Journal of Economic Perspectives—Volume 35, Number 3—Summer 2021—Pages 157–174

Statistical Significance, *p*-Values, and the Reporting of Uncertainty

Guido W. Imbens

Theodor Geisel, better known by the nom de plume Dr. Seuss, published *The Sneetches* in 1961. In this children's story, the Star-Belly Sneetches viewed themselves as superior to the Plain-Belly Sneetches. When the character of Sylvester McMonkey McBean arrives with a machine that can add or remove belly stars (for a modest fee), social upheaval results. In empirical work in economics, stars have long been attached to numbers in tables and figures to indicate the level of statistical significance: one star typically refers to an estimate that is statistically significant at a 10 percent level; two stars, the 5 percent level; and the coveted three stars, the 1 percent level. In the word of Dr. Seuss: "Those stars weren't so big. They were really so small./You might think such a thing wouldn't matter at all./But, because they had stars, all the Star-Belly Sneetches/Would brag, 'We're the best kind of Sneetch on the beaches.'" In empirical studies, estimates with one, two, or three stars are often viewed as superior to those without such adornments.

The statistical significance indicated by stars in tables of empirical results is a concept that is at the same time widely used, widely misunderstood, and widely decried, probably more than any other statistical notion. In this essay, I begin with a short overview of the current controversies among some academic journals and professional societies in reporting *p*-values and statistical significance. Some

■ *Guido W. Imbens is Professor of Economics, Graduate School of Business, Professor of Economics, Department of Economics, and Senior Fellow, Stanford Institute for Economic Policy Research (SIEPR), all at Stanford University, Stanford, California. He is also a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. His email address is imbens@stanford.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.35.3.157.

journals, following the "star-off" machine of Sylvester McMonkey McBean, have started removing indicators of statistical significance before publication, or even further, any use of hypothesis testing.

I then turn to three distinct concerns that have been raised—against or going even further to disallow—the use of statistical significance and *p*-values. The first concern is that often *p*-values and statistical significance do not answer the question of interest. In many cases, researchers are interested in a point estimate and the degree of uncertainty associated with that point estimate as the precursor to making a decision or recommendation to implement a new policy. In such cases, the absence or presence of statistical significance (in the sense of being able to reject the null hypothesis of zero effect at conventional levels) is not relevant, and the all-too-common singular focus on that indicator is inappropriate. Statistical education has arguably failed in clarifying to decision makers, even those with a reasonable degree of statistical sophistication, the key issues involved in decision making under uncertainty.

The second concern arises if a researcher is legitimately interested in assessing a null hypothesis versus an alternative hypothesis: say, the efficient market hypothesis or the permanent income hypothesis. Such cases do commonly arise in economics, although perhaps not as often as in physical sciences, and certainly not as often as the prevalence of null hypothesis testing in empirical work would suggest. As Abadie (2020) writes, "in economics . . . there are rarely reasons to put substantial prior probability on a point null." Questions have been raised whether *p*-values and statistical significance are useful measures for making the comparison between the null and alternative hypotheses. The use of a uniform standard (the ubiquitous 5 percent level for statistical significance) irrespective of context has been questioned. In addition, alternatives to *p*-values have been proposed for this setting, including Bayes factors. Here, I do think there is a limited but important role for *p*-values. Although I agree with much of the sentiment that small *p*-values are not *sufficient* for concluding that the null hypothesis should be abandoned in favor of the alternative hypothesis, I do think that small *p*-values are *necessary* for such a conclusion. More specifically, in cases where researchers test null hypotheses on which we place substantial prior probability, it is difficult to see how one could induce anyone to abandon that belief without having a very small *p*-value. Reporting such a *p*-value would seem a reasonable way to summarize evidence.

The third concern is the abuse of *p*-values. Because in practice much importance is attached to small *p*-values and statistical significance—the number of stars in a table—there are strong incentives for researchers to obtain more favorable *p*-values. To put it bluntly, researchers are incentivized to find *p*-values below 0.05. This has led to concerns about researchers searching for specifications (whether consciously or unconsciously) that lead to such *p*-values in ways that invalidate the meaning and interpretation of those *p*-values. This has become known as *p*-hacking. On the other side of the publication process, there are concerns that results without statistical significance are less likely to be accepted for publication. There is interesting recent work on detecting the presence of *p*-hacking and/or publication bias

(Andrews and Kasy 2019; Elliott, Kudrin, and Wuthrich 2019; Brodeur, Cook, and Heyes 2018). One approach to avoid issues of *p*-hacking relies on the use of pre-analysis plans in which a researcher specifies in advance how data will be gathered and analyzed (Casey, Glennerster, and Miguel 2012; Duflo et al. 2020; Olken 2015), as supported by the AEA registry for randomized experiments.

In this essay, I argue that I find the first concern the most compelling. Statistical significance has been over-emphasized in empirical research.[1] In many cases where decision makers are faced with deciding whether to implement a new policy or not, confidence intervals are a more useful way of communicating uncertainty of point estimates. It would be even better, in my view, to report Bayesian posterior intervals, but in many cases confidence intervals can be interpreted as posterior intervals, and so this often becomes a minor quibble. In cases where Bayesian posterior intervals and confidence intervals differ substantially, I would more strongly prefer posterior intervals.

