Estimating the false finding rate across scientific fields

Scott W. Piraino

September 30, 2017

Abstract

The possiblity of large amounts of false positives within the scientific literature has gained significant attention, particular in light of several replication projects in which large proportions of published studies have failed to replicate. I show through simulation that low replication rates can occur even when the published literature contains mostly true (non-null) findings. Using conservative estimates of the proportion of true null hypotheses within published studies based on replications, I show that the results of recent replication projects are consistent with the possiblity that most published research is true.

Introduction

Across many scientific disciplines, there is a growing concern that the scientific literature contains many “false positives”, meaning statistically significant positive results where the tested null hypothesis is actually true (i.e. the study commits a type I error). The article “Why Most Published Research Findings Are False” [1] popularized a theoretical argument which suggests that many claims that make their way into the literature may be wrong. This concern was given a possible empirical basis when several major collaborative projects were undertaken to attempt to independently perform empirical replications of published experiments [2,3]. Some of these projects produced replication rates which were lower than some scientists may have desired or expected, and although in many cases the original authors of these projects have been appropriately measured in their interpretations, these results do raise the concerning possiblity that a significant number of studies to replicate because the underlying conclusions of those studies are incorrect.

Do major replication projects actually support the empirical claim that most published research is false? Although much attention has been focused on the topic of false positive findings, extremely little work has acutally attempted to empirically assess this issue. Jager and Leek [4] performed one of the few analyses aimed at empirically estimating the proportion of findings which are false. The analysis by Jager and Leek [4] and the associated commentaries [5–10] raised several issues that might have rendered the estimate of the rate of false positives misleading. Here, I take advantage of data from several replication projects [2,3] which avoids many of the potential issues with the data used by Jager and Leek. My analysis suggests that data from replication projects does not conclusively show that most published research is false, and in some cases suggests that a the rate of false positives among the replicated studies is reasonably low.

Some readers may wonder why worrying about the accurate estimation of this proportion is worthwhile. Surely there are things about the scientific process that can improved. Is arguing about whether or not most research is false just a distraction that is derailing important efforts at reform? I argue that understanding whether or not most research is false is important not because there is question about whether change is needed, but because there are important questions about what reforms will actually lead to improvements, as well as what the goals of reform should be. Many reforms either explicitly or implicitly aim to decrease the proportion of false positives within the literature. For example, a recent proposal to change the interpretation of p-value thresholds [11] is based on this concern. If the rate of false positives in the literature is not actually large, this may raise questions about the usefulness of some policies, while potentially supporting the usefulness of others. Likewise, some reforms could be implemented differently depending on whether the main goal is to address false postives, or to address other issues with the scientific process. I discuss the role that the proportion of false findings could have on concrete policies in more detail in the discussion.

Results

Several recent replication projects have produced rates of replication that some may consider disappointing. If most published results are true, shouldn't a large proportion of those findings successfully replicate in independent replications? Here I show that this intuition is not nessarily accurate, that even when a literature contains mostly true (non-null) findings, many of these findings can fail to replicate. To demostrate this, I simulated the results of a replication project under a plausible model of the scientific publication process. I assume that scientists draw possible ideas to test from a pool that is half null and half alternative. Results that are statistically significant with a p-value less than 0.05 are published, while other results are not (i.e. studies are subject to publication bias). Replications are performed on published studies, and have an identical true effect size to the original study that they replicate. Replications have their sample size determined by a power analysis targeting 80% power based on the published effect size (similar to how replications in the Reproducibility Project: Psychology [2] were designed). Figure 1 shows the results of simulating from this model. The blue curve is a kernel density plot of the distribution of the true proportion of null effects among published results which is known within each simulation iteration. The green curve shows a plot of the replication failure rate (proportion of replications that fail to replicate). In this simulation, while the true false positive rate is relatively low, the failed replication rate is high.

