To P or not to P? The Usefulness of P-values in Quantitative Political Science Research *

Vera E. Troeger†

March 12, 2019

*This contribution represents my own opinion and does not reflect the view of the European Political Science Association (EPSA) or Political Science Research and Methods (PSRM).

†Department of Economics, University of Warwick, Coventry, CV4 7AL, United Kingdom, e-mail: v.e.troeger@warwick.ac.uk

1 Introduction: The use and misuse of p-values

Star-gazing and p-hacking are just two of the commonly used pejorative descriptions of publication or favoured-hypothesis bias. The so-called replication crisis (Gelman 2011, Benjamin et al. 2018; Lakens et al. 2018; McShane et al. 2017; Tramow and Marks 2015; Nuzzo 2014) in quantitative social science research is often attributed to the (mis-)use of p-values when presenting inferential statistical results to empirically support a previously stated hypothesis or theoretical argument.

A quote from the infamous Political Science Rumors website exemplifies this problem: "Third-year AP here. Starting to realize that there is no way I can demonstrate a meaningful relationship between my two variables without manipulating P-values. Two questions: 1) Is this unethical? 2) What are the consequences if I get caught? At this point, I've sunk too much time into the project, so abandoning it simply isn't an option."

Recent research has shown that the distribution of presented p-values in published research significantly differs from that distribution in unpublished work (Gerber and Malhotra 2008a,b). In published empirical research p-values bunch up at the (arbitrarily) set α- value of 0.05 (Esarey and Wu 2016, Gerber and Malhotra 2008a,b, Gerber et al. 2001). This research finds that statistically significant results are overrepresented in academic articles. If significant results are consistently favoured in the review process, published empirical findings could systematically overstate the magnitude of the effects even under ideal conditions (Esarey and Wu 2016, Gerber and Malhotra 2008a,b, Gerber et. al 2001). Gerber and Malhotra (2008a) analyze empirical articles in the two leading political science journals, the American Political science Review (APSR) and the American Journal of Political Science (AJPS), and conclude that there is publication bias due to the reliance on the 0.05 significane level in empirical research. Geber et al. (2001), in addition, argue that to achieve statistical significance, the effect size must be larger in small samples. If published work is frequently biased against statistically insignificant findings, we should observe that the effect size reduces as sample sizes increase. And they show exactly this.

The new editor of the prime political methodology journal, Political Analysis, recently banned the usage of p-values and significance stars from articles published in PA (Gill 2017). This kicked loose a general debate about the usefulness of employing statistical hypothesis testing in general and presenting p-values as indication of statistical significance more specifically. This debate cannot be treated independently of a more general discussion of replicability, robustness and reproducibility of empirical research and ultimately academic misconduct.

After the American Statistical Association published their statement on the use of p-values (Wasserstein and Lazar 2016), I, as then editor-in-chief of the EPSA journal PSRM, initiated a debate with the editorial board about the use and mis-use of p-values. The debate concluded that p-values as such are not the problem, they provide more or less useful information for the consumer of scientific research. However, they cannot be used as sole criterion for the reliability, significance or economic/political relevance of the empirical findings. This information needs to be coupled with information on effect size, e.g. real world relevance of the empirical results, robustness of the estimates, as well as a discussion of coverage and potential effect heterogeneity. In combination these different sets of empirical information can paint a more complete picture of the credibility of the presented statistical results.

Certainly, t-tests and p-values are not more or less useful than providing confidence intervals or credibility intervals in Bayesian statistics. Bayesian statisticians argue that credibility intervals are more useful because they are generated by simulating the posterior distribution of the estimates. The underlying philosophy differs but Bayesians make equally strong assumptions about prior and posterior distributions that - if violated - have equally negative effects on inference. Gelman (2011) argues that so-called Bayesian hypothesis testing is just as bad as regular hypothesis testing.

In what follows, I will quickly present the logic of statistical inference and significance testing, discuss the implications of significance testing in linear models, and will then turn to the bigger question of what the profession can do to deal with academic misconduct, since p-value hacking is just a symptom of a larger credibility crisis.

2 The econometrics of p-values: hunting for inference

Inference - the potential to draw conclusions beyond the analysed data sample to the population - is one of the main goals of empirical analysis in the social sciences. Researchers want to know whether the relationships they find in the sample at hand can predict the relationships between the same variables but drawn from a different sample. What we are ultimately interested in are out-of-sample predictions.