With regard to the second issue, in which cases where the questions of interest are naturally formulated as hypothesis tests, I think that advantages of Bayes factors over *p*-values are relatively minor. In such cases, it is my view that *p*-values are a reasonable and standardized way of communicating the strength of the evidence.[2] Summarizing the strength of that evidence by a binary indicator—whether a statistically significant at the 5 or 1 percent level—seems to serve little purpose.

Concerning the third issue, *p*-hacking, it would be useful both to lower the incentives for *p*-hacking by de-emphasizing statistical significance thresholds (not reporting stars in tables), and to make it more difficult to *p*-hack by rewarding pre-analysis plans whenever feasible.

Given that this debate over statistical significance and *p*-values has gone on for a long time, I will say little that is new, and perhaps little that is controversial. My aim is to help readers understand the basic issues and why various recommendations have been made in the literature. Cox (2020) offers another recent discussion of some of these issues.

Controversy about the Reporting of *p*-values and Significance Levels

Despite the widespread use of statistical significance and *p*-values, there is much controversy in the academic literature over its appropriate role. Many authors—including multiple journal editors in empirical fields (as opposed to journals devoted to theoretical statistics)—have weighed in on the merits of reporting (in decreasing order of controversy) statistical significance, *p*-values, confidence

[1] The alleged importance of statistical significance has even entered into fiction, as in Nesbø's (2012, p. 93) crime novel *The Bat*: "Trying to find a pattern . . . is hopeless without statistics. Cold, concise statistics. Keyword number one is statistical significance. In other words, we're looking for a system that cannot be explained by statistical chance."

[2] In fact, I have written papers focused primarily on the calculation of *p*-values: including, for example, Athey, Eckles, and Imbens (2018).

intervals, and Bayesian intervals.[3] At the same time, theoretical work on properties of tests continues to attract much attention. The 1995 paper by Benjamini and Hochberg on controlling the false discovery rate when multiple statistical tests are being carried out has been cited well over 70,000 times in 25 years.

The editor of the journal *Basic and Applied Social Psychology* (BASP) went the furthest in terms of restricting the reporting of tests, ultimately banning the use of significance levels, including *p*-values as well as confidence intervals. In 2014, the editor of BASP wrote, "prior to publication, authors will have to remove all vestiges of the NHSTP [Null Hypothesis Statistical Testing Procedures] (*p*-values, *t*-values, *F*-values, statements about 'significant' differences or lack thereof, and so on)" (Trafimov 2014, p. 1). The next year the editors went further and also banned confidence intervals, although, "Bayesian procedures are neither required nor banned" (Trafimow and Marks 2015, p. 1). Back in 1986, the *American Journal of Public Health* included an "Editor's Note" (1986, p. 587, in response to Fleiss 1986) that drew a line between *p*-values and confidence intervals: "We . . . have encouraged the use of confidence intervals. We believe that the quantitative message that they convey is less subject to misinterpretation than significance testing or *p*-values." Editors of some economics journals have drawn the line between reporting indicators of statistical significance and *p*-values. Both *Econometrica* and the *American Economic Review* have policies on their website discouraging the use of stars to indicate statistical significance. *Econometrica* does explicitly encourage standard errors and confidence intervals: "Please do not use asterisks or bold face to denote statistical significance. We encourage authors to report standard errors and coverage sets or confidence intervals."[4]

The actual act of banning a probability calculation in a scientific journal is quite striking. As Hal Stern (2016 p. 23) writes,

> The *p*-value is a probability calculation giving the probability of an event (observing a more extreme t statistic) under specific assumptions: The statistical model is correct and H_0 is true. Probability calculations do not seem particularly objectionable. Why then would BASP [*Basic and Applied Social*

[3]To know the views of the authors, it often suffices to read the titles of such editorials or articles. A partial list of examples includes: "P-values and Confidence Intervals: Two Sides of the Same Unsatisfactory Coin," in *Journal of Clinical Epidemiology* (Feinstein 1998); "The Cult of Statistical Significance" (Ziliak and McCloskey 2011); "That Confounded P-value," in *Epidemiology* (Lang, Rothman, and Cann 1998); "A Dirty Dozen: Twelve *P*-value Misconceptions" (Goodman 2008); "An Investigation of the False Discovery Rate and the Misinterpretation of *p*-values" (Colquhoun 2014); "Toward Evidence-Based Medical Statistics. 1: The *P* value Fallacy" (Goodman 1999a); "The End of the p value" (Evans, Mills, and Dawson 1988); "The Difference between 'Significant' and 'Not Significant' Is Not Itself Statistically Significant" (Gelman and Stern 2006); "Confidence Intervals Rather than P values: Estimation Rather than Hypothesis Testing" (Gardner and Altman 1986); "In Praise of Confidence Intervals" (Romer 2020); and "Testing a Point Null Hypothesis: The Irreconcilability of *P* Values and Evidence" (Berger and Sellke 1987). In "Why Most Published Research Findings Are False," Ioannidis (2005) writes: "Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles should be interpreted based only on *p*-values."

