The Use and Usefullness of $P$-values in Political Science: Intro

Daniel Bischof†
Mariken van der Velden.
27th May 2019

$P$-values are the most frequently employed metric to assess the significance of statistical findings in the social sciences. Since the earliest years of their usage the meaning and usefulness of $P$-values were topics of heated discussion (Fisher 1935; Berkson 1942). Lately the reproduction/replication crisis resuscitated this debate (Gelman 2018; Benjamin et al. 2018; Lakens et al. 2018; McShane et al. 2017; Trafimow and Marks 2015; Nuzzo 2014). Meanwhile, the skepticism has not stopped at the gates of political science. Most prominently the journal “Political Analysis” banned $P$-values “in regression tables or elsewhere” after the new editor took over the board of editors in 2017 (Gill 2018: 1).¹ Also political scientists contributed to a swelling debate suggesting to lower the threshold for $P$-values to 0.005 (Benjamin et al. 2018; Esarey 2017).

This special issue seeks to contribute to the debate on $P$-values by summarizing the main arguments of it, providing an encompassing discussion of $P$-values – also from an epistemological perspective – as well as advice for the discipline about the Do’s and Don’ts for $P$-values. In February 2018 the Department of Political Science at the University of Zurich invited several political methodologists to discuss the matter in a series of public lectures. The present contribution summarizes these public lectures but also goes beyond them by presenting the arguments two other distinguished colleagues.

Our introductory piece summarizes the discussion around $P$-values in the discipline and situates this into the larger debate on the replication crisis; Vera Träger discusses the logic

†Department of Political Science, University of Zurich (CH); corresponding author: bischof@ipz.uzh.ch.
¹Only to back-pedal later on and deciding to allow $P$-values in specific cases.
of statistical inference and significance testing and the implications of significance testing for empirical research; Susumu Shikano provides a detailed insight into two Bayesian approaches to hypothesis testing; Marco Steenbergen takes a stronger epistemological view on to $P$ or not to $P$; and finally, Simon Hug critically discusses the contributions of this issue by emphasizing that none of the discussed approaches appear to provide a “silver bullet” for the issues they seek to address.

A brief history of $P$

So, what are $P$-values and why do we as researchers care so much about them? $P$-values date back to the 18th century where Pierre Simon-Laplace came-up with first ad hoc definitions of $P$-values. Later on in the 20th century Pearson formalized them in his $\chi^2$ test and then Fisher popularized them by also proposing the threshold of 0.05 for statistical significance. The first references from Fisher on the threshold of 0.05 stems from his well-known “lady testing tea” experiment. Muriel Bristol, a phycologist and enthusiastic tea drinker, claimed to be able to differentiate tea which was poured on milk from milk which was poured on tea – of course keeping the amount of tea and milk constant.

Fisher and his friend William Roach decided to test Bristol’s tea tasting skills with a simple experiment: Muriel Bristol was provided eight cups of tea (four prepared by first adding milk; four prepared by first adding tea). Bristol then was asked to name the four cups prepared by her be-liked method. Thus, the null hypothesis was that Burial did not have the ability to distinguish the preparation of tea. Given $n=8$ cups and $k=4$ chosen cups the experiment results in 70 possible combinations. In order to reject the null hypothesis Fisher suggested that Bristol needed to get four out of four cups right. The combination of four correctly classified cups has a chance to occur in one out of 70 combinations. Bristol eventually got all eight cups correct.

Fisher discusses the threshold of 5% in close relation to the lady testing tea experiment. As outlined above Bristol’s performance had a chance to occur in only 1.4 per cent, while if she had missed only a single cup the chances to observe such a performance would have increased

$$\frac{8!}{4!(8-4)!} = 70$$
drastically to 24.3 per cent. In the latter case Fisher believed the likelihood of observing such a performance just by chance was too high. Future research built on Fisher’s reasoning – without him necessarily having had the intention to be used as the default approach to hypotheses testing – and eventually stopped discussing the reasons for the 5% threshold entirely. As this example illustrates, what $P$-values then really tell us is how likely our data are, assuming that our $H_0$ (Bristol not having the skills to tell the difference between the two tea preparation methods) is true (Wasserstein and Lazar 2016). A standard for empirical testing was born and until today this standard guides social scientists’ behavior, evaluations of research and most prominently publication standards.

But why have researchers recently ‘seen the light’ and payed attention to the shortcomings of $P$-values? In 2011, the renowned social-psychologist Diederik Stapel was found guilty of fraudulent research practices. As it turned out Stapel had faked his entire data collection. He simply answered to his questionnaires himself and thereby created the data he and his research team then analyzed. His research fraud sparked a larger debate within psychology: To what extend was there a culture of “sloppy” science, in which some scientists did not understand the essentials of statistics, reviewers for journals encouraged researchers to leave unwelcome data out of their papers, and even the most prestigious journals printed results that were obviously too good to be true? The Open Science Collaboration (2015) replicated 100 studies published in psychology journals. Using high-powered designs, they found that their mean effect size was approximately half of the size of the original articles. Moreover, while 97% of the original studies had demonstrated significant results ($p < 0.05$), only 47% of the replicated studies had significant results – indicating that 53% of the studies could not be replicated. This is not only a problem of psychology, where the norm is to publish based on experimental studies. In economics, also half of the studies could not be replicated (Chang and Li 2015). Chang and Li (2015) replicate 67 original articles published in 13 well-regarded macro-economics and general interest economic journals and demonstrate that replication issues are not tied to using

\[
\frac{16 + 1}{70} = 0.243
\]


experimental data, but equally apply to studies using publicly available data sets. In political science the debate caught fire with the Mike LaCour case. LaCour did not only follow “sloppy” research practices but committed fraud by inventing data he never had collected in the first place (Broockman, Kalla, and Aronow 2015).

These happenings suggest that practices of ‘P-hacking’ are more likely to occur in an environment focusing so much on the question of whether \( P < 0.05 \). They are then amplified by the human tendencies of apophenia - seeing patterns in random data - and of conformation bias - focusing on evidence that is in line with our (favored) explanation. These human tendencies are likely to affect how we walk through the ‘garden of forking paths’ when conducting analysis (Gelman and Loken 2013) and how we interact with the question of ‘researcher degrees of freedom’ (Simmons, Nelson, and Simonsohn 2011). And in many instances making the threshold is just one tiny step away – e.g. by adding/dropping a control, an interaction term or dropping some unfavorable outliers.

Thus, from our point of view not only the practices of how we engage and interpret \( P \)-values need to change, but eventually the environment under which we conduct research needs to adapt as well. As all contributions to this issue show: lowering the threshold for significance is unlikely to achieve this goal (Benjamin et al. 2018). A design-based derivation of the threshold might be better-equipped to achieve this goal (Lakens et al. 2018), but similar to the issue of the ‘garden of forking paths’ leaves researchers potentially with too many ‘degrees of freedom’. Proposals calling for a purely Bayesian approach to questions of significance tend to ignore that eventually we will run in very similar questions and issues irrespective if we choose a Bayesian or Frequentist perspective. Thresholds and guidelines are also the likely outcome if we go Bayesian – as Simon Hug’s discussion in this issue correctly points out.

Instead, we understand the replication/reproduction crisis as a symptom for a larger, systematic problem in the Social Sciences. This problem speaks to all aspects of what we are as a profession. It speaks to: how we teach empirical research practices, how we engage with changing practices in data sciences and how we question our own past and present behavior as scientists. But most importantly it suggests that no matter how we engage with \( P \)-values in the future as a profession, proposals for change need to take into account how much public-
References


Esarey, Justin. 2017. "Lowering the Threshold of Statistical Significance to $p < 0.005$ to Encourage Enriched Theories of Politics." *The Political Methodologist* 24 (2): 13–19.


Gelman, Andrew, and Eric Loken. 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." *Downloaded January* pp. 1–17.


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 \(\alpha\) 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 significance 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. 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 \( \beta \) 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 a or Type-I errors as compared to $\beta$ 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 doesn’t work, however, if researchers regard the experimentally generated data as private property which don’t 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-
stills 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


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.


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/](http://andrewgelman.com/2011/04/02/so-called_bayes/).


Nuzzo, R. 2014. Statistical errors: P values, the “gold standard” of statistical validity, are not as reliable as many scientists assume. *Nature* 506 (7487): 150-152.


Hypothesis testing in Bayesian framework

Susumu Shikano

February 2019

Abstract
While the Bayesian parameter estimation has gained a wider acknowledgement among political scientists, discussion about the Bayesian hypothesis testing is still not enough visible. This paper introduces two Bayesian approaches to hypothesis testing: one based on the Bayesian credible intervals and the other based on the Bayes factor. By using an example based on a linear regression model, I demonstrate similarities and differences not only between the null-hypothesis significance tests and Bayesian hypothesis tests, but also those among two different Bayesian approaches.

1 Introduction
Since decades, there have been always criticisms against the null-hypothesis significance test (NHTS) and in particular against the use of p-value. Some of them problematize that a not ignorable amount of researchers misuse and/or misinterpret the p-value (see the literature cited in the introductory paper of this special issue). The other kind of criticism is more fundamental that the basic logic behind the null-hypothesis significance test should be flawed (e.g. Gill, 1999). Most of such critics suggest Bayesian approaches as alternative to the NHST with p-value.

Thanks to multiple introductory books (e.g. Gill, 2002; Gelman and Hill, 2007; Jackman, 2009) and articles (e.g. Western and Jackman, 1994; Jackman, 2000) for political scientists, the Bayesian statistics has gained a wide acknowledgment in our discipline and many colleagues have become familiar with the basic concepts of Bayesian inference. While such discussion about Bayesian statistics focuses rather on parameter estimation techniques via Markov-Chain-Monte-Carlo, Bayesian hypothesis testing seems to be still not enough visible
at least in political science literature. Against this backdrop, this paper aims to introduce the hypothesis testing in the Bayesian framework and discuss its pros and cons.