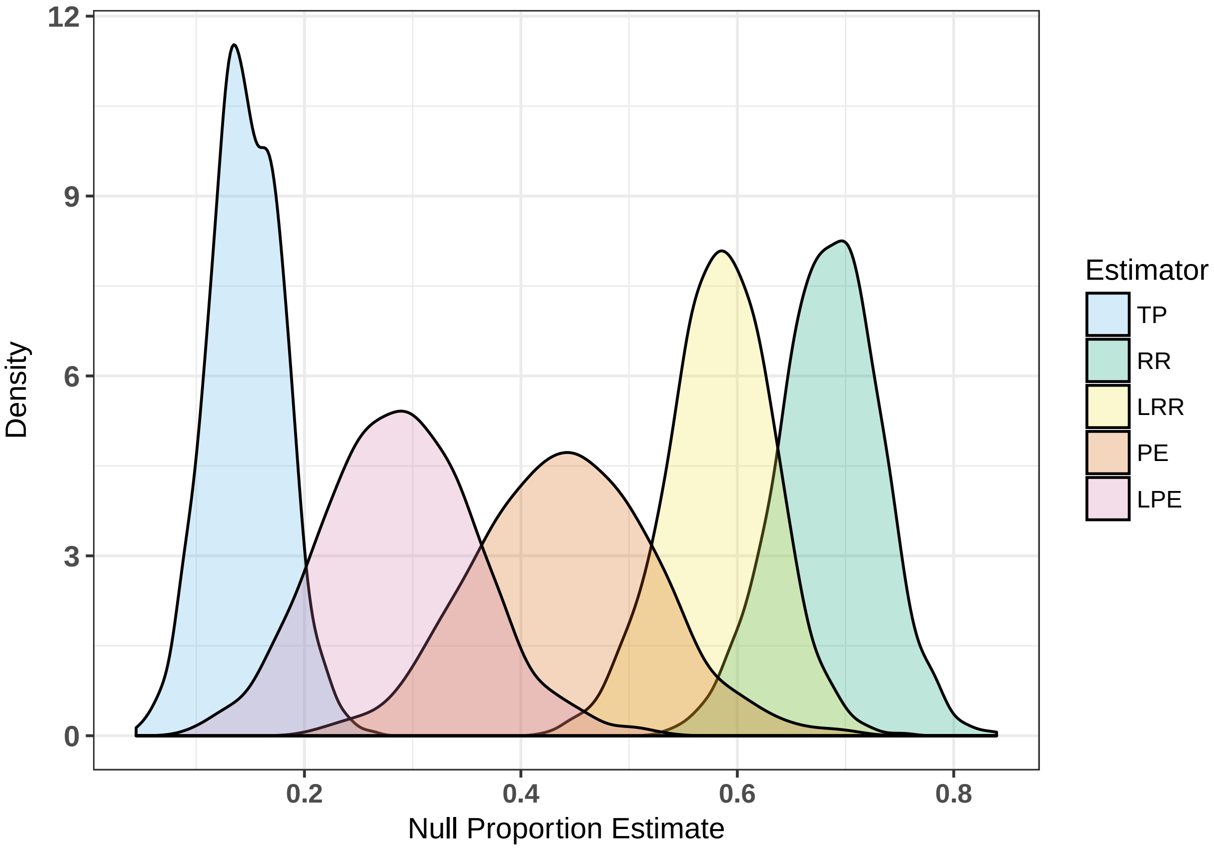

Figure 1: Kernel density plots of the true proportion of null hypotheses (TP, blue), an estimate this proportion from Storey [12] (PE, orange), a bootstrap 95% CI lower bound if this estimate (LPE, red), the failed replication rate (RR, green), and a lower 95% CI bound for this estimate from a proportions test (LRR, yellow), based on 1000 simulated replication projects