[4]This policy predates my term as Editor of *Econometrica*, and I had no involvement in its formulation.

> *Psychology*] ban *p*-values? . . . It is true that *p*-values are often misinterpreted and abused . . . but that by itself does not seem like a compelling reason to ban them.

Perhaps even more striking, the American Statistical Association put out an official statement on *p*-values that included the following (Wasserstein and Lazar 2016):

> Underpinning many published scientific conclusions is the concept of "statistical significance," typically assessed with an index called the *p*-value. While the *p*-value can be a useful statistical measure, it is commonly misused and misinterpreted. This has led to some scientific journals discouraging the use of *p*-values, and some scientists and statisticians recommending their abandonment, with some arguments essentially unchanged since *p*-values were first introduced.[5]

It is surely quite unusual for a professional society to weigh in on a specific scientific issue like the merit of a given statistic. In a blog post on the website of *Nature,* Monya Baker (2016, p. 151) writes: "'This is the first time that the 177-year-old ASA has made explicit recommendations on such a foundational matter in statistics,' says executive director Ron Wasserstein. 'The society's members had become increasingly concerned that the *p*-value was being misapplied in ways that cast doubt on statistics generally,' he adds."

A subsequent article by Wasserstein, Schirm, and Lazar (2019, p. 1), although not a formal statement of the American Statistical Association, went further than the original words of caution by explicitly recommending against the use of statistical significance indicators:

> The ASA Statement on *p*-values and statistical significance stopped just short of recommending that declarations of "statistical significance" be abandoned. We take that step here. We conclude, based on our review of the articles in this special issue and the broader literature, that it is time to stop using the term "statistically significant" entirely. Nor should variants such as "significantly different," "$p < 0.05$," and "nonsignificant" survive, whether expressed in words, by asterisks in a table, or in some other way.

To put this in perspective, I find it difficult to imagine the American Economic Association issuing an edict that a certain statistical approach would be banned (say, the use of instrumental variables) or the editor of the *American Economic Review* prohibiting researchers from mentioning a method or economic theory (say, the

[5] In the spirit of full disclosure, I was part of the committee that was tasked with crafting the statement.

permanent income hypothesis or rational expectations) solely because the editor felt that these methods or theories have at times been misapplied.

Although the use of statistical significance is common in economics, these discussions about statistical significance and *p*-values have not generated quite as much excitement in the economics profession as in other fields using statistical methods. In one recent exception, David Romer (2020) carefully documents that the majority of empirical papers in three leading economics journals (*American Economic Review*, *Quarterly Journal of Economics*, and *Journal of Political Economy*) focus primarily on point estimates and statistical significance in various forms. He argues against this practice and recommends reporting confidence intervals instead to summarize the uncertainty in the point estimates: "Focusing on point estimates and statistical significance obscures the implications of the findings for those values [values other than the point estimate and zero]. In addition, as discussed below, this focus also leaves out important information even about the strength of the evidence against a parameter value of zero." Another exception in the economics literature is Abadie (2020), who points out that in some cases, nonsignificant results may be much more informative than significant results in terms of changing beliefs about plausible values of the parameters of interest.

Estimation versus Hypothesis Testing

I will begin with some comments about the general nature of empirical work in economics and the relative importance of estimation versus hypothesis testing. Although hypothesis testing is routinely used in economics, I would submit that many of the substantive questions are primarily about point estimation and their uncertainty, rather than about testing. However, many studies where estimation questions should be the primary focus present the results in the form of hypothesis tests. Romer (2020) presents a specific example—the return to schooling—where testing a null hypothesis of no effect is common, yet arguably of little or no substantive interest. One would be hard-pressed to find an economist who believes that the return to education is zero. As Romer (p. 56) notes, "[T]he vast previous work in this area already provides overwhelming evidence that the rate of return is positive." Imagine for a moment that the abstract of a paper in an economics journal claimed, along the lines of the abstracts of many medical papers: "We show that an increase in education causes significantly higher earnings." One rarely sees such abstracts, because such a finding would not be surprising or interesting. For the same reason, such claims should not feature prominently in the paper. What is of interest in such papers is the magnitude and uncertainty of the estimates, and the robustness to identification concerns, not whether the data allow for the rejection of a zero effect.

Given this distinction between estimation and testing problems, in the next two sections I will discuss the role of *p*-values and statistical significance in analyses for such problems.