This paper proceeds as follows: The next section briefly introduces the basic logic of Bayesian inference. It will be quite brief due to the space, therefore the interested readers are advised to consult the other introductory texts (e.g. Shikano, 2014; Lambert, 2018). In the subsequent sections, we will discuss two different approaches to hypothesis testing in the Bayesian framework. The first approach is more widely practiced procedure: we estimate the parameters of interest and evaluate whether the parameter value corresponding to the null-hypothesis falls in the credible interval. The second approach is based on Bayesian model selection. More specifically, it relies on Bayes factors. Recently a prominent paper advocated to adopt the threshold value $p = 0.5\%$ in NHTS (Benjamin et al., 2018). Their arguments are mainly based on the Bayes factor concept, therefore we should take a closer look at the Bayes factor. After discussing advantages and limits of both approaches, the last section concludes by discussion how we should deal with both Bayesian and NHTS approach.\footnote{To focus on hypothesis testing, this paper does not discuss the so-called Bayesian p-value in the context of posterior predictive checks. It is an important tool for model checking, however, a rather off-topic concerning hypothesis testing.}

## 2 The basic logic of Bayesian inference

Suppose we are interested in the effect of variable $X$ on $Y$ and the null-hypothesis says that there is no effect. In such situation, many political scientists would start with estimating the parameters of a linear regression model as follows:

$$y_i \sim \text{Normal}(\alpha + \beta x_i, \sigma^2) \quad (1)$$

Many first estimate the parameters by using ordinary least squares (OLS). Figure 1 presents the result based on an example data, which we use throughout this paper. Given such results, we conduct a t-test to assess whether the true value for $\beta$ equals zero. The p-value in such a t-test refers to the probability that we obtain the data if the null-hypothesis is true.\footnote{More exactly, p-values refer to the probability that the test statistics has the value based on the observed data or the more favorite value for the alternative hypothesis, given the null hypothesis is true.} This is not the probability that a hypothesis is true or false, but the probability of observed data.
Call:
```
lm(formula = Y ~ X)
```

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-2.43</td>
<td>-0.43</td>
<td>0.2</td>
<td>0.57</td>
<td>1.33</td>
</tr>
</tbody>
</table>

Coefficients:

|              | Estimate | Std. Error | t value | Pr(>|t|) |
|--------------|----------|------------|---------|---------|
| (Intercept)  | 3.93     | 0.088      | 44.57   | <2e-16  *** |
| X            | -0.26     | 0.123      | -2.09   | 0.038   *  |

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.84 on 184 degrees of freedom
Multiple R-squared: 0.023, Adjusted R-squared: 0.018
F-statistic: 4.35 on 1 and 184 DF, p-value: 0.038

Figure 1: A result of OLS-regression. The data comes from Shikano, Küntzler and Kim (2019), which investigates the effect of an information flyer about online platform for political participation on citizens’ internal efficacy levels.

In contrast, the Bayesian inference directly approaches the probability that a hypothesis is true or false. More precisely, researchers aim to obtain the probability of the unknown parameter (not only $\beta$, but also $\alpha$ and $\sigma^2$ in Equation 1) given observed data. Such kind of probability is called “posterior probability” and can be expressed as $p(\theta|y)$ with $\theta$ being unknown parameters and $y$ being data.

By applying the Bayes theorem, posterior probability can be obtained as follows:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{2}$$

$$= \frac{p(y|\theta)p(\theta)}{\int_{\theta} p(y|\theta)p(\theta) d\theta} \tag{3}$$

The p-value in NHTS corresponds rather to $p(y|\theta)$, which is called likelihood. The equation makes it clear that the posterior probability is not necessarily identical with the likelihood. $p(\theta)$ is the so-called prior, which refers to the probability how likely individual parameter values are before data analysis.
And the denominator of the right hand side of the equation is the normalizing constant, which ensures that the posterior probability has total probability of one.

As stated above, the goal of Bayesian inference is to obtain the posterior probabilities of unknown parameters and describe them. In the above example with a regression model, we are interested mostly in the posterior probability of $\beta$. The corresponding distributions are presented in Figure 2. Note that depending on the prior distribution, the posterior distributions’ locations are slightly different even though both are based on the same dataset. As a rule of thumb, an uncertain prior (i.e. flat distribution as in the left-hand side panel) has less impact on the posterior than the data, and vice versa.

![Figure 2: Marginal posterior distribution of $\beta$ (solid curve) based on two different prior distributions (dotted curve). For each model, a normal-inverse-gamma prior was used. Both priors have zero mean for $\alpha$ and $\beta$, but differ in their dispersion. The less dispersed prior (the right hand side panel) has the variance of 0.5 for $\alpha$ and $\beta$. See for more detail about prior specification the online appendix.](image)

Once we have obtained the posterior distribution of $\beta$, we can describe e.g.:

- the most likely value of $\beta$ is . . . (by using the mode of the posterior)
- the expected value of $\beta$ is . . . (by using the mean of the posterior)
- the most credible values of $\beta$ are . . . (by using the interval based on certain percentiles of the posterior)
One of the challenges in Bayesian inference is that it is not always easy to obtain the posterior probability. First, Equation 3 has an integral in the denominator. Further, in practical situations, we have in general multiple parameters (e.g. three parameters in the above regression example). Consequently, our posterior becomes multi-dimensional, which makes it more difficult to obtain the posterior probability. There are possibilities to obtain the posterior in analytical ways e.g. by using conjugate priors. For the sake of flexibility in modelling as well as convenience, most researchers rely on the Markov-Chain-Monte-Carlo techniques such as Gibbs sampling, Metropolis-Hasting algorithm.

3 Hypothesis testing via Bayesian parameter estimation

The above introduction only dealt with Bayesian parameter estimation. How can we decide about the null-hypothesis and alternative hypothesis given the posterior of the interested parameter is obtained? The widely practiced procedure is something like following:

1. Specify the parameter value or region, which correspond to the null-hypothesis (in the above example \( \beta = 0 \)).
2. Specify the level of credibilities for the decision (e.g. 95%).
3. Calculate the posterior and obtain the interval estimate for the level of credibilities.
4. Reject the null-hypothesis if the value/region specified in Step 1 is outside of the interval obtained in Step 3.

If we apply this rule to the posteriors in Figure 2, we can just compare the 95%-credible intervals with the dotted line at \( \beta = 0 \). Accordingly, we can reject the null hypothesis in both panels. Here, we are oriented toward the credible intervals, which cover 95% in the mid of posterior. Each of intervals are based on a t-distribution with a location and a scale parameter. Based on such symmetrical distributions, the intervals can also be conceived as highest probability density (HPD) regions, which have higher densities than outside of them. A credible interval is not always identical with the HPD region. In case of an asymmetric posterior distribution, the mid region of a credible interval is
unlikely to cover the higher densities than outside of it. For this reason, the HPD region is in general recommended (Box and Tiao, 1965), however it is more computationally demanding to identify the HPD region than the credible interval, which only needs its corresponding percentiles.

For the null-hypothesis with a point parameter value (such as $\beta = 0$), some researchers have additionally suggested to consider some interval around the parameter value. Spiegelhalter, Freedman and Parmar (1994) for example suggested “ranges of equivalence”, which should corresponds to the cost to take the alternative hypothesis. Also here, we can reject the null-hypothesis if the whole range is outside of the interval obtained from the posterior. Only in the case that the range of equivalence is partly included in the credible interval, neither null-hypothesis nor alternative hypothesis will be chosen (see also Kruschke, 2011).

3.1 Advantage and limits of decision based on Bayesian interval estimates

The most important advantage of the above procedure is its simple and intuitive implementation. Many researchers are surely familiar with a similar procedure in the non-Bayesian framework, in which researchers rely on confidence intervals. It is, however, tricky to interpret a frequentist confidence interval. Like the p-value, it does not describe the probability of the corresponding parameter since the true parameter value is assumed to be unique and has no distribution. Its probability is the chance that the interval has the true parameter value in its range. In contrast, interpretation of Bayesian credible intervals is much more intuitive and easy to understand because the credible intervals assume existence of the posterior distribution of the interested parameter. Therefore, the Bayesian credible intervals are just a description of the distribution of the interested parameters.

At the same time, there are also some limits in this approach. First, we have to decide for a certain level of credibility, which is always an arbitrary choice just like $p = 0.05$. Second, the interval estimates and therefore the decision depends always on the priors. Figure 2 showed that the different priors lead to two different posterior distributions, while they are based on the same data set. In the above example, both credible intervals are completely on the same side of $\beta = 0$. If we however use more informative prior in favor of $\beta = 0$ (i.e. prior with a smaller dispersion), the situation can be different. For
many researchers who wish clear-cut definite decision about hypotheses, such a situation is surely frustrating. Some advocate to use diffuse and non-informative priors (Box and Tiao, 1965), which however leads to the almost identical result with the frequentist parameter estimation.

4 Hypothesis testing as model selection by using Bayes factor

Differently from the above widely used approach based on the parameter estimation, we can now more directly approach to hypothesis testing and translate it into the Bayesian framework. Our starting point is that we regard hypothesis testing as model selection. That is, we set up two distinct models corresponding to the null and alternative hypothesis: \( M_0 \) for \( H_0 \) and \( M_1 \) for \( H_1 \).

In the above example with a regression model, we can set up two models as follows:

\[
M_0 : y_i \sim \text{Normal}(\alpha + \beta x_i, \sigma^2) \quad (4)
\]

\[
\beta = 0 \quad (5)
\]

\[
M_1 : y_i \sim \text{Normal}(\alpha + \beta x_i, \sigma^2) \quad (6)
\]

\[
\beta \neq 0 \quad (7)
\]

Here, the decision between two hypotheses is equivalent with that between two models. Just like in parameter estimation described above in Equation 2, we first start with prior probability about both models. If we have no idea prior to data analysis, we can start with even probabilities: \( p(M_0) = p(M_1) = 0.5 \).

Our goal is to obtain the posterior probability of both models: \( p(M_0|y) \) and \( p(M_1|y) \). They can be obtained by applying the Bayes theorem:

\[
p(M_0|y) = \frac{p(y|M_0)p(M_0)}{p(y)} = \frac{p(y|M_0)p(M_0)}{p(y|M_0)p(M_0) + p(y|M_1)p(M_1)} \quad (8)
\]

\[
p(M_1|y) = \frac{p(y|M_1)p(M_1)}{p(y)} = \frac{p(y|M_1)p(M_1)}{p(y|M_0)p(M_0) + p(y|M_1)p(M_1)} \quad (9)
\]

Table 1 shows the posterior of both models based on the same dataset and
two different priors in Figure 2. In the above analysis, we could reject the null hypothesis in the t-test above (1) and obtain 95% credible intervals completely on the negative side. In contrast, the posterior based on a diffuse prior, which should resemble to the likelihood-based non-Bayesian analysis, is over 25%, which clearly exceeds 5%.

