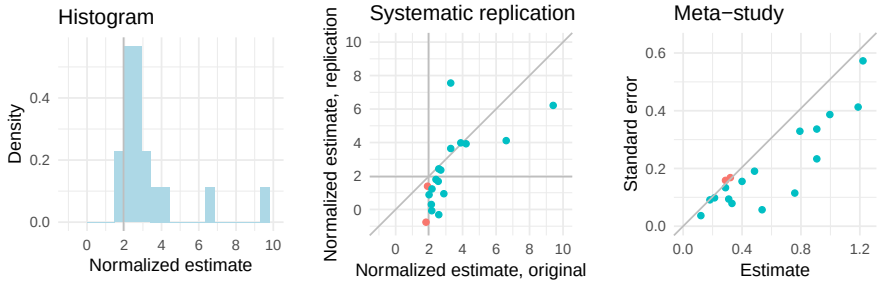
The most commonly discussed and criticized form of selection is based on significance. Studies might be more likely to be published if their headline finding corresponds to a test-statistic exceeding the 5% critical value, or some other conventional value. The leftmost plot in Figure 1 illustrates. The function $p(.)$ shown is the probability of a finding to be written up and published, depending on its z-statistic. In the example shown, the publication probability jumps up once the z-statistic exceeds the critical value of 1.96.

But there are many other forms of selection that might be operational in various fields, which are less often discussed in debates about replicability. In addition to significance, referees might for instance care about whether the estimate has the "right sign," according to theory or conventional beliefs; this might lead to functions $p(.)$ as in the second plot of Figure 1.

Researchers or referees might also compare findings to a reference point other than zero. They might for instance value surprisingness relative to some prior mean, as in the third example shown in Figure 1. Or they might do the opposite, and consider findings implausible when they deviate a lot from prior believes, which might lead to selection as in the last example shown.

Which findings get published? A number of methods have been used to detect the presence of selective publication. These methods include (1) checking the distribution of published p-values for discontinuities, which can only arise because of selective publication, (2) meta-regressions, regressing estimates across studies on (inverse) standard errors, and checking whether the slope differs from 0, and (3) discussions of the "rate of replication" of significant findings, where a low rate is taken as indicative of problems. The first two of these methods provide valid tests of the null hypothesis of no selection, but they do not have power for all forms of selection. Method (3) is more limited, since the "rate of replication" of significant findings, taken by itself, does not tell us much about selection. Low "rates of replication" can arise without any selectivity if most true effects are small or equal to 0.

Figure 2: Evidence for selective publication in economics lab experiments



In [Andrews and Kasy, 2018], we develop two alternative methods for identifying and estimating the form and magnitude of selection in the publication process. This is a more ambitious goal than just testing for the presence of some form of selection. For illustration of our methods, consider the data of [Camerer et al., 2016], who replicated 18 laboratory experiments published in top economics journals in the years 2011 to 2014. Figure 2 plots data from this systematic replication in different ways. The left figure shows the distribution of z-statistics based on the original studies. The middle figure shows (normalized) original and replication estimates. The right figure shows original estimates and their standard errors.

Our first proposed method is based on systematic replication estimates, as shown in the middle figure. In the absence of selective publication, there should be no systematic difference between originally published estimates and replication estimates, so that flipping the axes in the figure should not systematically change the picture (leaving differences in sample size aside). Selective publication, however, breaks this symmetry. Our second proposed method is based on only the estimates and standard errors from the original studies, and is illustrated in the right figure. In the absence of selective publication, estimates for studies with higher standard errors (smaller sample sizes) should be more dispersed. Deviations from this prediction again allow to fully pin down (estimate) the mapping from estimates to publication probabilities.

Estimates using the data of [Camerer et al., 2016] based on either of these methods suggest a similar form and degree of selectivity, where the probability of publication jumps 30-fold at the cutoff 1.96. For other fields in economics, we find different forms of selectivity; findings might for instance be more likely to be published if their sign conforms with standard theory.

Which findings should be published?