Significance tests have been developed to answer exactly the question whether it is possible to generalize the regression results for the sample under observation to the universe of cases. However, for significant tests to produce reliable results a host of assumptions has to hold. In linear (OLS) regressions this set of underlying assumptions is called

full ideal conditions or Gauss-Markov assumptions [1]. These assumptions ensure that the data sample under observation matches the characteristics of the universe of cases or the so-called population. For this to work the researcher has to define the population. This is usually a theoretical question and harder than most applied researchers expect: To what set of cases does the formulated theory or theoretical argument apply? All countries over all periods of time? A set of countries over a defined time-span? All individuals across geographical entities, sex, age, time?

The underlying assumption for significance tests to produce reliable results, is that the sample is randomly drawn from the underlying population and thus mirrors all relevant characteristics of the universe of cases. All deviations are due to random sampling error. Gauss-Markov assumptions ensure that this is the case. If deviations from the population are non-random, the standard errors of the estimated coefficients are estimated incorrectly and the resulting significance tests are therefore wrong and lead to false conclusions.

Bayesian statisticians strongly criticise the assumption, underlying inferential statistical significance testing, that standard errors depict the sampling variation of the estimated coefficient, i.e. the distribution of all effects estimated with a large number of different randomly drawn samples. This criticism is fuelled by the observation that a) we often don't know what the actual population is from which we are drawing a sample, b) samples are often not randomly drawn even if Gauss-Markov assumptions hold, and c) we often cannot draw a sample from a population, especially when we analyse a fixed set of countries or other geographical identities. These issues are certainly present and affect inferential statistical analysis, however standard errors can be interpreted as the precision with which the relationship in the sample can be estimated. For example, they depict random noise whose source is not necessarily random sampling but random measurement error and others.

2.1 The T-test: a quick discussion

The t-test is the most commonly used significance test in linear OLS regression analysis. It tests whether the estimated coefficient is significantly different from zero, e.g. there is not effect of x - the right-hand-side variable - on y - the dependent variable. The Null-Hypothesis (H0) thus states that $\beta = 0$, where by β denotes the estimated effect of x on y. There are two variations, a one sided alternative (HA) with $\beta > 0$ or $\beta < 0$ or a two sided

[1] I will not re-iterate here the technical aspects of significance testing. This can be found in all popular textbooks, e.g. Wooldridge 2015.

alternative hypothesis with $\beta \neq 0$. the test statistic follows a student-t distribution under the Null-Hypothesis, if and only if all Gauss-Markov assumptions are met.

$$t_{(n-2)} = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

t is the critical value of the student t distribution for a specific number of observations n and a specific level of significance. This is the p-value. The level of significance in theory can be set by the researcher but in practice the convention in statistics and quantitative data analysis in general is a significance level of p= 5% , or 2.5% on each side of the t-distribution for a 2-sided t-test.

The p-value itself is an arbitrary number and but the stated convention has lead to the discussed problem of p-hacking, star-gazing and publication bias because the profession has been conditioned for decades to accept results that are significant on the 5% level. In order to combat this publication bias, several political scientists (Benjamin et al. 2018; Esarey 2017) suggested to lower the threshold for p-values to 0.005. However, in my opinion, a mechanical lowering of the accepted threshold will not solve the problem.

Why is this the case? P-values adjudicate the frequency with which the researcher allows his statistical analysis to make α or Type-I errors as compared to β or Type-II errors. Statistical testing adopts the legal philosophy "in dubio pro reo": to rather acquit the defendant even though s/he might be guilty than convict an innocent. In this sense, the statistical profession has decided that it is more important to avoid Type-I errors - wrongly rejecting the Null-hypothesis and conclude that there is a non-zero effect, than avoiding Type-II errors - wrongly accepting the null that the coefficient is zero. Whether this is reasonable for every single empirical analysis, remains debatable. Selecting p-values increases or decrease the probability of type I and Type-II errors. The smaller the significance level (0.05, 0.01) the lower the probability of making Type-I and the higher the probability of Type-II errors.

Under ideal conditions, the t-test has good statistical power. However, as most applied researchers understand, ideal conditions are just that and are frequently violated in real data analysis. It is therefore useful to discuss and question the mechanical convention of a p-value of 0.05. Since different set ups, different data types and samples meet these ideal conditions differently well, it does not seem helpful to set another static significance level that is lower to solve the problem of publication bias (Esarey 2017).