<table>
<thead>
<tr>
<th>prior</th>
<th>( M_0 )</th>
<th>( M_1 )</th>
<th>odds ( \left( \frac{M_1}{M_0} \right) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>prior</td>
<td>.5000</td>
<td>.5000</td>
<td>1.0000</td>
</tr>
<tr>
<td>posterior (diffuse prior)</td>
<td>.2723</td>
<td>.7277</td>
<td>2.6724</td>
</tr>
<tr>
<td>posterior (more informative prior)</td>
<td>.8338</td>
<td>.1662</td>
<td>0.1993</td>
</tr>
</tbody>
</table>

Table 1: Prior and posterior probabilities of models corresponding to the alternative and null hypothesis. The data and priors for the unknown parameters are identical with those in Figure 2. The odds in the last column are prior and posterior odds. The posterior odds are identical with the Bayes factor as we discuss below.

This kind of discrepancies, known as Lindley-paradox (Lindley, 1957), is not surprising if we consider the difference between the NHTS and Bayesian approach. As stated above, the p-value in the OLS result is the probability of data given null-hypothesis. In contrast, we are focusing here the posterior probability of the Null-hypothesis given data. To obtain this, we also took explicitly the alternative hypothesis into account as can be found in the denominator of Equation 8. And we also have to note that the posterior probability of the alternative hypothesis is higher than the null hypothesis.

More disturbing result is actually the posterior based on the more informative prior. Here, the posterior of the null hypothesis is much higher than that of the alternative hypothesis, while the credible interval in the right-hand side panel in Figure 2 signaled the opposite. Before discussion this problem, we first introduce a new tool, Bayes factor, in the following section.

4.1 Bayes factor

Above, we compare two posterior probabilities with each other. Equivalently, we can also look at its their odds: \( \frac{p(M_1|y)}{p(M_0|y)} \). The odds larger one means that the posterior probability of the alternative hypothesis is higher than the opposite. From Equations 9 and 8, the odds can be described as follows:

---

3To obtain the posterior probabilities, \textit{stan} and an R-package \textit{bridgesampling} was used. We briefly discuss how we can obtain the posterior probabilities of both models below. The corresponding code appears in the online appendix.
\[ \frac{p(M_1|y)}{p(M_0|y)} = \frac{p(y|M_1)}{p(y|M_0)} \frac{p(M_1)}{p(M_0)} \]  \hspace{1cm} (10)

We call the first fraction of the right-hand side of the equation “Bayes factor”. That is, the posterior odds are the product of the prior odds and the Bayes factor. Analogously to the posterior odds, \( M_1 \) has more evidence than \( M_0 \) if \( BF > 1 \). Further, the larger \( BF \) is, the more evidence in favor of \( M_1 \). And in case of \( BF = 1 \), both models have the same level of evidence. Jeffreys suggested the following scale for interpretation as of Table 2:

<table>
<thead>
<tr>
<th>( BF_{10} )</th>
<th>Support for ( M_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 1</td>
<td>Negative</td>
</tr>
<tr>
<td>1 to 3</td>
<td>Barely worth mentioning</td>
</tr>
<tr>
<td>3 to 12</td>
<td>Positive</td>
</tr>
<tr>
<td>12 to 150</td>
<td>Strong</td>
</tr>
<tr>
<td>&gt; 150</td>
<td>Very strong</td>
</tr>
</tbody>
</table>

Table 2: Interpretation of Bayes factor

The important difference between the posterior odds and the Bayes factor is that the prior odds affect only the posterior odds. And only if the prior odds equal one, that is, we give the same probability to both models in the prior, the posterior odds and the Bayes factor are identical. For this reason, the posterior odds in Table 1 are identical with the Bayes factor. If we now apply the Jeffreys’ scale to our results, the alternative hypothesis has an evidence which is barely worth mentioning, if we use the diffuse prior. However, if we use the informative prior, the evidence is negative and in favor of the null hypothesis. The contradicting results due to the priors, therefore, remain. Below, we discuss where this sensitivities to the priors come from.

4.2 Sensitivities of the Bayes factor to priors

In the first sight, the Bayes factor may remind many readers of the likelihood ratio test, which the frequentist model selection often relies on. The likelihood ratio can be obtained as follows:

\[ LR_{10} = \frac{p(y|\theta = \hat{\theta}_{M_1})}{p(y|\theta = \hat{\theta}_{M_0})} \]  \hspace{1cm} (11)
This equation looks similar to the Bayes factor in Equation 10, but with an important difference. The likelihood ratio is based on the probability of data given specific parameter values $\hat{\theta}$. These parameters are those, which maximize the likelihood given a certain model.

In contrast, the Bayes factor seemingly contains neither $\theta$ nor $\hat{\theta}$. However, $\theta$ is hidden in the equation. To find them, we can reformulate Equation 2 corresponding to the context of model selection:

$$p(\theta_{M_0}|y, M_0) = \frac{p(y|\theta_{M_0}, M_0)p(\theta_{M_0}|M_0)}{p(y|M_0)}$$  \hspace{1cm} (12)

$$p(\theta_{M_1}|y, M_1) = \frac{p(y|\theta_{M_1}, M_1)p(\theta_{M_1}|M_1)}{p(y|M_1)}$$  \hspace{1cm} (13)

The Bayes factor is the odds of the denominators of these equations. From Equation 3, we know that the denominators are the normalizing constants, which can be obtained as follows:

$$p(y|M_0) = \int_{\Theta_{M_0}} p(y|\theta_{M_0}, M_0)p(\theta_{M_0}|M_0)d\theta_{M_0}$$  \hspace{1cm} (14)

$$p(y|M_1) = \int_{\Theta_{M_1}} p(y|\theta_{M_1}, M_1)p(\theta_{M_1}|M_1)d\theta_{M_1}$$  \hspace{1cm} (15)

They are called marginal likelihood since the possible parameter values are integrated out.

At this point, the difference between Bayes factors and likelihood ratio should be clear: First, Bayes factors take into account the parameter space ($\Theta_M$), while the likelihood ratio is solely based on the specific parameter values ($\hat{\theta}_M$). Second, in integrating out $\theta$, the priors ($p(\theta_{M_0}|M_0)$ and $p(\theta_{M_1}|M_1)$) play a crucial role. Here, we should not confuse the priors for the models ($M$) and the priors for their unknown parameters ($\theta$). As is shown in Equation 10, the Bayes factor is independent of prior odds, where the priors for the model are at stake. However, the Bayes factor is sensitive to the priors for the unknown parameters in each model. This sensitivity caused the discrepancy in the posterior odds, which was identical with the Bayes factor, in Table 1.
4.3 Advantage and limits of the Bayes factor

The main advantage of the Bayesian models selection is that Bayes factors can be an evidence in favor of the null-hypothesis as well as the alternative hypothesis. That seems to be more direct evidence on which we can rely on in hypothesis testing. Bayes factors are further known to be consistent. That is, if the sample size is infinite large, Bayes factors provide the correct evidence. Another additional advantage in terms of model selection is that marginal likelihood tends to be larger at a simpler model than at a more complex model if both models predict data to a similar degree (the so-called Occam’s Window).

At the same time, there are also some limits of this approach. First, this approach only compares two models which are selected by researchers. If both models are inadequate, the evidence is not useful at all. Second, once researchers rely on certain criteria to interpret and report Bayes factors e.g. based on Table 2, some arbitrary discontinuities are introduced in the continuous scale of Bayes factors, just like \( p = 0.05 \). Third, Bayes factors are sensitive to the prior choice concerning \( \theta \) (Kass, 1993). We have already seen above that the posterior distribution is relatively insensitive to different priors (Figure 2), the same set of priors lead to very different Bayes factors (Table 1). For this reason, there are also many Bayesian researchers, who are skeptical about Bayes factor (e.g. Lambert, 2018, 235-37). Fourth and probably most importantly for many applied researchers, computation of Bayes factors is not an easy exercise in particular due to the integrals in the enumerator and the denominator. In cases that one model is nested in the other model (just for the likelihood-ratio-test), there is an analytical solution known as Savage-Dickey-Method (Dickey and Lientz, 1970; Dickey, 1971). In other cases, we rely on the numerical approaches, e.g. naive Monte Carlo methods, the product space method (Carlin and Chib, 1995; Lodewyckx et al., 2011), bridge sampling (Gronau et al., 2017). A further possibility is the approximation by using the Bayesian information criterion (Raftery, 1995; Wagenmakers, 2007):

\[
BF_{10} = \exp \left( \frac{BIC(M_0) - BIC(M_1)}{2} \right)
\]  

(16)

One advantage is that researchers do not have to specify any priors, to which the Bayes factor is sensitive. However, BIC also implies some assumption about prior, which tends to make the null-hypothesis more plausible.
5 Is Bayesian inference a substitute for NHST and p-value?

This paper did not aim to advocate for replacement of NHST/p-values with Bayesian inference. Both approaches are based on different philosophies. In particular, their approaches differ in how we should deal with uncertainty, which is inherent in any kinds of inferences. In selecting an approach, researchers should be aware about the difference.

In the NHST framework, we have two clearly defined options: an alternative hypotheses and its opposite null-hypothesis. In this context, p-value is the chance that we make the type-I error in a long run if we would repeat our study with an identical design for many times. Therefore, a decision for or against the null-hypothesis is inherent in this process. Further, the NHTS in its basic form does not presuppose such sequential and cumulative processes, in which data collection and data analysis are conducted successively (see however as exception sequential testing suggested by Wald, 1945).