Decision Making under Uncertainty

Consider a decision maker choosing whether to implement a new policy—perhaps mandating a new early childhood educational program (Krueger and Whitmore 2001; Schanzenbach 2006; Chetty et al. 2011), or making micro credit available to communities in developing countries (Banerjee, Karlan, and Zinman 2015; Crépon et al. 2015; Meager 2019), or changing a search algorithm for a tech company (Gomez-Uribe and Hunt 2015; Gupta et al. 2019). Suppose the only unknown component of the utility of implementing the policy is the average treatment effect (the difference in the average outcome if everybody was exposed to the intervention versus the average outcome if nobody was exposed). To inform this decision, suppose that a randomized experiment was conducted. In this experiment, a sample of units is randomly divided into two sub-samples, with units in the first sub-sample exposed to the intervention and the units in the second sub-sample exposed to the old regime. (I am focusing here on an example with a randomized experiment because it abstracts from some other concerns about internal validity that would also come up in such discussions in the absence of randomization: for general discussions of these issues, see Deaton 2010; Imbens 2010, 2018; and Deaton and Cartwright 2018.)

A question is what information should the statisticians bring to the meeting with the (sophisticated) decision maker after having analyzed the data. In my experience, it is common in such settings for the statistician to present point estimates of the average effect, together with some combination of statistical significance, standard errors, confidence intervals, subgroup analyses, and robustness checks. A discussion might then ensue concerning the magnitude of the effect and the precision of the estimated effect, where the latter discussion would cover the degree of statistical significance and standard errors. There would also be a discussion regarding the credibility of the findings (especially in settings where the estimates are not based on randomized experiments), as well as their external validity and any evidence of heterogeneity. Kohavi, Henne, and Sommerfield (2007), Kohavi, Tang, and Xu (2020), and Gupta et al. (2019) discuss in more detail the process of decision making in the context of randomized experiments in a business setting. Kohavi views experiments in this setting, and data-driven decision making more generally, as helping reduce the importance of what he has called the Highest Paid Person's Opinion (HIPPO) in less formal versions of these discussions.

In this setting of providing information to decision makers, I want to make two claims. First, what is most relevant for the decision maker is the point estimate with some measure of the uncertainty of that point estimate, and some sense of the robustness and identification issues. The second claim is that the testing of statistical hypotheses—and thus the reporting of p-values or statistical significance—is essentially irrelevant in this case. The common practice of prominently reporting these measures is therefore largely misguided. As the statement of the American Statistical Association claims, correctly in my view, "Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold" (Wasserstein and Lazar 2016).

To provide further support for the view that in this case the appropriate focus is on point estimates and measures of uncertainty, consistent with the view of some econometricians of econometrics as applied decision theory (Leamer 1978; Chamberlain 2000; Manski 2013; Hirano 2010; Dehejia 2005), let me make this example more specific. Suppose the point estimate of the treatment effect is $\hat{\tau} > 0$ (with positive values preferred relative to negative values by the decision maker), and suppose the standard error is σ. Let us also suppose that the analysts are confident that the sampling distribution of the estimator is approximately normal, so that the 95 percent confidence interval is plus or minus 1.96 standard deviations from the point estimate $\hat{\tau}$. Given these numbers, the discussion of the decision makers would typically center on the plausibility of the estimates, the magnitude of the cost relative to the estimated benefits, the external validity of the estimates (will they actually generalize to the population they might be applied to), evidence of heterogeneity in the effects, and the possibility (or explicitly, the probability) of effect sizes that would render the decision to be clearly wrong after it was taken, possibly taking into account prior beliefs. These topics have an implicitly Bayesian flavor: the decision maker is in various ways confronting the point estimates with prior beliefs. The use of confidence intervals as the basis for a discussion in a Bayesian spirit is (approximately) justified by the interpretation of the confidence intervals as Bayesian intervals, although this is rarely made explicit.[6]

In addition, identication issues may arise, for example, from lack of randomization, or via uncertainty about differences between the study population and the target population, or uncertainty about differences between the future and the past. These are often dealt with informally by just acknowledging that some degree of additional uncertainty exists, rather than by using more principled ways of calculating bounds along the lines of the work by Manski (2013).

Although the topic of statistical significance is often brought up in these discussions, it often is used inappropriately by implicitly interpreting insignificant estimates as true zeros. To illustrate the lack of a role for the significance level, suppose the utility from the general implementation of the treatment is equal to the true treatment effect, so that implicitly the cost of implementing the treatment is zero, and there is no risk aversion. In this case, the right decision given a treatment effect equal to τ would be to implement the intervention if the estimated value of $\tau > 0$, and not otherwise. From a Bayesian perspective, the only reason not to implement the intervention given a positive estimate $\hat{\tau}$ would be that the prior distribution for τ implies that the posterior expected value for τ is

[6]This is based on the Bernstein-Von Mises theorem that, informally, says that in many cases confidence intervals can be viewed as approximate Bayesian posterior intervals (Van der Vaart 2000). Although there are multiple settings where confidence intervals are not based on asymptotic normality (for example, in instrumental variables settings with weak instruments, or with settings with unit roots), I have not seen analysts attempt to explain such confidence intervals to policy-makers, and I would expect that to be a challenging task. In such cases where the Bernstein-von Mises Theorem does not hold and confidence intervals are *not* similar to (Bayesian) posterior intervals I would strongly prefer the Bayesian intervals over confidence intervals. See Sims and Uhlig (1991) for a related discussion in the context of unit roots.

negative, despite that positive value for the estimated $\hat{\tau}$. If one really believes that a flat prior is appropriate, then even the value of the standard error σ does not actually matter. In practice, of course, a flat prior is almost always implausible and the prior standard deviation is often modest. Moreover, one may a priori be skeptical about the proposed intervention, so that the prior mean is negative. In that case, one needs not just a positive point estimate, but also a sufficiently positive and precise point estimate to justify the implementation of the proposed intervention. In some cases, such a prior distribution could be justified more systematically using data from prior experiments using an empirical Bayes approach (Morris 1983). Although I am pushing for a more Bayesian approach than is typically reported, I would be comfortable with the statisticians just reporting the point estimates and confidence intervals, because decision makers can then combine that with their own prior distributions (for example Andrews and Shapiro 2020).