Researchers often know which of the Gauss-Markov assumptions are violated and how these violations affect the estimation of the standard errors and thus the significance tests. A multitude of solutions to these specification issues like robust standard errors,

such as clustering etc. controlling for (group) heteroscedasticity, serial correlation, spatial correlation amongst other issues, as well as small and non-normal sample corrections have been developed and are frequently employed by applied researchers. The problem, quite often, with manipulating the standard errors only is, that most violations of full ideal conditions affect the estimation of both the coefficients and standard errors. Just treating the standard errors might increase the potential for wrong inferences.

While these solutions go some way in reducing the potential for overestimating the statistical significance of effects, because they usually are more conservative estimates of standard errors, they do not necessarily solve the problem of p-hacking and preferred hypothesis bias. The incentives set by the profession, journals, and the research community remain untouched.

3 The bigger debate: academic misconduct

The debate about the mis-use of p-values in empirical research is intimately intertwined with the more recent debate on academic fraud and thus reproducibility, reliability, credibility, and robustness of published empirical findings. Why is there an incentive to engage in academic mis-conduct and risk the career? Like doping in sports, cheating allows to reach the goal (publications, citations, tenure, promotion) faster. With probability of detection low still very low, incentives for cheating remain high. But the costs are borne by honest academics both personally (competition) and as a profession (reputation).

DART (Data Access and Research Transparency) and COPE (Committee on Publication Ethics) initiatives help to raise awareness and define standards for replication and robustness. Many journals in political science have developed dedicated replication guidelines for empirical research and some of them have implemented in-house replication of quantitative Analysis (PSRM, PA, AJPS).

Yet, this doesn't seem to be enough. Academic research produces (positive) results that hinge on our credibility and reputation. We need to maintain this credibility and reputation by implementing self-control mechanisms that prevent academic fraud and misconduct. We cannot leave it to the (criminal) justice system, since the fraud of a few produces negative externalities for the whole profession.

It seems almost impossible to detect subtle kinds of fraud like p-hacking and non-robust empirical results through the typical peer review process, which is supposedly the main instrument of quality assurance in the academic profession. In most cases, authors dont have to provide their data to the reviewers. This often might even have good reasons

when data is original, sensitive, or even personalized. Yet, the peer review process only evaluates the plausibility of results, it assumes honesty.

What are the solutions? Banning p-values from articles doesn't seem to help much or it is only a drop on the hot stone of publication bias and academic fraud, since it only treats a symptom but not the disease itself. Raising the costs of mis-conduct is one way forward. Solutions have to increase the perceived probability of detection for the single researcher. Let discuss a few possibilities that come to mind, without claiming to be exhaustive.

Publishers can easily implement Plagiarism software into their online submission systems to screen articles and books for potential copying of existing work without proper citation. A few journals like PSRM have implemented this.

Since the incentives cannot be denied, researchers must bind themselves to the mast like Ulysses through pre-registration: Disciplines that are less affected by spectacular fraud seem to be leading. In Political Science the EGAP registry holds 1128 pre-registered research designs, as compared to only 80 in 2014. In economics the RCT Registry of the American Economic Association contains 2370 registered studies, as compared to 240 in 2014. Registration of research designs is exponentially increasing. This is a great development since registered experiments cannot be changed ex-post in order to adapt the design to the empirical results. However, not all studies lend themselves to pre-registration. Again editors have to step up and make pre-registration compulsory in order to make this practice the norm in the profession. Registration doesnt work, however, if researchers regard the experimentally generated data as private property which dont have to be published or made available to reviewers. In this case researchers can in principle remove cases that do not fit the argument.

Another potential measure is to make all data publicly available. Again many journals require data and code to be make available to the public before publication. But often there are no requirements whether source data has to be included. When source data is original, confidential, or personalized publication might not be possible or undesirable. However, new avenues to make this kind of data available for replication need to be explored.

Given that the collection of original data is time consuming, costly, and creates public goods for the discipline, data citation must be improved. Data are intellectual products for which citation should be required (Mooney 2011). This practice increases incentives for scholars to publish data because it will affect their citation count. Original data collection should also be valued more by the profession and our journals to make it both more attractive to collect but also to share data.

The DART initiative and leading journals and editors have institutionalized the publication of replication material. When it comes to replication journals and their editors are key because they set the standards for good practise in the profession. One way is to strengthen the review process with actual replication of empirical results. This might not be always feasible due to the reasons discussed above. That is why journals need to conduct their own replication analysis of accepted empirical studies, as several leading journals in the discipline now do (PSRM, AJPS, PA).