Validity of inference Why is selection of findings for publication, whether by researchers or by journals, a problem? Because it makes all our standard inferences invalid. As we show in [Frankel and Kasy, 2018], standard inference methods are valid if and only if publication probabilities do not depend on findings in any way (dependence on standard errors is allowed). Any form of selection leads to biased estimates, distortions of size for tests and confidence sets, and incorrect Bayesian posteriors, if not properly accounted for. For the sake of discussion, consider the extreme case where only findings exceeding the 5% significance threshold of 1.96 (for standard normal estimates) are published. Figure 3 illustrates this case.

The figure on the left shows the resulting bias of point estimates as a function of the true effect, conditional on publication. For very large true effects no bias occurs, since such studies are published with very high probability. For a true effect of zero, no bias occurs either, by symmetry of selection. For intermediate effect sizes of around 1 standard error, however, point estimates are biased upward by up to 1.5 standard errors, conditional on publication. The middle

Figure 3: Distortions induced by selective publication

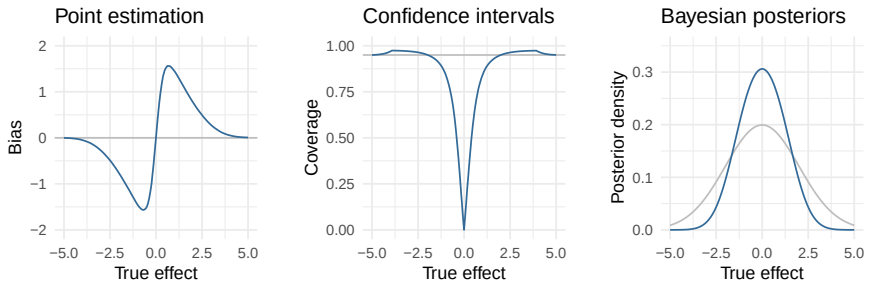


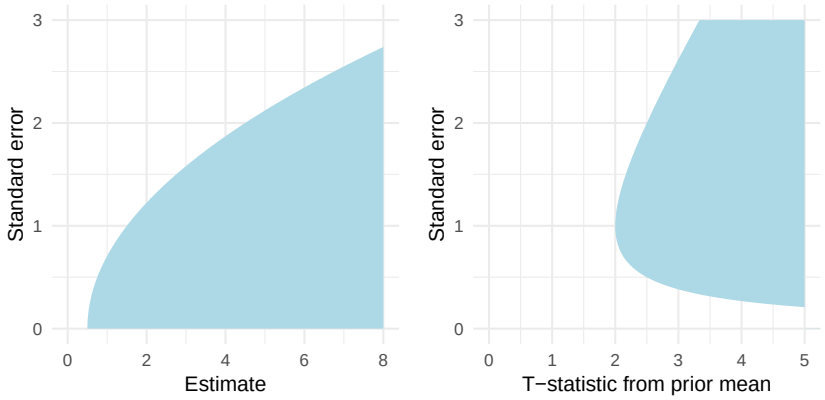
figure similarly plots the probability of containing the true effect (coverage) of a nominal 95% confidence interval, conditional on the true effect and conditional on publication. For large true effects, again no size distortions happen. For small effects, however, coverage conditional on publication can go all the way down to 0. Consider, lastly, a Bayesian reader of the literature. How should she update her beliefs? When observing a published finding, she need not take into account selection based on that finding, but she needs to update her beliefs in the *absence* of a publication. *Not* observing a publication makes it more likely that the true effect is close to zero. The figure on the right contrasts the Bayesian posterior (for a normal prior) in the absence of publication to the corresponding naive posterior, which ignores selection. Note in particular, that the distortions induced by selective publication affect both frequentist and Bayesian inference.

To summarize, selective publication can heavily distort statistical inference, whether frequentist or Bayesian. And there is ample evidence that publication is selective, albeit to different degrees and in different ways across different empirical fields. Recognition of these facts has motivated reform initiatives aimed at increasing the replicability and credibility of published research by reducing selection in the publication process. These are valuable initiatives that are likely to improve the standards of scientific evidence. They do raise, however, the question of what it is that reforms of academic research institutions and norms wish to ultimately achieve?

So let us take a step back and consider this question. Clearly, validity of inference should not be our only goal. Presumably we also care about ultimate objectives such as scientific progress, social learning, or helping decision makers in medicine, public policy, and technology. To put it starkly, publishing only estimates calculated based on a random number generator will yield statistical inference that is valid, but presumably not very interesting.