In the case I just outlined, presenting the implicitly Bayesian decision makers with *p*-values or conventional indicators of statistical significance does them a disservice and in practice underestimates their sophistication. In practice it often leads decision makers to act as if statistically insignificant results are truly zero. In doing so, it confuses the matter at hand by distracting the decision maker from the real issues: what are the costs of type I and type II errors, what are their prior beliefs, and how much the estimates change those beliefs. As Abadie (2020) shows, statistical significance need not change those beliefs very much.

Assessing the Relative Merits of the Null Hypothesis versus an Alternative Hypothesis

Although I have argued that in many cases point estimates and confidence intervals are the most useful summary statistics from a statistical analysis, there are settings in economics where it may be reasonable to focus on testing null hypotheses, often about a particular economic theory. We may be interested in testing the permanent income hypothesis, the efficient market hypothesis, whether there are constant returns to scale, whether there is a "sheepskin effect" of graduation in the returns to education, or whether particular groups are discriminated against. Although in all these examples one can still argue how seriously to take such a sharp null hypothesis (that is, with sufficiently large samples we might expect to reject most of such hypotheses), it may still be useful to assess whether there is clear evidence in the available data against such theories. To make the discussion specific, let me focus on an (non-economics) example where testing whether the null hypothesis holds may be more relevant than the magnitude of deviations from the null hypothesis if it is violated, and where the testing has generated much controversy. A similar example is the hot-hand fallacy (Ritzwoller and Romano 2020).

This example attracted great controversy in the psychology literature. In the *Journal of Personality and Social Psychology*, Bem (2011) studies whether precognition

exists: that is, whether future events retroactively affect people's responses. Reviewing nine experiments, he finds (from the abstract): "The mean effect size (d) in psi performance across all nine experiments was 0.22, and all but one of the experiments yielded statistically significant results." This finding sparked considerable controversy, some of it methodological. The title of a response by Wagenmakers et al. (2011) sums up part of the critique: "Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi: Comment on Bem (2011)." A *New York Times* article on the controversy was titled, "Journal's Paper on ESP Expected to Prompt Outrage," which states: "Many statisticians say that conventional social-science techniques for analyzing data make an assumption that is disingenuous and ultimately self-deceiving: that researchers know nothing about the probability of the so-called null hypothesis" (Carey 2011). The same issue is addressed in the statement by the American Statistical Association: "By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis" (Wasserstein and Lazar 2016).

In this case, it would appear there is reasonable interest in testing the sharp null hypothesis irrespective of the magnitude of the effect: that is, the question of whether precognition exists at all is interesting. The same can be argued for drug trials, where some cases have found that a particular drug or medical procedure has some effect on a medical condition, even if the effect is very small and possibly far below a cost-effective level. Such a finding is informative about possible mechanisms and may suggest further research into alternative treatments. I see these settings as qualitatively different from the decision problem discussed in the previous section, where the question was whether to implement a particular intervention. Here the decision question is whether to investigate a particular scientific question further. In this setting I disagree with Neyman's (1935) comment that knowing that a treatment has some effect, even if the average effect is zero, is purely academic. Here, such a finding is important even if it is not of immediate policy relevance.

Even if we agree that assessing the null hypothesis relative to an alternative hypothesis is for certain questions a matter of interest, one might argue as to whether the *p*-value is the most useful statistic for assessing that null hypothesis. Arguments have been put forward in favor of an explicitly Bayesian approach, as in, for example, Wagenmakers et al. (2011), Goodman (1999b), and Carey (2011). Using a probability that a null hypothesis that precognition does not exist equal to 10^{-20} (a prior distribution more or less in agreement with my own), Wagenmakers et al. (2011) show that the posterior probability that precognition exists, given some of Bem's experiments, remains very small so as to make it unlikely. I agree with the premise of Wagenmaker's argument that a small *p*-value alone is not *sufficient* to reject the null hypothesis in favor of the alternative hypothesis. However, I do think a small *p*-value is *necessary* for this. It is difficult to imagine a dataset that would contain enough information to reject the null hypothesis of no precognition *without* a small *p*-value. Here I agree with Benjamini (2016, p. 1) who writes: "[The *p*-value] offers a first line of defense against being fooled by randomness, separating signal from noise."