In addition to the true proportion on null results and the replication rate, I also show several other quantities that may be thought to estimate the proportion of the literature that is null. In yellow I plot a a density estimate for the lower bound of a 95% confidence interval for the replication failure rate, showing that even this lower bound is much large than the true proportion of null hypotheses among the replicated studies. In orange I plot an estimate of the proportion of true null hypotheses among the replicated studies based on a method developed by Storey [12], that is widely used in multiple-testing correction, along with a lower bound of a boostrap 95% confidence interval for this estimate in purple. This etimator and it's lower CI bound are also larger than the true proportion, was expected given the known conservativeness of this estimator [12]. However, this estimate is much less conservative as an estimator of the true proportion of null hypotheses compared to the replication failure rate.

Overall, when the replication failure rate is considered as an estimator of the number of "false positives" (significant findings that are truely null), under plausible assumptions about the publication process (low power and publication bias), the replication failure rate can be large even when the acutal proportion of false positives is small. The estimate from Storey [12] applied to replication p-values also overestimates this proportion, but not as severely as the replication failure rate.

If the failed replication rate does not necessarily reflect the proportion of replications testing true null hypotheses, what can be said about this proportion using data from recent replication projects? To make some progress towards an answer, I apply the estimator from Storey [12] to the p-values from replications in recent replication projects [2,3]. While this estimator is still conservative (i.e. overestimates the proportion of true null hypotheses) it is less conservative compared to using the failed replication rate. In Figure 2, I show estimated proportions of false positives along with bootstrap 95% confidence intervals for three fields. For cognitive psycholoy (CP, grey) and experimental economics (EE, orange) the point estimate of the false positive rate is les than 25%, and the lower end of the CIs, which from the simulation above are often still quite conservative, are near 0. This suggests that in these fields it is not nessearily the case that many published findings are false positives, with the conservative estimates presented here generally suggesting that the majority of published research do not examine true null hypotheses. For social psychology (SP, blue), the point estimate for the false positive rate exceeds 75%, suggesting that there may be a concern about high false positive rates is this field. Although this result suggests that high false positive rates can not be ruled out in social psychology, the data does not definitively show this, because the estimator I use is conservative.