In Bayesian inference, we primarily aim to update our prior belief about unknown parameters and/or models through data analysis. Our belief after the data analysis, posterior belief, should have less uncertainty. From Equation 10, we can clearly see that the prior odds is updated by the Bayes factor to the posterior odds. This process does not require to decide between two different hypotheses. The Bayes factor only serves to report the relative evidence for both hypotheses. In Bayesian approach, we can even extend the idea of Bayes factors for more than two models. For example, Bayesian model averaging can be conceived as generalization of Bayes factors. Here, multiple models are compared by their marginal likelihood, but none of them will be rejected, but they are aggregated based on their marginal likelihood as weight.

Having discussed both NHTS and Bayesian approaches, a few final remarks should be in order. Most importantly, I like to emphasize that the hypothesis testing is not the only way to make inference and its value should not be overstated. From this point of view, I share the sceptism on the Bayes factor with some Bayesian researchers, who may also be skeptical due to the other reasons (e.g. sensitivity to priors). Further, the proposal of a more strict threshold \( p = 0.5\% \) based on the Bayes factor, as suggested by (Benjamin et al., 2018) is superfluous given the above differences in NHTS and Bayesianism. Instead to sticking to the Bayes factor and hypothesis testing, we should better turn to a
wider variety of Bayesian tools.

References


Null hypothesis significance testing (NHST) is under attack and \( p \)-values, long seen as the quintessence of the scientific method, are receiving particularly bad press. The criticisms are not new (Berkson, 1942; Rozeboom, 1960) and they cover a variety of fields (Kline, 2013; Wasserstein and Lazar, 2016; Ziliak and McCloskey, 2008). They cover a variety of aspects, including misunderstandings of \( p \)-values (Goodman, 2008) and what one might call the .05 fetish (Benjamin et al., 2018; Yates, 1951). One other aspect of the criticisms is that NHST has contributed to black-and-white thinking about scientific findings. In this paper, I focus on this aspect and discuss some novel and not-so-novel approaches that permit for a more nuanced interpretation of \( p \)-values.

The \( p \)-Value Dichotomy

In scientific practice, \( p \)-values have come to be used as an arbiter of sorts—a statistical criterion resulting in a binary decision (for a rare exception, see Cox and Donnelly, 2011). In its mildest form, the resulting dichotomy distinguishes between more and less reliable effects. More often than not, however, \( p \)-value-based dichotomies go much further. Thus, \( p \)-values are used to decide between interesting and uninteresting findings, worthwhile and worthless experiments (in the broadest sense of that term), publishable and unpublishable findings, and, ultimately, truth and falsehood. (The latter represents a particularly egregious misunderstanding of \( p \)-values).

There are numerous problems with this practice. First, it is questionable whether \( p \)-values were ever intended to play this role. In itself, the \( p \)-value is a continuum that captures the consistency of empirical evidence with the null hypothesis. It became linked to a binary decision in the Neyman-Pearson framework on hypothesis testing, in particular (Neyman and Pearson, 1928a,b), although Fisher (1925) also contributed to it. It should
not be forgotten, however, that Neyman and Pearson conceived their approach in terms of repeated experiments and not a single study, as is characteristic of much political science. In the latter context, the wisdom of reaching a dichotomous decision based on some significance level is questionable.

Second, the dichotomous use of \( p \)-values incentivizes scholars to engage in less than desirable scientific practices in order to reach significance. HARKing (hypothesizing after results are known) and \( p \)-hacking are textbook examples of practices that hamper reproducible science (Munafò et al., 2017). They do not contribute to a robust scientific canon and, indeed, can hurt in a world in which science deniers have gained a strong political voice.

Third, the practice results in the well-known publication bias (Munafò et al., 2017). Significant findings make it into the scientific discourse, whereas non-significant findings, no matter how valuable, land in filing drawers never to see the light of day. The baseline rate of failed experiments thus remains hidden, making it all the more difficult to assess the replicability of research.

I am certainly not the first one to lament the dichotomous use of \( p \)-values (see Goodman, 1999a; Stern, 2016). But the question is how to change scientific practice. One could abandon \( p \)-values altogether, as Political Analysis did for a short while and other journals still do. However, this may be throwing out the baby with the bathwater, at least if one agrees that the problem may not so much be \( p \)-values themselves as their use in scientific practice. One could also lower \( p \)-values but this might just shift the threshold at which dichotomous decisions are made (Benjamin et al., 2018).

My modest proposal is not to abandon \( p \)-values wholesale, which may anyway be difficult in light of the generations of scholars who have learnt (and often come to love) \( p \)-values. Rather, my plea is to move away from its dichotomous use. Fortunately, there exist both older and newer techniques that allow us to take a more nuanced view of \( p \)-values. Here, I focus on the Bayes factor (Jeffreys, 1998; Kass, 1993; Kass and Raftery, 1995) and the analysis of credibility (Matthews, 2018), which both derive from Bayesian inference.

The Bayesian View of Inference

Shikano (this issue) provides an extensive discussion of the Bayesian perspective on hypothesis testing. Thus, it suffices to highlight a few aspects that we shall need to develop the Bayes factor and analysis of credibility. In general, the Bayesian approach can be summarized as follows (e.g., Jackman, 2011):

\[
\text{Prior + Data} \Rightarrow \text{Posterior} \tag{1}
\]

Work on the Bayes factor emphasizes the data aspect, whereas the analysis of credibility focuses on the prior.
In NHST à la Neyman-Pearson, we focus on a pair of hypotheses, $H_0$ and $H_1$. When testing means across treatment (indexed by 1) and control (indexed by 0) groups we may, for example, test $H_0 : \mu_1 = \mu_0$ against $\mu_1 \neq \mu_0$. We can now ask, a posteriori how much support do $H_0$ and $H_1$ receive. A standard result gives the posterior as (Jeffreys, 1998; Kass, 1993; Kass and Raftery, 1995)

$$\frac{p(H_0|\text{Data})}{p(H_1|\text{Data})} = BF_{01}(\text{Data}) \cdot \frac{p(H_0)}{p(H_1)}$$

(2)

The left-hand side gives the relative beliefs in $H_0$ and $H_1$ after the experimental evidence has been collected. The second term on the right-hand side gives the beliefs in $H_0$ and $H_1$ prior to conducting the experiment. Finally, BF is the Bayes factor. This is the ratio of the marginal likelihoods for $H_0$ and $H_1$ and, as such, captures the empirical evidence vis-à-vis both hypotheses—what some have called the “weight of the evidence” (Good, 1950). It influences whether and how beliefs about the hypotheses should be adjusted.

Many test quantities asymptotically follow a normal distribution. This is true of means, differences in means, and log-odds ratios, for example. Imagine we have a normally distributed quantity $Y$ with a known variance of $\sigma^2$. We are interested in the posterior distribution of the mean, $\mu$, of $Y$. With a normal prior $\mu \sim \mathcal{N}(\mu_0, \phi_0)$, the posterior is given by

$$\mu|\text{Data} \sim \mathcal{N}(\mu_P, \phi_P)$$

$$\mu_P = \phi_P \left[ \frac{\mu_0}{\phi_0} + \frac{\bar{y}}{\phi} \right]$$

$$\phi_P = \left[ \phi_0^{-1} + \frac{1}{\phi} \right]^{-1}$$

(3)

Here, $\phi = \sigma^2/n$ is the sampling variance of the sample mean, $\bar{y}$. We can simplify the posterior by dividing by $\phi_P$ so that $\mu_P/\phi_P = \mu_0/\phi_0 + \bar{y}/\phi$ and $1/\Phi_P = 1/\phi_0 + 1/\Phi$. These expressions play an important role in the analysis of credibility (Matthews, 2018).

The Bayes Factor

One of the major problems with NHST is that it cannot say anything about the relative merits of $H_0$ and $H_1$. After all, the $p$-value is computed under the assumption that $H_0$ is true (cf. Stern, 2016). The Bayes factor rectifies this problem. It is given by (Jeffreys, 1998; Kass, 1993; Kass and Raftery, 1995)

$$BF_{01}(\text{Data}) = \frac{\int p(\text{Data}|\theta, H_0)p(\theta|H_0)d\theta}{\int p(\text{Data}|\theta, H_1)p(\theta|H_1)d\theta}$$

(4)
By focusing on the parameters $\theta$, which drive the marginal likelihoods on the right-hand side of Equation 4, we shift our focus from hypothesis testing to estimation. Note that we can also state the Bayes factor as the weight of evidence in favor of $H_1$: $\text{BF}_{10}(\text{Data}) = 1/\text{BF}_{01}(\text{Data})$.

If our concern is the dichotomous interpretation of $p$-values, then the Bayes factor already offers a clear advantage because it typically is classified in a more nuanced manner. Table 1 shows four different classifications of the Bayes factor ([Goodman 1999b; Held and Ott, 2016; Jeffreys 1998; Kass and Raftery 1995]). We see that there is more nuance in the manner in which evidence is interpreted. If dichotomous interpretations of the $p$-value are black and white, then the interpretation of the Bayes factor is in terms of shades of grey. We are forced to think about the empirical merits of the null and alternative hypotheses in a nuanced manner. Obviously, this is not the decisive selling point of the Bayes factor but it certainly helps to put empirical evidence against the null hypothesis into some perspective.

One could raise two objections at this point. First, there hardly seems to be any consensus on the classification of the Bayes factor. Sure, it may be nuanced but no one seems to agree on what those nuances are. I do not view this as a particularly troublesome argument. The different classifications vary on how much evidence they require against $H_0$ in order to upgrade the verbal designation of that evidence. An author or journal could settle on a particular scheme and take it from there. I will be using the scheme of [Held and Ott, 2016].

A second objection is that the Bayes factor is notoriously difficult to compute in most cases (see [Kass, 1993]). They also require choices about priors that may require expertise beyond the typical skill level among political scientists (again, see [Kass, 1993]). How can such a difficult technique ever be a practical replacement for the $p$-value?

In fact, it does not have to be. In the context of precise null hypotheses (e.g., $\mu_1 = \mu_0$), [Berger and Delampady, 1987] showed that it is possible to derive lower bounds on the Bayes factor from $p$-values. [Vovk, 1993; Sellke, Bayarri and Berger, 2001; and Held and Ott, 2018] formally derived the lower-bounds under different assumptions about the priors. These lower-bounds allow us to establish a link between $p$-values and the more nuanced interpretation that flows from Bayes factors.