Relevance Consider, as an example, clinical research on new therapies. Suppose that in some hypothetical area of medicine a lot of new therapies, say drugs

Figure 4: Example of optimal publication regions for binary decisions



or surgical methods, are tested in clinical studies. Suppose that most of these trials don't work out - the new therapies just don't deliver. Absent a publication of successful clinical research, no doctor would implement these new therapies. And doctors have limited time - they are not going to read hundreds of studies every month. But they might read some.

In this hypothetical scenario, which findings should be published? That is, which subset of studies should doctors read? In order to improve medical practice, it would arguably be best to tell doctors about the small subset of new therapies which were successful in clinical trials. Those are the ones they should incorporate into their practice. Figure 4 illustrates optimal publication regions for an example of this form; see [Frankel and Kasy, 2018] for details.

If this is the selection rule used for publication, however, published findings are biased upward. Replications of the published clinical trials will systematically find smaller positive effects or even sometimes negative effects. This reasoning suggests that there is a deep tension between relevance (for decision making) and replicability in the design of optimal publication rules.

In [Frankel and Kasy, 2018], we argue that this type of logic holds more generally, in any setting where published research informs decision makers and there is some cost which prevents us from communicating all the data. Such a cost clearly must be present; otherwise it would be optimal to simply publish all data, without any role for statistical inference, researchers, or journals. Given such a cost, it is not worth it to publish "null-results," understood as results that do not change decisions relative to the default absent publication. Surprising results, on the other hand, especially results that lead to large changes of optimal decisions, are of great value to decision makers, and should thus be preferred for publication. Furthermore, some notions of social learning, such as reducing the variance of posterior beliefs, are isomorphic to informing decision makers. Therefore similar conclusions go through when our goal is to maximize social

learning, subject to attention constraints.

Plausibility We have argued that validity of standard inference requires that we eliminate selection on findings, while (policy) relevance compels us to publish surprising findings. But what about the plausibility of findings? Often we would think that extreme (surprising) findings indicate that there might be some problem with the study design. For instance, if a study reports that a very minor intervention has major health benefits, it might be more likely that the reported findings are biased than that the authors stumbled upon a miracle cure.

To formalize this intuition, suppose that the referees and readers of a study are uncertain about the bias of a study, where the latter might arise for all kinds of reasons. Correct updating of beliefs will imply that the posterior expected bias is increasing in the surprisingness of a finding. Suppose we are again interested in the relevance of findings for decision makers. As before, unsurprising findings are not relevant for decisionmakers and should not be published. But very surprising findings are implausible, suggesting issues with the study, and should also not be published. Only intermediate findings satisfy the requirements of relevance and plausibility.

Achieving multiple objectives in a functionally differentiated publication system

These considerations leave us with the practical question of what to do about the publication system. How shall we trade off these conflicting objectives? Can we have validity, relevance, and plausibility at the same time? A possible solution might be based on a functional differentiation of publication outlets, which could build on the present landscape, while making the differences of objectives and implied publication policies across outlets more explicit. Such a differentiation avoids having to sacrifice one of these objectives (e.g. relevance) for the sake of another (e.g., validity and replicability). The following provides a sketch.

There might be a set of top outlets focused on publishing surprising ("relevant") findings, subject to careful quality vetting by referees. These outlets would have the role of communicating relevant findings to attention-constrained readers (researchers and decision makers). A key feature of these outlets would be that their results are biased, by virtue of being selected based on surprisingness. In fact, this is likely to be true for prominent outlets today, as well. Readers should be aware that this is the case: "Don't take findings published in Science at face value."

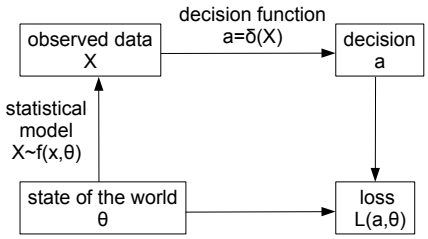
There might then be another, wider set of outlets that are not supposed to select on findings, but have similar quality-vetting as the top-outlets, thus focusing on validity and replicability. For experimental studies, pre-analysis plans and registered reports (results-blind review) might serve as institutional

safe-guards to ensure the absence of selectivity by both researchers and journals. Journals that explicitly invite submission of "null results" might be an important part of this tier of outlets. This wider set of outlets would serve as a repository of available vetted research, and would not be subject to the biases induced by the selectivity of top-outlets. Hiring and promotion decisions should take care to give similar weight to this wider set of publications as to top publications, so as to minimize the incentives for researchers to distort findings, whether by p-hacking or other means.