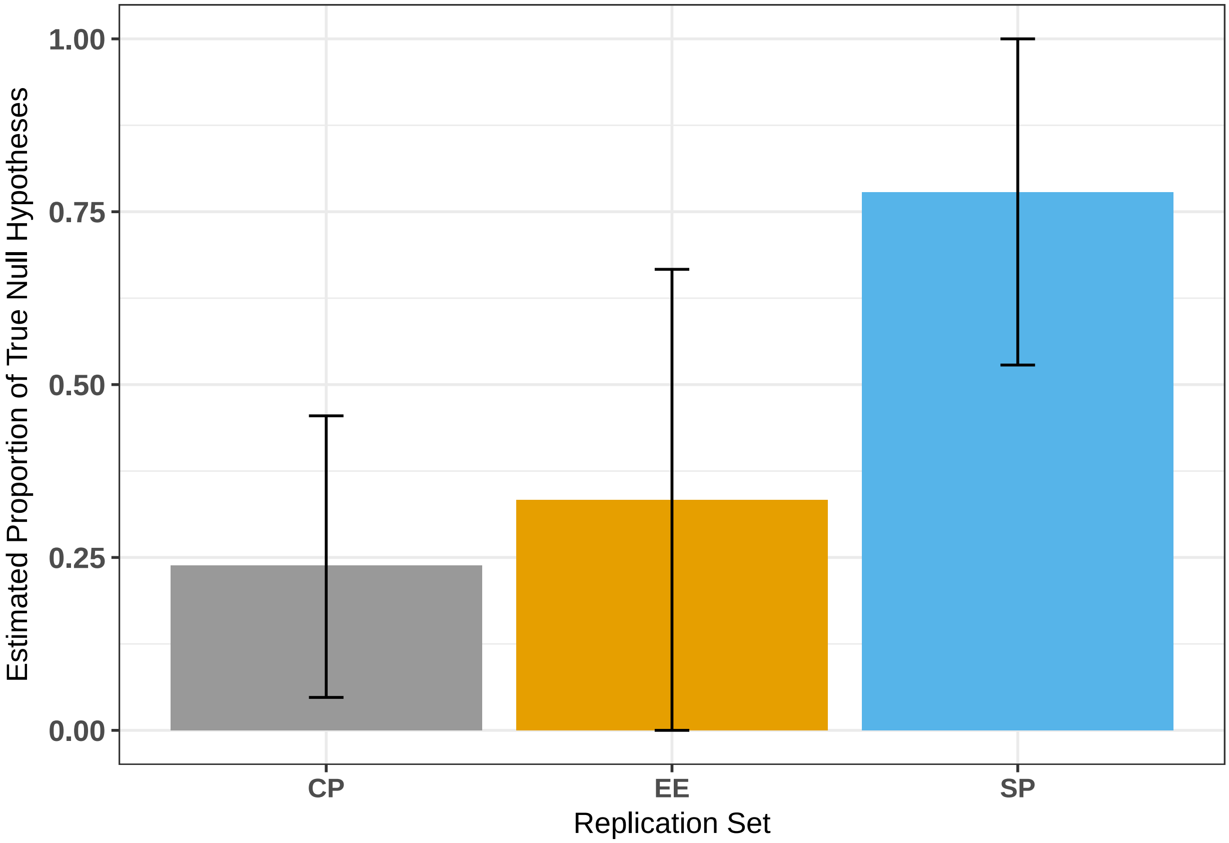

Figure 2: Point estimates and 95% boostrap confidence intervals (black whiskers) for the proportion of null hypotheses among replications in cognitive psychology (CP, grey), experimental economics (EE, orange), and social psychology (SP, blue)

The results I present in Figure 2 also suggest heterogeneity across fields, as has been observed in the Reproduciblity Project: Psychology [2] from which some of these data originate. It is worth noting that because these estimates can differ in their degree of conservativeness, this analysis does not necessarily show that these fields differ in their false positive rate *per se*, because it may be the case that the false positive rates across field are similar but that some fields have features that result in greater degrees of conservativeness. The analysis I perform here can not distinquish increased conservativeness from true differences in false positive rates.

Disscussion

The results that I present here show that high replication failure rates do not nessearily imply that most of the replicated studies are false. Conservative estimates suggests that in some fields most published results may be true. If most published research is true, then what explains failed replications? The simulations that I present here offer a simple model where the failed replication rate is high even though the false positive rate is low. Under this model, low power and publication bias result in published effect sizes overestimating true effects sizes. As a result, replications are under powered to detect small but non-null effects.

The possiblity that low replicability is caused by something other than high false positive rates has important implications for potential reforms to the scientific process. Many proposed changes are explicitly or implicitly premised on the idea that many studies are false positives, and therefore seek the improve the scientific process by decreasing the prevalence of false positives within the literature. For example, a recent proposal to change standards for considering a finding "statistically significant" [11] is based at least in part on the idea that such a change would descrease the likelihood of false positives. The possiblity that much of the literature is true potentially casts doubt on this justification, and the possiblity that lack of replicability is caused primarily by effect size inflation rather than false positives suggests that there is a risk that the proposal to lower p-value threshold could even be harmful if it results in increased publication bias.

The possiblity that most published effects are overestimated rather than being null also has many potentially implications for scientific practice and policy. Even for systematic changes that are widely viewed as beneficial, focusing on the underlying goals that change is meant to achieve can help guide changes are implemented. For example, replications aimed at weeding out false positives and replication aimed at accurately estimating effect sizes might be designed differently. The work that I present here is one step towards narrowing down what goals these types of reforms might aim for. Following [13], I wish to emphasis that the scientific process may legitimately aspire to multiple different goals, which may sometimes involve tradeoffs. It is tempting to view results such as those I have present here as needlessly blocking important changes. My aim isn't to block change, but rather to clarify these issues so that necessary changes can be designed optimally.