The starting point for all of the approaches is to assume that the $p$-values a priori follow the uniform distribution $U(0, 1)$ under $H_0$. $H_1$ a priori favors low $p$-values, which can be captured by selecting an appropriate beta distribution. [Vovk, 1993] and [Sellke, Bayarri and Berger, 2001] assume $p \sim \text{BE}(\alpha, 1)$, where $\alpha \leq 1$. The resulting priors are generally uninformative and produce

$$\min \text{BF}_{01}(p) = \begin{cases} -ep \ln p & \text{if } p < 1/e \\ 1 & \text{Otherwise} \end{cases}$$

A graphic depiction is shown in the left panel of Figure 1. Arguing that the selected beta
Table 1: Classifying the Bayes Factor

<table>
<thead>
<tr>
<th>Range</th>
<th>Goodman Meaning</th>
<th>Held &amp; Ott Range</th>
<th>Held &amp; Ott Meaning</th>
<th>Jeffreys Range</th>
<th>Jeffreys Meaning</th>
<th>Kass &amp; Raftery Range</th>
<th>Kass &amp; Raftery Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 to 1/5</td>
<td>Weak</td>
<td>1 to 1/3</td>
<td>Weak</td>
<td>&gt; 1</td>
<td>Negative</td>
<td>1 to 1/3</td>
<td>Not worth more than a bar mention</td>
</tr>
<tr>
<td>1/5 to 1/10</td>
<td>Moderate</td>
<td>1/3 to 1/10</td>
<td>Moderate</td>
<td>1 to ≈ 32/100</td>
<td>Barely worth mentioning</td>
<td>1/3 to 1/20</td>
<td>Positive</td>
</tr>
<tr>
<td>1/10 to 1/20</td>
<td>Moderate to strong</td>
<td>1/10 to 1/30</td>
<td>Substantial</td>
<td>≈ 32/100 to 1/10</td>
<td>Substantial</td>
<td>1/20 to 1/150</td>
<td>Strong</td>
</tr>
<tr>
<td>1/20 to 1/100</td>
<td>Strong to very strong</td>
<td>1/30 to 1/100</td>
<td>Strong</td>
<td>1/10 to ≈ 3/100</td>
<td>Strong</td>
<td>&gt; 1/150</td>
<td>Very strong</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1/100 to 1/300</td>
<td>Very strong</td>
<td>≈ 3/100 to 1/100</td>
<td>Very strong</td>
<td>≈ 3/100 to 1/100</td>
<td>Very strong</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt; 1/300</td>
<td>Decisive</td>
<td>&lt; 1/100</td>
<td>Decisive</td>
<td>&lt; 1/100</td>
<td>Decisive</td>
</tr>
</tbody>
</table>

**Notes:** Table shows the support for $H_0$ relative to $H_1$ [Goodman 1999](#), [Held and Ott 2016](#), [Jeffreys 1998](#), [Kass and Raftery 1995](#).
prior is not always so uninformative, favor $\mathcal{BE}(1, \beta)$ with $\beta \geq 1$. This results in

$$\min BF_{01}(p) = \begin{cases} -e(1 - p) \ln(1 - p) & \text{if } p < 1 - 1/e \\ 1 & \text{Otherwise} \end{cases}$$

(6)

Since $\ln(1 - p) \approx -p$, we may also write $\min BF_{01}(p) = ep$. This is shown in the right panel of Figure 1. In the range between 0 and 0.5, the Held-Ott formula consistently produces more conservative estimates of the minimal Bayes factor than the Vovk formula. Indeed, to date offer the most conservative minimal Bayes factor for $p$-values below 0.50.

Having defined the minimum Bayes factors in this manner, how could one offer a nuanced interpretation of a $p$-value? Table 2 classifies the Bayes factor according to $\beta$. It also shows the approximate ranges of $p$-values that fall into a particular class according to the Held-Ott formula. If we look at this table, we see that the conventional 0.05 significance level at best offers only moderate evidence against the null hypothesis. The minimum Bayes factor is 0.132, meaning that the evidence favors $H_1$ only by a factor of $1/0.132 \approx 7.5$. By contrast, the 0.005 significance level favored by Benjamin et al. (2018) yields strong support against $H_0$, with a minimum Bayes factor of $\min BF_{01} = 0.014$ or $\min BF_{10} = 73.761$.

Other cutoffs that we see in political science are 0.10 (sometimes indicated by a plus

### Table 2: The Weight of Evidence Contained in Different $p$-Values

<table>
<thead>
<tr>
<th>Designation</th>
<th>$p$-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decisive</td>
<td>0.000-0.001</td>
</tr>
<tr>
<td>Very Strong</td>
<td>0.001-0.004</td>
</tr>
<tr>
<td>Strong</td>
<td>0.004-0.012</td>
</tr>
<tr>
<td>Substantial</td>
<td>0.012-0.037</td>
</tr>
<tr>
<td>Moderate</td>
<td>0.037-0.132</td>
</tr>
<tr>
<td>Weak</td>
<td>0.132-1.000</td>
</tr>
</tbody>
</table>
Table 3: Alternative Presentation of Empirical Results

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Est</th>
<th>SE</th>
<th>p</th>
<th>Evidence Against Null</th>
<th>Maximum Weight Favoring Alternative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left-Right</td>
<td>-0.629</td>
<td>0.066</td>
<td>0.000</td>
<td>Decisive</td>
<td>∞</td>
</tr>
<tr>
<td>Left-Right Squared</td>
<td>0.005</td>
<td>0.006</td>
<td>0.405</td>
<td>Weak</td>
<td>1.191</td>
</tr>
<tr>
<td>Economy</td>
<td>0.315</td>
<td>0.078</td>
<td>0.000</td>
<td>Decisive</td>
<td>6837.722</td>
</tr>
<tr>
<td>Protestant</td>
<td>0.265</td>
<td>0.216</td>
<td>0.733</td>
<td>Weak</td>
<td>1.000</td>
</tr>
<tr>
<td>Catholic</td>
<td>-0.073</td>
<td>0.214</td>
<td>1.191</td>
<td>Weak</td>
<td>1.159</td>
</tr>
<tr>
<td>Secular</td>
<td>0.181</td>
<td>0.226</td>
<td>0.423</td>
<td>Weak</td>
<td>1.000</td>
</tr>
<tr>
<td>Education</td>
<td>0.142</td>
<td>0.016</td>
<td>0.406</td>
<td>Weak</td>
<td>1.000</td>
</tr>
<tr>
<td>Income</td>
<td>0.008</td>
<td>0.020</td>
<td>0.689</td>
<td>Weak</td>
<td>1.000</td>
</tr>
<tr>
<td>Male</td>
<td>-0.786</td>
<td>0.100</td>
<td>0.000</td>
<td>Decisive</td>
<td>∞</td>
</tr>
<tr>
<td>Age</td>
<td>0.025</td>
<td>0.017</td>
<td>0.141</td>
<td>Weak</td>
<td>2.810</td>
</tr>
<tr>
<td>Age Squared</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>Weak</td>
<td>1.000</td>
</tr>
<tr>
<td>Urban</td>
<td>0.262</td>
<td>0.152</td>
<td>0.085</td>
<td>Moderate</td>
<td>4.538</td>
</tr>
<tr>
<td>German</td>
<td>0.377</td>
<td>0.325</td>
<td>0.246</td>
<td>Weak</td>
<td>1.728</td>
</tr>
<tr>
<td>French</td>
<td>0.701</td>
<td>0.344</td>
<td>0.042</td>
<td>Moderate</td>
<td>9.040</td>
</tr>
<tr>
<td>Constant</td>
<td>6.356</td>
<td>0.584</td>
<td>0.000</td>
<td>Decisive</td>
<td>∞</td>
</tr>
</tbody>
</table>

Notes: Based on Steenbergen (2010, p. 419, results for the Greens). Evidence Against Null shows the classification of p-values that follows from Held and Ott (2016) and shows, in qualitative terms, how much evidence there is against the null hypothesis of no effect. Maximum weight favoring the alternative gives \( \max BF_{10}(p) \). It shows the maximal factor by which prior beliefs about the hypotheses should be updated in favor of \( H_1 \). The symbol \( \infty \) means that the Bayes factor approaches infinity.

Symbol in tables) and 0.01 (sometimes indicated with two stars). Like the 0.05 threshold, 0.10 qualifies only as moderate support against \( H_0 \). The minimum Bayes factor is only 0.258, however, meaning that the evidence favors \( H_1 \) only by a factor of roughly 3.9, about half of what 0.05 can offer. A significance level of 0.01 again qualifies as strong evidence against \( H_0 \) with \( BF_{10}(p) \approx 37 \).

To show the reader how one might use the Bayes factor in a publication, I replicate Table 5 from Steenbergen (2010). This table lists results from a hierarchical linear model of vote propensities for left parties in Switzerland. For the sake of simplicity, I only report the fixed effects estimates and to save space, I only report the results for one party: the Greens. Table 3 shows the replicated results with the kind of annotation one can derive from the p-based Bayes factor.

The original table contained five predictors for which \( p < 0.05 \): left-right self-placements, economic retrospections (economy), education, male, and domicile in the French-speaking part of Switzerland. As we can see in Table 3, however, the cluster of “significant” effects is quite heterogeneous. For four of the predictors, we have decisive evidence against the null hypothesis of a null effect. In one instance, the evidence against the null is less compelling,
receiving the moniker of moderate. These nuances are also reflected in the Bayes factors, which range from about 9 (for the variable French) to values approaching infinity. Such differences matter. After conducting the research, I am much less convinced of the divide between the French-speaking part and the rest of Switzerland (max BF$_{10} \approx 18$) than a gender gap the divide between urban and rural areas, even when both divides are significant at the .05-level.

There is nuance, too, among the non-significant findings. Some can still be classified as moderate evidence against $H_0$, whereas in other cases the evidence is weak. For example, the urban-rural divide in Green party support fails to be statistically significant at the .05-level. Still, the Bayes factor here is clearly more suggestive of an effect than, for example, the divide between German-speaking and other cantons. It is useful to communicate such nuances because there is information even in non-significant findings.