There is a substantial literature on whether the use of a "Bayes factor" would be more informative than p-values, part of an even larger literature on alternatives to p-values.[7] Given a null hypothesis and an alternative hypothesis, the Bayes factor is the ratio of the marginal likelihood of the data under the null hypothesis and the marginal likelihood of the data under the alternative hypothesis (Kass and Raftery 1995). Unlike the fully Bayesian calculation of the posterior probability that the null hypothesis is true given the data, the calculation of a Bayes factor does not require a prior probability that the null hypothesis is true. A couple of points are worth noting about this measure of the evidence. First, an attractive feature of the Bayes factor is that it is symmetric in its dependence on the two hypotheses, whereas the p-value conditions on the null hypothesis being true. Second, to calculate the actual probability of one of the hypotheses being true, the Bayes factor is not sufficient: we also need the prior probabilities of either hypotheses being true. Such prior probabilities are likely to be controversial. Finally, and this is probability the biggest reason the use of the Bayes factor is less common in practice than the p-value, it also requires a prior distribution to deal with nuisance parameters. For example, if the null hypothesis is sharp—say, that a coin is fairly balanced between heads and tails—the alternative hypothesis is typically not sharp: all values for p other than $p = 1/2$ are consistent with the alternative hypothesis. The calculation of the Bayes factor requires the specification of a prior distribution under the alternative hypothesis, that is, a prior distribution for p on the interval [0,1] excluding the value 1/2. Although in specific cases there may be natural prior distributions to consider (for some discussions, see Goodman 1999b; Berger and Pericchi 1996), in general this makes the Bayes factor calculations more challenging and controversial. For example, if we wish to test the null hypothesis that a drug has no effect on a health outcome, there is no natural prior distribution for the treatment effect under the alternative hypothesis. In the end, I do not see the advantages of Bayes factors over p-values as sufficient to convince researchers to adopt this technology more widely.

Finally, if one is comfortable with the use of p-values in settings such as these, the question remains whether the use of a standardized threshold of 5 percent is useful to indicate statistical significance. At some level, it is not surprising that researchers adopt a standard—whether 5 percent or some other level—to facilitate communication. However, it is difficult to justify a single standard across a wide range of applications that may differ enormously: for example, in terms of size of datasets, costs of type I and type II errors, the number of tests performed, and the prior beliefs about the null hypotheses. Such concerns have led researchers in genetics to move to substantially lower significance thresholds (Storey and Tibshirani 2003). In high-energy physics, statistical significance is commonly ascribed only

[7] As one example, "Lindley's paradox" concerns the discrepancy between frequentist testing and Bayesian calculations of the probability that the null hypothesis is true. The paradox is that for a given significance level p, a test can be statistically significant, even though the posterior probability of the null hypothesis can be arbitrarily high. This can happen when the prior probability of the null hypothesis is non-negligible, the sample is large, and the prior distribution over values consistent with the alternative hypothesis is sufficiently spread out.

to findings with *p*-values below 3×10^{-7}, corresponding to estimates more than five standard errors away from zero (for example, Sinervo 2002). Benjamin et al. (2018) suggest using 0.005 (corresponding approximately to estimates more than three standard errors away from zero), rather than 0.05, as a standard for indicating statistical significance in cases where the question of interest is whether to override a strong prior belief.[8]

Publication Bias and *p*-hacking

For academic researchers, the presence or absence of a statistically significant result may influence the chance of publication and thus career success. For drug companies, a *p*-value less than or more than 0.05 can mean a difference in revenues of billions of dollars. Thus, researchers may be tempted to shape or change their analyses to reach the unstated goal of a statistically significant result.

One of the most striking examples of such abuse is that of Scott Harkonen, the fomer CEO of InterMune. Intermune did a randomized trial for a drug that Harkonen called "a $2 billion market opportunity for InterMune" (Brown 2013). Comparing survival rates for all treated and control patients in the study led to a *p*-value of 0.08, not statistically significant at conventional levels. However, by creatively looking for subgroups (who had not been included in any pre-analysis plan), InterMune found that for the subsample of participants with mild to moderate (but not severe) cases of the disease, the drug had an effect on survival with a highly significant *p*-value of 0.004. The company sent out a press release: "InterMune Announces Phase III Data *Demonstrating* [my italics] Survival Benefit of Actimmune in IPF . . . . Reduces Mortality by 70 percent in Patients with Mild to Moderate Disease." As Mayo (2020) describes this episode, which ultimately led to a conviction for issuing a misleading press report, Harkonen "reported statistically significant drug benefits had been shown, without mentioning this referred only to a subgroup he identified from ransacking the unblinded data." Indeed, Brown (2013) reports on a follow-up study carried out by InterMume that "enrolled only people with mild to moderate lung damage, the subgroup whose success was touted in the press release. And it failed. A little more than a year into the study, more people on the drug had died (15 percent) than people on placebo (13 percent). That was the death knell for the drug. Most insurers stopped paying for it."