To make the findings from this wider set of publications available to attention-constrained decision makers, systematic efforts at aggregation of findings in review articles and meta-studies by independent researchers would be of great value. Lastly, systematic replication studies can serve as a corrective for the biases of top publications, and as a further safe-guard to check for the presence of selectivity among non-top publications, cf. [Andrews and Kasy, 2018].

Statistics as decision making versus statistics as communication

Figure 5: Statistics as a decision problem



Let us conclude by taking a step back to consider what the debates around replicability and selective publication imply for the foundations of statistics. One of the leading textbook models of statistics is statistical decision theory. Figure 5 summarizes the basic framework of decision theory.

The activity of statistics as conceived by decision theory is a rather solitary affair. It's just you and the data, and you have to make some decision based on the data - estimate a parameter, test a hypothesis, etc. This perspective can be extremely useful. It forces us to be explicit about what our objective is, what the action space is, and what prior information we wish to incorporate (in terms of the statistical model, in terms of a Bayesian prior, or in terms of a set of parameters for which we wish to control worst-case risk). And it makes explicit the tradeoffs involved in the choice of any statistical procedure. But this perspective also has severe limitations, as evidenced by the discussions around

p-hacking, publication bias, and pre-analysis plans. It is hard to make sense of these discussions from the vantage point of decision theory.

For instance, why don't we simply communicate all the data to the readers of research? If we took decision theory literally, that would indeed be optimal. Just communicating all the data avoids any issues of selection as well as any waste of information. In practice, as consumers of research we of course do prefer to read concise summaries of findings ("X has a big effect on Y, when W holds."), rather than staring at large unprocessed datasets. There is a role for researchers who carefully construct such summaries for readers. But it is hard to make sense of such a role for researchers unless we think of statistics as communication, and unless there is some constraint on the attention or time or information-processing capacity of readers.

Or, to take another example, what is the point of pre-analysis plans? Their purpose is often discussed in terms of multiple hypothesis testing (where reported p-values should account for selective reporting of tested hypotheses), or more generally in terms of the "garden of forking paths" of specification searching. But, taking the perspective of decision theory literally again, there is no obvious role for publicly committing to a pre-analysis plan in order to resolve this issue. Frequentist inference indeed requires knowledge of the mapping from data to reported statistics, for all counterfactual realizations of the data. But researchers might just communicate this mapping at the time of publication. To rationalize publicly registered pre-analysis plans, we again need to consider the social dimension of research, and assume (1) that there is some conflict of interest between researchers and readers, and (2) that there is some (attention) constraint that prevents the reporting of all data.

What these examples illustrate is that statistics (and empirical research more generally) is a social endeavor, involving different researchers, journal editors and referees, readers, policymakers, and others. Taking the social dimension seriously suggests a perspective on statistics where the task of empirical researchers is to provide useful summaries of complex data to their readers in order to promote some form of collective learning. This task is subject to costs of time and attention of researchers, referees, and readers, as well as constraints on social learning in terms of limited information, strategic behavior, the sociology of research, etc. Elaborating this perspective, where statistics gives normative recommendations for empirical practice while taking into account these social constraints, is an exciting task for the years ahead. This endeavor will have to draw, in particular, on microeconomic theory, psychology, and the sociology and history of science.

References

[Andrews and Kasy, 2018] Andrews, I. and Kasy, M. (2018). Identification of and correction for publication bias.

[Camerer et al., 2016] Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436.

[Camerer et al., 2018] Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E.-J., and Wu, H. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644.

[Frankel and Kasy, 2018] Frankel, A. and Kasy, M. (2018). Which findings should be published? *Working Paper*.

[Gelman and Loken, 2013] Gelman, A. and Loken, E. (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. *Department of Statistics, Columbia University*.

[Ioannidis, 2005] Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med*, 2(8).

[Open Science Collaboration, 2015] Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.