**The Analysis of Credibility**

The idea that there is information in non-significant findings plays an important role in the analysis of credibility (AnCred; Matthews, 2018). AnCred effectively asks what kind of prior would be needed to challenge a significant finding a posteriori or to make a non-significant finding a posteriori significant. The answer flowing from such an exercise can then be evaluated in terms of its plausibility, for example, in the light of published results. The beauty of the approach is that it only requires knowledge of the confidence intervals, which many believe should be published anyway.

Consider again Equation (1). We create a posterior distribution such that the a posteriori $100 \cdot (1 - \alpha)$ credible interval barely includes a null effect. On the left-hand side of the equation, the information in the data are given by the confidence interval. Following Good (1950), we now work backwards to derive the prior. Specifically, we derive the prior credible interval (CPI) that would be needed to generate the posterior.

Obviously any significant finding can be rendered a posteriori void by setting a prior with (almost) all of its probability mass located at 0. Any non-significant finding can be rendered a posterior significant by putting the mass at a value different from 0. In both cases, however, we effectively rule out that the data could ever cause us to revise our beliefs concerning the null hypothesis. This would not constitute a fair-minded—one could also say, open-minded—challenge to the empirical evidence. As Matthews (2018) argues, the key is to launch fair-minded challenges and advocacies, i.e., ones that leave room for the data.

So what would a fair-minded challenge look like? The starting point is that we have obtained a statistically significant result. Specifically, given a significance level of $\alpha$, we have obtained a $100 \cdot (1 - \alpha)$ percent confidence interval, $(L, U)$ that excludes 0. Here, $L = \bar{y} - z_{\alpha/2} \sqrt{\sigma}$, $U = \bar{y} + z_{\alpha/2} \sqrt{\sigma}$, and $z_{\alpha/2}$ is the critical value of the standard normal
distribution corresponding to $\alpha$. We posit ourselves as skeptics, meaning that we do not believe in an effect a priori. Hence, under normality, our prior is $\mu \sim \mathcal{N}(0, \phi_0)$. The corresponding CPI is $(-SL, SL)$, where $SL = z_{\alpha/2}/\sqrt{\phi_0}$ is the skepticism limit.

The key is now to derive $\phi_0 > 0$. We define the posterior so that the $100 \cdot (1 - \alpha)$ credible interval barely includes 0. This implies $\mu_P - z_{\alpha/2}/\sqrt{\phi_P} \geq 0$ or $\mu_P \geq z_{\alpha/2}/\sqrt{\phi_P}$. We can now use equation (3) to solve for $\phi_0$. Substituting the implied values for $\mu_0$ and $\mu_P$, as well as the empirical estimates of $\bar{y}$ and $\phi$, it is easily shown that

$$\phi_0 = \frac{(U - L)^4}{16 \cdot z_{\alpha/2}^2 \cdot U \cdot L}$$

Consequently,

$$SL = \frac{(U - L)^2}{4 \cdot \sqrt{U \cdot L}}$$

Let us apply the idea of skepticism to one of the results in Table 3. Let us distinguish between two groups: French-speaking (F) versus Italian-speaking (I) voters. Our null hypothesis is that the two groups do not differ in their proclivity to vote for the Greens: $H_0 : \mu_F - \mu_I = 0$. The estimate of the mean difference is 0.701, has a standard error of 0.344, and is statistically significant at the .05-level. The 95 percent confidence interval runs from 0.027 to 1.375 (see Table 4). Applying Equation 8, we obtain $SL = 2.369$ and a CPI that runs from -2.369 to 2.369 (see Table 4).

How would we interpret this? If we want to turn the significant finding into an a posteriori insignificant one, we would need to include 0 in the 95 percent posterior credible interval. In light of the evidence, we can accomplish this only by letting the a priori difference between French and Italian speakers to range between -2.369 and 2.369. This credible interval is rather wide, meaning that we have to allow all sorts of mean differences a priori. More specifically, a priori we would have to allow for the possibility that Italian speakers would give up to a 2.4 point higher vote propensity score to the Greens than their French speaking peers, holding all else constant. This does not seem very plausible because, for one, the Greens have never performed better in the Ticino than in the Romandie. Thus, we would conclude that it is difficult to challenge the statistically significant effect of the variable French in Table 3.

Let us now turn to the idea of advocating for a statistically non-significant finding.
Table 5: Applying Advocacy Limits to a Non-Significant Finding

<table>
<thead>
<tr>
<th></th>
<th>95% Conf. Int.</th>
<th>95% CPI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L</td>
<td>U</td>
</tr>
<tr>
<td>Urban versus Rural</td>
<td>-0.036</td>
<td>0.560</td>
</tr>
</tbody>
</table>

Imagine, we anticipate a positive effect. The *advocacy limit*, AL, is a value such that the CPI covers the open interval from 0 to AL. We can recover AL by again working backwards from the posterior. As with the skepticism limit, we have \( \mu_P \geq z_{\alpha/2}\sqrt{\phi_P} \). We further know that \( \mu_0 - z_{\alpha/2}\sqrt{\phi_0} = 0 \) or \( \mu_0 = z_{\alpha/2}\sqrt{\phi_0} \). Further, in mathematical terms \( AL = \mu_0 + z_{\alpha/2}\sqrt{\phi_0} \). Substituting the result for \( \mu_0 \), this means \( AL = 2z_{\alpha/2}\sqrt{\phi_0} \).

The key is now once more to derive the expression for \( \phi_0 \). Again using equation (3), it can be shown that, in the case of advocacy,

\[
\phi_0 = \frac{(U + L)^2(U - L)^4}{16z_{\alpha/2}^2U^2L^2} \quad (9)
\]

Consequently,

\[
AL = -\frac{(U + L)(U - L)^2}{2LU} \quad (10)
\]

(The negative multiplier is there because L and U are of opposing signs, rendering the denominator in equation (10) negative. To offset this, we retain only the negative root of \( \phi_0 \), which ensures that \( AL > 0 \).) Once the advocacy limit is obtained, we once more assess its plausibility by considering relevant past evidence and theory.

As an illustration, consider the result for the predictor urban in Table 3. As Table 5 shows, the 95 percent confidence interval runs from -0.036 to 0.560. We expect that urban voters are more inclined to vote for the Greens than rural voters. Accordingly, our prior is that the urban effect is positive. But how positive should it be to turn over the lack of statistical significance? If we want to ensure that the 95 percent posterior credible interval excludes a null effect, then \( AL = 4.625 \). This means that the 95 percent CPI runs from 0 to 4.625. Any effect of urban in this prior region would suffice to turn over the non-significance a posteriori.

What does that mean? Those who believe in an urban-rural divide in Green party support can still credibly challenge the non-significance of the urban predictor by arguing that the anticipated effect is somewhere in the range between 0 and 4.6 points. Of course, they could also challenge the non-significance by claiming a much larger effect. However, such a claim would lack credibility since the CPI suggests that effects greater than 4.6 are a priori unlikely. In the present case, a credible challenge of the null effect seems possible. The value constellation of urban voters is likely to favor environmental causes and parties.
more than that of rural voters. At the same time, few would argue that the divide should amount to more than 5 points on a 0-10 scale such as the vote propensity scale.

The advocacy CPI also sheds light on the amount of information contained in an empirical finding. Imagine that $0 < CPI < \infty$. This would mean that nearly anything goes a priori against the lack of significance. In this case, we had better not put much stock in the failure to reject $H_0$. The tighter the CPI becomes, the smaller the range of possible values that would challenge non-significance and the more difficult it becomes to launch a credible claim at the behest of $H_1$.

Thus, we see that the analysis of credibility is another way of moving beyond the dichotomous interpretation of $p$-values. What looks to be statistically significant may actually be an easy target for a challenge. Conversely, what appears to be non-significant might plausibly be rescued, receiving the benefit of, for example, a second test with new data. The nice thing about analysis of credibility is that one can adjust the formulas easily to a variety of generalized linear models (Matthews, 2018). In this sense, it is as flexible as the use of Bayes factors.

Conclusions

Many reasons exist why one might challenge the current emphasis on $p$-values in scientific practice. Not least is the fact that such reliance has often been accompanied by the compulsion to divide research into worthwhile and worthless efforts, findings that should be sent out for review and those that should not, and papers that deserve to be published and those that do not. As a result, highly surprising and significant results receive a great deal of play, even when reproducibility frequently turns out to be problematic. By contrast, non-significant findings, even when they are extremely robust, may never see the light of day.

In this paper, I have shown that one can think about $p$-values in a much more nuanced way. So-called significant findings may contain less evidence against the null hypothesis than meets the eye, while non-significant findings might actually be more meaningful than one thinks. Bayes factors and analysis of credibility can be used to extract more information from $p$-values and to allow for a more nuanced view of scientific evidence. Authors should consider adding them to their tables and journal editors might want to ask for such information, if only to place research findings in a more realistic perspective and temper publication bias.

References

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin
URL: http://www.nature.com/articles/s41562-017-0189-z

URL: http://projecteuclid.org/euclid.ss/1177013238

URL: http://www.tandfonline.com/doi/abs/10.1080/01621459.1942.10501760

URL: Statistics


URL: https://www.sciencedirect.com/science/article/pii/S0037196308000620


URL: [http://www.nature.com/articles/s41562-016-0021](http://www.nature.com/articles/s41562-016-0021)


**URL:** http://content.apa.org/journals/bul/57/5/416


**URL:** http://www.tandfonline.com/doi/abs/10.1198/000313001300339950


**URL:** http://doi.wiley.com/10.1002/j.1662-6370.2010.tb00435.x


**URL:** https://www.tandfonline.com/doi/full/10.1080/00273171.2015.1099032


**URL:** https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108


**URL:** http://www.tandfonline.com/doi/abs/10.1080/01621459.1951.10500764

Just say no to $p < x \ (\forall x \in (0, 1])$, *s and other evil things*

Simon Hug†
Département de science politique et relations internationales,
Université de Genève

First version: December 2018; this version: May 26, 2019
prepared for the Swiss Political Science Review

---

* The title is inspired by the web-commentary on the March 2019 special issue of *The American Statistician* on “Statistical Inference in the 21st Century: A World Beyond $p < 0.05$” entitled “Stats Experts Plead: Just Say No to P-Hacking” ([https://undark.org/article/statisticians-p-hacking/](https://undark.org/article/statisticians-p-hacking/), accessed May 13, 2019). To be precise, the contributors of this special issue go much further than rejecting p-hacking, which is just an outgrowth of much more egregious practices. I gratefully acknowledge comments on an earlier draft by Kurt Annen, Simone Dietrich, Luzia Helfer, Bjørn Høyland, Adrien Petitpas, and Reto Wüest. It is likely that what comes below will not allow these colleagues to update “significantly” their prior belief $p(mule|me)$.