Replication of empirical results is a necessary but not sufficient condition for detecting and reducing misconduct. The example of the excel-spreadsheet mistakes of Rogoff and Reinhard, as well as the problem of how to treat missing values in the Piketty case show that simple replication of results will remain insufficient to prevent the publication of unreliable empirical findings. Robustness checks can close part of the gap. They have become increasingly standard in the social sciences. Robustness checks do not just replicate empirical results but take into account that researchers have to take many decisions about estimation and specification. Many published studies read as if the presented specification was the only plausible one. Robustness checks, however, assume that alternative specifications are no less plausible and test whether results and conclusions hold for alternative assumptions. The problem still remains that it is in the hands of the authors to decide which robustness and sensitivity checks to include. This implies the same logic as for p-hacking.

The problem that is faced by the profession is feasibility. Even if we could agree on a set of necessary robustness and sensitivity tests, the question remains who should be in charge of checking that these rules have been followed and at what stage of the publication process?

There is much to do. the profession, publishers and editors need to decide on joint policies with respect to replication and robustness and journals need to start accepting and publishing null findings and replication studies more. This also requires that the scientific community, publishers and journals need to provide the necessary resources to generate an infrastructure which increases the probability of detecting academic fraud, much more so than it is the case at present.

4 Conclusion

Researchers always have an incentive to select results that confirm their favoured hypotheses. No requirement for robustness and sensitivity checks, or banning of p-values can change this incentive. Unless the profession renders academic fraud more costly, in-

stils better norms of replicability and reproducibility, pre-registration of research designs not just for experimental studies, and encourages publication of none or negative findings. Banning p-values cannot and will not solve the replication crisis.

References

Benjamin, DJ, Berger, JO, Johannesson, M, Nosek,BA, Wagenmakers, EJ, Berk, R, Bollen, KA, Brembs, B, Brown, L, Camerer,C. et al. 2018. Redefine statistical significance. *Nature Human Behaviour* 2 (1): 6.

Berkson, J. 1942. Tests of Significance Considered as Evidence. textitJournal of the American Statistical Association 37 (219): 325335.

Chang, A, and Li, P. 2015. Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say Usually Not. *Finance and Economics Discussion Series Divisions*, `https://www.federalreserve.gov/econresdata/feds/2015/files/2015083pap.pdf`.

Esarey, Justin. 2017. Lowering the Threshold of Statistical Significance to p ¡ 0.005 to Encourage Enriched Theories of Politics. *The Political Methodologist* 24 (2): 1319.

Esary, J. and Wu, a. 2016. Measuring the effects of publication bias in political science. *Research & Politics*. July-Septmeber: 1-9.

Gelman, A. 2011. So-called Bayesian hypothesis testing is just as bad as regular hypothesis testing. Available at: `http://andrewgelman.com/2011/04/02/so-called_bayes/`.

Gerber, AS and Malhotra, N. 2008a. Do statistical reporting standards affect what is published? Publication bias in two leading political science journals. *Quarterly Journal of Political Science* 3(3): 313326.

Gerber, AS and Malhotra, N. 2008b. Sociological methods & publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociological Methods & Research* 37(3): 330.

Gerber, AS, Green, DP and Nickerson, D. 2001. Testing for publication bias in political science. *Political Analysis*: 385392.

Gill, J. 2018. Comments from the new Editor. *Political Analysis* 26 (1): 12.

Lakens, D, Grange, JA, Adolf, F, Albers, C, Anvari, F, Apps, M, Argamon, S, Baguley, T, Becker, R, Benning, S et al. 2018. Justify your alpha. *Nature Human Behavior*.

McShane, BB, Gal, D, Gelman, A, Robert, C, and Tackett, JL. 2017. Abandon statistical significance. arXiv preprint arXiv:1709.07588.

Mooney, H. 2011. Citing data sources in the social sciences: do authors do it? *Learned Publishing* 24:99108

Nuzzo, R. 2014. Statistical errors: P values, the "gold standard" of statistical validity, are not as reliable as many scientists assume. *Nature* 506 (7487): 150152.

Simmons, JP, Nelson, LD, and Simonsohn, U. 2011. False-positive psychology: Undisclosed Flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22 (11): 13591366.

Tramow, D, and Marks, M. 2015. Editorial. *Basic and Applied Social Psychology* 37 (1): 12.

Wasserstein, RL, and Lazar, NA. 2016. The ASAs Statement on p-Values: Context, Process, and Purpose. *The American Statistician* 70 (2): 129133.

Wooldridge, JM. 2015: Introductory Econometrics: A Modern Approach. South-Western, 6th ed.