The suspicion is that there are many more cases that do not have billions of dollars at stake, but where researchers also search for specifications that lead to *p*-values that cross the threshold into the territory that allows them to be referred to as statistically significant (Head et al. 2015). Concerns about searching through specifications for statistically significant results have been prominent in econometrics at least since the work of Edward Leamer (1978, 1983). In particular, there

[8] I am sympathetic to this proposal, and in fact was one of the many authors on this paper.

may be substantial incentives for researchers to come up with surprising findings of effects where prior beliefs put a high probability on these effects being absent. Such findings are more likely to be picked up by the popular press and, in general, gather attention as well as lead to publications in academic journals. Andrew Gelman has eloquently criticized many examples on his blog *Statistical Modeling, Causal Inference, and Social Science*, focusing on the concerns that even if researchers do not deliberately set out to calculate misleading *p*-values, they make many specification choices (the "garden of forking paths") that affect these measures, so the reported results should not be taken at face value (Gelman and Loken 2013).

One example that Gelman presents involves the "hurricanes versus himmicanes" controversy: is damage greater from hurricanes with female names rather than male names? The finding seems implausible on its face, given that female and male names are assigned to hurricanes on an alternating basis. However, Jung et al. (2014) apply a 5 percent significance standard and write in their abstract: "We use more than six decades of death rates from US hurricanes to show that feminine-named hurricanes cause significantly more deaths than do masculine-named hurricanes." If the paper had been submitted to the *Proceedings of the National Academy of Sciences* with an abstract reading "We use more than six decades of death rates from US hurricanes to show that the damage of hurricanes is not related to the gender of their name," would the paper have been accepted for publication? If the authors had not found a statistically significant result, would they have simply moved on to another project?

One direction that has been explored in the literature is to assess evidence for possible abuse of *p*-values by exploring specifications that are not reported, or what is typically referred to as "*p*-hacking" (Andrews and Kasy 2019; Elliott, Kudrin, and Wuthrich 2019; Brodeur, Cook, and Heyes 2018). A related issue is publication bias, where reviewers and editors may be more inclined to accept for publication papers with low *p*-values and/or statistically significant results. The presence of *p*-hacking and publication bias can be detected using data on a large number of published articles: for example, if there is a discontinuity in the distribution of *p*-values, with a larger number of *p*-values just below 0.05 relative to the number of *p*-values just above 0.05.

Detecting *p*-hacking is one thing; addressing it is a different matter (Simmons, Nelson, and Simonsohn 2013). One possible approach is to use replication studies (as in Makel, Plucker, and Hegarty 2012), which can focus on what choices were made behind the scenes in reaching the statistically significant result. Such studies do not directly prevent *p*-hacking but can show that the announced results have less support than it might seem. De-emphasizing *p*-values (and perhaps also statistical significance more broadly) may decrease the incentives for *p*-hacking, and thus lower its prevalence. In some contexts, in particular with randomized experiments, filing a pre-analysis plans that specifies how the data will be analyzed can also help to prevent *p*-hacking (Casey et al. 2012; Chang and Li 2017; Duflo et al. 2020). Such pre-analysis plans are required by the Food and Drug Administration in its drug approval process and are becoming increasingly used in social sciences. The

American Economic Association has operated a registry for randomized experiments since 2012 that provides all the essential benefits from pre-analyses plans.

Publication bias may be more difficult to deal with. In some cases, journals are willing to pre-commit to publishing studies based on pre-analysis plans, but it is difficult to imagine that practice becoming widespread. Consider an editor approached with a proposal to investigate precognition through a well-designed, large-scale trial. Given a very strong prior belief that precognition does not exist, it is difficult to see why an editor would pre-commit to publishing such a study. On the other hand, if the study was well-designed and did find a substantial and precisely estimated effect, there would be clear arguments after the work was completed to publish such a study—if only to encourage other researchers to further investigate the topic.

Conclusion

The use of *p*-values and indicators for statistical significance has become a matter of substantial controversy. Some journals have established policies banning the use of such measures. In my view, banning *p*-values is inappropriate. As I have tried to argue in this essay, I think there are many settings where the reporting of point estimates and confidence (or Bayesian) intervals is natural, but there are also other circumstances, perhaps fewer, where the calculation of *p*-values is in fact the appropriate way to answer the question of interest. Moreover, there is little evidence that a blanket ban on *p*-values improves the quality of statistical reporting. When the journal *Basic and Applied Social Psychology* banned *p*-values, the editors wrote that, "We hope and anticipate that banning the NHSTP [null hypothesis statistical testing procedures] will have the effect of increasing the quality of submitted manuscripts by liberating authors from the stultified structure of NHSTP thinking thereby eliminating an important obstacle to creative thinking" (Trafimow and Marks 2015, p. 2). However, a study assessing statistical studies published in the journal following the *p*-value ban concludes the opposite. Quoting from the abstract: "We found multiple instances of authors overstating conclusions beyond what the data would support if statistical significance had been considered. Readers would be largely unable to recognize this because the necessary information to do so was not readily available" (Fricker Jr. et al. 2019).