† Département de science politique et relations internationales, Faculté des sciences économiques et sociales; Université de Genève; 40, Bd du Pont d’Arve; 1211 Genève 4; Switzerland; phone ++41 22 379 83 78; email: simon.hug@unige.ch
1 Introduction

When starting to ponder about what I could add to this debate on statistical significance and p-values in this journal I noted (admittedly with some shock) that during my academic life the discipline had gone (at least) full circle. While in graduate school the hot debate was about how important substantive significance of quantitative results is compared to statistical significance. As determining the latter relies in the classical frequentist linear world on the Gauss-Markov assumptions (as (with the exception of the normality assumption for the error terms) we are reminded by Troeger, 2019), we were taught to consider with circumspection each and every assumption individually in the well known “violations in assumptions” perspective (e.g., Hanushek and Jackson, 1977). As at least some of these assumptions (including the normality assumption) were considered with a healthy dose skepticism (e.g., Achen, 1982), statistical significance, we were taught, is not the main goal in scientific inference, and, if it plays a role, it is only in conjunction with substantive significance and consequently the theoretical basis. Nevertheless, in our readings we encountered many problematic uses of significance testing, which Barnett (1973, 196), in his influential textbook characterizes in the following terms:\footnote{Ward, Greenhill and Bakke (2010) advance, amongst others, a similar point in their more recent article focusing on a specific research area, namely studies on civil conflicts. Amrhein and Greenland (2018) in their commentary on Benjamin et al.’s (2018) proposal to “[r]edefine statistical significance” highlight with an epidemiological example the ludicrousness of this idea.}

This is the basic dilemma: that tests of significance inevitably will be employed as aids to, or even as the basis for, the taking of practical action. And yet their method of construction does not allow for the costs of possible wrong actions to be taken into account in any precise way.

Retrospectively, and perhaps not completely correctly, the raise of maximum likelihood estimations (with the explicitly required distributional assumptions, see King, 1989) in our discipline obviously quickly made this (healthy) skepticism almost an indication of heretism. Major journals (among them APSR and AJPS, to be more precise, their editors) started at that time to insist on systematically indicating with those infamous stars the statistical significance of
The pinnacle of this idea, i.e. that all empirical (and not necessarily only) work would fit into a single mold, came with the period when the editor of AJPS required that all abstracts follow the same structure neatly subdivided into Theory, Methods and Results (with the ensuing demands for a standardized presentation of empirical results). This led to a period during which stargazing became the favorite pass-time of political scientists (and researchers of other related disciplines).

Now the debate has moved on to the question whether “Statistical Inference in the 21st Century: A World Beyond p < 0.05” (title of a special issue of The American Statistician, see Wasserstein, Schirm and Lazar, 2019) should lead us to abandon p-values (and possibly much more) or not (see also Bischof and van der Velden, 2019; Shikano, 2019; Steenbergen, 2019; Troeger, 2019). Interesting to note in this debate is that the critique addressed to the current practices is not new. Again, in the late 20th century and early 2000s several publications alerted us to publication biases (e.g., Rosenthal, 1979; Iyengar and Greenhouse, 1988) and the implicit p-hacking, the “cult of statistical significance” (title of the book by Ziliak and McCloskey, 2008), etc. but as Hubbard (2019) nicely shows, these warnings were happily ignored by large majorities in our disciplines.

Thus, calls for rendering users of statistical tools more alert to these well-(but not widely-)known pitfalls, will not lead us out of the current conundrum. While the contributions to the debate in this journal all agree on the basic problem, and thus also share some commonalities, they differ, however, also with respect to the way forward to address the undeniable problems. To state it from the outset, I have qualms with all these suggestions as they, in my view, address only part of the problem and leave, however, others largely or completely unaddressed (which is also the problem (as the authors admit) of a highly visible suggestion by Benjamin et al., 2018).

As I am not without sin either, I still want to set the record straight. My early quantitative work after graduate school systematically refrained from reporting p-values, significance tests etc. As many readers will have doubted, I received many reviews and letters from journal editors complaining about crucial information missing from my work.

Even textbooks published at that time alerted students to these problems (see also the quote above from Barnett, 1973), by referring to catchy quotes from Nobel-prize winners, which I normally show my students: “If you torture the data long enough, Nature will confess.” (Coase cited by Leamer (1983, 37)); “Beware of testing too many hypotheses; the more you torture data, the more likely they are to confess, but confession obtained under duress may not be admissible in the court of scientific opinion.” (Stigler, cited by Gujarati (1995, 115)).

The proposal does not address multiple-hypothesis testing, P-hacking, publication bias,
where the authors differ in their views, I will organize my commentary in three parts before concluding. In the first part it seems useful to highlight what p-values actually are and where they come from. Second, I return to the issue how these p-values (and other summaries) are used in significance testing (significant, non-significant) or any other categorical decisions (significant, suggestive, non-significant), or any other permutation of terms you can imagine). Third, I focus on alternative ways to proceed to scientific inference, before concluding in the last section.

2 p and its interpretation

Bischof and van der Velden (2019) in their introduction remind us of the historical sources of, on the one hand p-values, and on the other the ominous “p<0.05,” namely Fisher’s (1925) work. In his discussion of the “[i]nsignificance of null hypothesis significance testing” Gill (1999) (also briefly referred to by Troeger, 2019) reminds us, however, that Fisher (1925) never intended his p-values to be used in arbitrating between two hypotheses, i.e. our proposed hypothesis and the corresponding null hypothesis. Only with Neyman and Pearson’s (1933) work on competing hypotheses appeared a second hypothesis in the discussion. As Gill (1999) nicely discusses, the conflation of these two approaches (which are not entirely compatible), allowed the infamous practice of statistical significance testing to enter our (and other) discipline(s), with all the nefarious consequences that we know.

Given these origins, it can surprise that this “bastard” (in all meanings of this term) of a tool has been misused and misinterpreted in the discipline. Surprising is that the authors of the call to “[r]edefine statistical significance” (Benjamin et al., 2018) only refer to Fisher’s (1925) work when mentioning the origins of “statistical significance,” thus ignoring the tensions that result from applying Fisher’s (1925) ideas to something for which he never intended it to be used (for an even more encompassing view on how we got to where we are, see Kennedy-Shaffer, 2019).

low power, or other biases (for example, confounding, selective reporting, and measurement error), which are arguably the bigger problems. . . solutions to these other problems, which include good study design, ex ante power calculations, pre-registration of planned analyses, replications, and transparent reporting of procedures and all statistical analyses conducted.” (Benjamin et al., 2018, 8)
In this section it is also useful to remind the readers that the p-values our favorite statistical packages so diligently spew out are based on the assumption that we estimate for the first time, without any previous look at the data, our model. While this ideal might, with the help of preregistration of experimental studies, be approximated in some of the finest experiments, this is almost never the case in studies using observational data. A conclusion from this might be that even observational work would need to be preregistered (see Humphreys, de la Sierra and Van der Windt, 2013; Burlig, 2018). Gelman and Loken (2014, 460), in their more broad-sweeping discussion, mention preregistration as one way out of the conundrum, highlight, however, also a more profound problem, namely “the garden of forking paths.” Each research from its initiation all the way to its publication is based on choices at numerous forking paths. Each of these choices (i.e., which path to choose), as Gelman and Loken (2014) argue, rules out other possibilities, which, when aiming at reporting the correct p-values, would need to be taken into account. Especially as, according to them, some choices in the “garden of forking paths” need to rely on data, making the perfect respect of preregistration plans, most notably for studies based on observational data, impossible.

Regarding the issues discussed in this section, I infer that all contributors to this debate share the critique addressed to p-values. More specifically, Troeger (2019) embraced, as she reminds us, as the founding-editor of a journal, an editorial policy that required a more explicit recognition of uncertainty to the detriment of p-values (see also Bischof and van der Velden, 2019). Shikano (2019) and Steenbergen (2019), on the other hand, search for solace in Bayesian statistics.

3 p, significance testing and other categorical decisions

Calculating the “correct” p-values, for whatever test, is, as the discussion above suggested, almost never achieved in empirical research. It is also useful to re-
member Achen’s (2002) ART- (A rule of three) principle: with more than three independent variables it is close to impossible to ensure that the Gauss-Markov (and any corresponding set of) assumptions are met. Thus, it becomes even more puzzling how our (and other disciplines) came to a point where a “bright-line” at 0.05 suddenly appeared as the dividing line between a “statistically significant” and a “statistically insignificant” result, which, as the numerous studies on publication biases highlight, also largely decided on whether a paper was publishable or not (and as a consequence submitted or not).