Methods

I performed all computational analyses in R [14], using ggplot2 [15] for visualization. I used to R package "qvalue" [16] to estimate proportions of true null hypotheses from p-values using the method of Storey [12], with "lambda" set to 0.5. I obtained p-values from several replication projects [2] to estimate true null hypothesis rates. I obtained p-values from the Reproducibility Project: Psychology [2,3] from publically available files on the Open Science Framework (https://osf.io/ezcuj/wiki/home/), using the package "RCurl" [17]. For data from the Reproducibility Project: Psychology I only include completed replications where the orginal publication reported a p-value less than 0.1. For the Experimental Economics Replication Project [3], I extracted replication p-values manually from Table S1 of [3]. Code to reproduce the analyses is this article are available at https://github.com/ScottWPiraino/false_finding_rate

References

1. Ioannidis JPA. Why Most Published Research Findings Are False. PLoS Medicine. 2005;2: e124. doi:10.1371/journal.pmed.0020124

2. Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015;349: aac4716–aac4716. doi:10.1126/science.aac4716

3. Camerer CF, Dreber A, Forsell E, Ho T-H, Huber J, Johannesson M, et al. Evaluating replicability of laboratory experiments in economics. Science. American Association for the

Advancement of Science; 2016;351: 1433–6. doi:10.1126/science.aaf0918

4. Jager LR, Leek JT. An estimate of the science-wise false discovery rate and application to the top medical literature. Biostatistics. 2014;15: 1–12. doi:10.1093/biostatistics/kxt007

5. Ioannidis JPA. Discussion: Why "An estimate of the science-wise false discovery rate and application to the top medical literature" is false. Biostatistics. 2014;15: 28–36. doi:10.1093/biostatistics/kxt036

6. Goodman SN. Discussion: An estimate of the science-wise false discovery rate and application to the top medical literature. Biostatistics. 2014;15: 23–27. doi:10.1093/biostatistics/kxt035

7. Gelman A, O'Rourke K. Discussion: Difficulties in making inferences about scientific truth from distributions of published p-values. Biostatistics. 2014;15: 18–23. doi:10.1093/biostatistics/kxt034

8. Cox DR. Discussion: Comment on a paper by Jager and Leek. Biostatistics. 2014;15: 16–18. doi:10.1093/biostatistics/kxt033

9. Benjamini Y, Hechtlinger Y. Discussion: An estimate of the science-wise false discovery rate and applications to top medical journals by Jager and Leek. Biostatistics. 2014;15: 13–16. doi:10.1093/biostatistics/kxt032

10. Schuemie MJ, Ryan PB, Suchard MA, Shahn Z, Madigan D. Discussion: An estimate of the science-wise false discovery rate and application to the top medical literature. Biostatistics. 2014;15: 36–39. doi:10.1093/biostatistics/kxt037

11. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, Berk R, et al. Redefine statistical significance. Nature Human Behaviour. Nature Publishing Group; 2017; 1. doi:10.1038/s41562-017-0189-z

12. Storey JD. A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology). Blackwell Publishers; 2002;64: 479–498.

doi:10.1111/1467-9868.00346

13. Finkel EJ, Eastwick PW, Reis HT. Replicability and other features of a high-quality science: Toward a balanced and empirical approach. Journal of Personality and Social Psychology. 2017;113: 244–253. doi:10.1037/pspi0000075

14. R Core Team. R: A Language and Environment for Statistical Computing. 2016.

15. Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2009.

16. Storey JD, Bass AJ, Dabney A, Robinson D. qvalue: Q-value estimation for false discovery rate control. 2015.

17. Lang DT, The CRAN Team. RCurl: General Network (HTTP/FTP/...) Client Interface for R. 2016.