The basic problem is that based on an uncertain summary of our uncertainty a decision is reached, independent of what threshold value for \( p \) is selected. Thus, Gill (1999, 669) argues that

\[\text{[t]he basic problem with the null hypothesis significance test in political science is that it often does not tell political scientists what they think it is telling them. . . . From the current presentation of null hypothesis significance testing in published work it is very easy to confuse statistical significance with theoretical or substantive importance.}\]

This broad critique applies equally to the widely discussed suggestion by Benjamin et al. (2018) to replace the threshold of 0.05 by 0.005. As the authors admit themselves, this will only solve a minor part of the problems due to the use of p-values and significance testing. First of all, replacing one arbitrary value by an other equally arbitrary value will lead to the exactly same publication biases and p-hacking. Interesting to note is that Benjamin et al. (2018) clearly envision as a consequence that sample sizes will have to increase, which will lead to a reduction in the number of studies. This consequence, which the authors seem rather willingly to accept, might be perhaps acceptable in some areas of experimental work (to which the authors oftentimes, but not exclusively, refer to), but not in others and certainly not in most work on observational data. Even some field

\[7\text{For a personal illustration: in a recent co-authored paper we reported a result on a quite important policy-relevant effect relying on a much improved dataset (compared to what other scholars have used so far). Our conclusion that our data did not offer support for the effect that is widely believed to be present \( (p>0.05) \) did not enchant the reviewers nor the editors of the journal (reviews and decision letter on file with the author). Thus, while the new editorial policy of Political Studies Review (Fisher et al., 2019) to allow the publication of shorter notes with “non-significant” results is certainly well-mean, it is still based on the idea that } p = 0.05 \text{ is the dividing line between article-length publications and work that only deserves smaller notes (I thank Bjorn Hoyland for having alerted me to this new policy).}\]
experiments, despite unlimited resources, will not be able to increase the sample size willy-nilly. Should the consequence be, if the study offers results of an effect with a p-value of 0.01 that we should not take it seriously (and not publish it)? Similarly, cross-national comparative work would run into problems of the same type. Amrhein and Greenland (2018) in a commentary highlight these issues with an example of a “non-significant effect” but with very consequential implications (i.e., the survival of a patient). The problem of arbitrary thresholds for p-values (or any other categorical decision reached based on a statistic) can be illustrated with two wagers that I would be willing (at the time of writing) to accept:

- “Give me any hypothesis and an unlimited amount of resources and I will provide you with empirical evidence that allows me to reject the corresponding $H_0$ with $p<0.005$.”

- “Give me any phenomenon that you would like to explain (i.e. a dependent variable), I will find you a hypothesis and data such that I can reject the corresponding $H_0$ with $p<0.005$.”

Regarding the first wager the issue is obviously quite easy to understand and explained by Wasserstein and Lazar (2016, 132) as follows: “Any effect, no matter how tiny, can produce a small p-value if the sample size or measurement precision is high enough, and large effects may produce unimpressive p-values if the sample size is small or measurements are imprecise.” The second wager, on the other hand, relies on the fact that for any phenomenon (i.e., dependent variable) a sufficiently close (in the causal chain) antecedent can be found which, even with rather small sample sizes, may be identified as a variable with a “statistically significant” coefficient in an analysis and this whatever the threshold for the p-values that is adopted (possibly due to an almost or even completely tautological effect). This, with admittedly extreme examples, illustrates that publication decisions (and others as well) can not and should not be taken on the basis of any categorical decision based on a statistic.

---

8Let me be clear that the unlimited resources would not be used to fabricate data as in LaCour and Green (2014).

9This leads several contributors to the special issue of The American Statistician edited by Wasserstein, Schirm and Lazar (2019) to suggest that paper submissions should only be evaluated for publication on the basis of the theory and the research design section, without considering the empirical results (including the p-values). I will return to this suggestion in the conclusion.
Wasserstein, Schirm and Lazar (2019, 2), in their summary of the contributions to *The American Statistician* conclude on this point in the following way:

To be clear, the problem is not that of having only two labels. Results should not be trichotomized, or indeed categorized into any number of groups, based on arbitrary p-value thresholds. Similarly, we need to stop using confidence intervals as another means of dichotomizing (based, on whether a null value falls within the interval). And, to preclude a reappearance of this problem elsewhere, we must not begin arbitrarily categorizing other statistical measures (such as Bayes factors).

This admonition fails to jibe well with a table (adorning my office door) drawn from a methods textbook, which offers two lists of adjectives to be used to describe values of (multiple) correlation coefficients that fall into precisely defined intervals, namely for analyses of individual data, respectively aggregate data. This leads Gelman (2016, 1) to note that “[c]onfidence intervals, credible intervals, Bayes factors, cross-validation: you name the method, it can and will be twisted, even if inadvertently, to create the appearance of strong evidence where none exists. . . . [T]he solution is not to reform p-values or to replace them with some other statistical summary or threshold, but rather to move toward a greater acceptance of uncertainty and embracing of variation.”

As readers of this debate will have noted, these admonitions equally fail to jibe with some of the recommendations offered by the contributors to this debate in this journal. While Bayesian statistics embraces uncertainty in a much more direct and intuitive way, just relying on elements derived from this perspective to arrive again at categorical decisions cannot be our solace. Thus, Shikano (2019), based on an insightful discussion of credible intervals, underlines (as already suggested in a review article by Jackman, 2004) underlines the parallel to confidence intervals, and thus suggests their use in similar ways. This step, however, is likely to lead most scholars applying his suggestions down the same dangerous

---

10Thus, if an analysis of individual data should yield an $r/R$ in the interval between 0.20 and 0.34, then the appropriate adjective is “strong”, while in an analysis relying on aggregated data an $r/R$ between 0.81 and 0.95 should be described as “spectacular.” Readers interested in using these adjectives in the correct way can obtain additional information on the source of this table from the author.
path. Similarly, both Shikano (2019) and Steenbergen (2019), after a detailed discussion of Bayes factors, suggest the latter’s use in model comparisons. As both highlight, Bayes factors require authors to clearly specify two hypotheses and the prior odds. Consequently, as with much else, a Bayesian approach offers a much more solid underpinning for a comparison of hypotheses (compared to the frequentist perspective, see discussion above). The idea, however, to classify values of Bayes factors in a holy trinity (Steenbergen, 2019) is just as helpful as replacing one threshold for $p$-values with two (it is useful to remind the reader that Bayes factors, and their classification, is one of the reasons for suggesting a lower threshold for “statistically significant” results, see Benjamin et al., 2018).

4 p and other measures and approaches

Thus, as discussed above, tweaking Bayesian statistics to allow us to return to the safe ground of looking up values and have certainty about the “significance,” the “suggestivity”, etc. of our results will hardly lead us out of the present conundrum. But this leaves us to some extent with the list of “Don’t’s from the American Statistical Association (ASA) on significance testing and $p$-values (Wasserstein and Lazar, 2016) (for a partial critique, see Gelman, 2016). Clearly this can not be satisfying either. Thus, the ATOM-principles proposed subsequently by Wasserstein, Schirm and Lazar (2019, 2) offer a (admittedly vague) way forward:

Accept uncertainty. Be thoughtful, open, and modest. Remember ‘ATOM.’

“Accepting uncertainty” is obviously the main element and corresponds to what Gelman (2016) in various contributions to this debate has advocated. Doing so also implies (as discussed above) that having recourse to falsely secure categorizations of statistical results, be they significance tests (Benjamin et al., 2018), categorizations of Bayes factors (Steenbergen, 2019), or of credible intervals (Shikano, 2019) will lead us nowhere. Thoughtfulness, openness and modesty,
however, are clearly in the eye of the beholder. Thus, it can not surprise that the various recommendations summarized by Wasserstein, Schirm and Lazar (2019, 4) under the heading “Thoughtful Alternatives and Complements to P-Values" go in quite different directions and at times are at odds. Among them is also the suggestion picked up by Steenbergen (2019), namely to rely on the analysis of credibility as proposed by Matthews (2018, 2019). The usefulness in this approach is in my view the notion that the decision to consider something as “credible” is returned to the substantive level, namely to the question what the prior belief has to be to put in doubt the credibility of the result. An important advantage in this regard is, that there is no way that these required prior beliefs can be meaningfully specified (and categorized) a priori across different fields of study etc. The danger is, however, that what is considered to be putting the credibility of our result in doubt presupposes again, implicitly, a dichotomous definition (i.e., in defining the relevant credible interval).

5 Conclusion

As discussed above, I do not believe that replacing ideas coming from a frequentist perspective with those stemming from a Bayesian perspective as proposed by Shikano (2019) and Steenbergen (2019) provide the silver bullet, despite the undeniable advantages of this perspective when it comes to dealing with uncertainty. And if they are sold as silver bullets, we will just enter again in a cycle of abuse and hacking of principles that are in essence sound (which with some caveats also applies to confidence intervals, etc. in a frequentist perspective), but not equally applicable in all circumstances.

As these few lines will hopefully appear in print in the *Swiss political science review*, I encourage the latter’s editors to choose the nuclear option: Insist that contributions to this journal adhere to the ATOMic principles and ignore any reviews that do not comply with the list of “Don’t”s proposed by the ASA (Wasserstein and Lazar, 2016) while downplaying the ATOMic principles (Wasserstein, Schirm and Lazar, 2019). In order not to be too categorical, I would suggest that studies with (credibly) preregistered plans (whether for experimental and/or observational studies) may report their favorite summary of uncertainty, provided that the authors donate 0.01% of their annual income to ASA or any other professional organization that aims at improving the training
in statistics and quantitative work. Finally, and in part this relates to another part of Achen’s (2002) “Agenda for New Political Methodology,” I would suggest that the editors allow authors of quantitative work to opt for an evaluation procedure in which reviewers only evaluate the theoretical part and the research design proposed by the authors (thus without seeing the empirical results, as suggested by Locascio, 2011, 2019). Locascio (2019) suggests that papers having survived such an evaluation would be “conditionally accepted,” pending an evaluation of the correct execution of the proposed empirical analysis (see relatedly Kmetz, 2019). This (which corresponds in part to the “pre-acceptance: original research” submissions recently introduced by the Journal of Experimental Political Science, see Dickson, 2016; Arceneaux et al., 2018), in my view, would refocus publication decisions again more on the importance of the theoretical claims (i.e., microfoundations in the terms proposed by Achen, 2002) and a sound (and replicable) research design, instead of the count of “s. This would also allow authors to adhere (without fear of punishing reviews) to embrace uncertainty and to be thoughtful, open, and modest.

13This might even be done without having recourse to reviewers, as many journals, among them Political Science Research & Methods, under the stewardship of Troeger (2019) adopted pre-publication replications of conditionally (pending successful replication) accepted papers.
References


Kennedy-Shaffer, Lee. 2019. “Before p<0.05 to Beyond p<0.05: Using History to Contextualize p-Values and Significance Testing.” The American Statistician 73(sup1):82–90.


Troeger, Vera E. 2019. “To P or not to P? The Usefulness of P-values in Quantitative Political Science Research.” *Swiss Political Science Review* .