Although I do not endorse a ban on the reporting of *p*-values, I do agree that over the years, and in some disciplines more than other, *p*-values and statistical significance have been overemphasized. In many cases, the *p*-value or the measure of statistical significance is not the relevant output from an analysis of a dataset. Therefore, its prominence in the abstracts of many empirical papers is misplaced. It would be preferable if reporting standards emphasized confidence intervals (as Romer 2020 suggests) or standard errors, and, even better, Bayesian posterior intervals.

■ *I am grateful for comments by Alberto Abadie and Kei Hirano and for generous support from the Office of Naval Research through ONR grant N00014-17-1-2131.*

References

Abadie, Alberto. 2020. "Statistical Nonsignificance in Empirical Economics." *American Economic Review: Insights* 2 (2): 193–208.

Andrews, Isaiah, and Maximilian Kasy. 2019. "Identification of and Correction for Publication Bias." *American Economic Review* 109 (8): 2766–94.

Andrews, Isaiah, and Jesse M. Shapiro. 2020. "A Model of Scientific Communication." NBER Working Paper 26824.

Athey, Susan, Dean Eckles, and Guido W. Imbens. 2018. "Exact p-values for Network Interference." *Journal of the American Statistical Association* 113 (521): 230–40.

Baker, Monya. 2016. "Statisticians Issue Warning over Misuse of *p*-values." *Nature News* 531 (7593): 151.

Banerjee, Abhijit, Dean Karlan, and Jonathan Zinman. 2015. "Six Randomized Evaluations of Microcredit: Introduction and Further Steps." *American Economic Journal: Applied Economics* 7 (1): 1–21.

Bem, Daryl J. 2011. "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect." *Journal of Personality and Social Psychology* 100 (3): 407–25.

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen et al. 2018. "Redefine Statistical Significance." *Nature Human Behaviour* 2 (1): 6–10.

Benjamini, Yoav. 2016. "It's Not the p-values' Fault." *The American Statistician* 70 (2).

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1): 289–300.

Berger, James O., and Luis R Pericchi. 1996. "The Intrinsic Bayes Factor for Model Selection and Prediction." *Journal of the American Statistical Association* 91 (433): 109–22.

Berger, James O., and Thomas Sellke. 1987. "Testing a Point Null Hypothesis: The Irreconcilability of *p*-values and Evidence. *Journal of the American statistical Association* 82 (397): 112–22.

Brodeur, Abel, Nikolai Cook, and Anthony G. Heyes. 2018. "Methods Matter: P-Hacking and Causal Inference in Economics." IZA Discussion Paper 11796.

Brown, David. 2013. "The Press-Release Conviction of a Biotech CEO and Its Impact on Scientific Research." *The Washington Post*, September 23. https://www.washingtonpost.com/national/health-science/the-press-release-crime-of-a-biotech-ceo-and-its-impact-on-scientific-research/2013/09/23/9b4a1a32-007a-11e3-9a3e-916de805f65d_story.html.

Carey, Benedict. 2011. Journal's Paper on ESP Expected to Prompt Outrage. *The New York Times*, January 06.

Casey, Katherine, Rachel Glennerster, and Edward Miguel. 2012. "Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan." *The Quarterly Journal of Economics* 127 (4): 1755–1812.

Chamberlain, Gary. 2000. "Econometrics and Decision Theory." *Journal of Econometrics* 95 (2): 255–83.

Chang, Andrew C., and Phillip Li. 2017. "Preanalysis Plan to Replicate Sixty Economics Research Papers that Worked Half of the Time." *American Economic Review* 107 (5): 60–64.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star." *Quarterly Journal of Economics* 126 (4): 1593–1660.

Colquhoun, David. 2014. "An Investigation of the False Discovery Rate and the Misinterpretation of *p*-values." *Royal Society Open Science* 1 (3): 140216.

Cox, David R. 2020. "Statistical Significance." *Annual Review of Statistics and its Application* 7: 1–10.

Crépon, Bruno, Florencia Devoto, Esther Duflo, and William Parienté. 2015. "Estimating the Impact of Microcredit on Those Who Take It Up: Evidence from a Randomized Experiment in Morocco." *American Economic Journal: Applied Economics* 7 (1): 123–50.

Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48 (2): 424–55.

Deaton, Angus, and Nancy Cartwright. 2018. "Understanding and Misunderstanding Randomized Controlled Trials." *Social Science & Medicine* 210: 2–21.

Dehejia, Rajeev H. 2005. "Program Evaluation as a Decision Problem." *Journal of Econometrics* 125 (1–2) : 141–73.

Duflo, Esther, Abhijit Banerjee, Amy Finkelstein, Lawrence F. Katz, Benjamin A. Olken, and Anja Sautmann. 2020. "In Praise of Moderation: Suggestions for the Scope and Use of Pre-analysis Plans for RCTs in Economics. NBER Working Paper 26993.

Editor's Note. 1986. *American Journal of Public Health* 76 (5): 587–88.

Elliot, Graham, Nikolay Kudrin, and Kaspar Wuthrich. 2019. "Detecting *p*-hacking. arXiv preprint arXiv:1906.06711.

Evans, S.J., Peter Mills, and Jane Dawson. 1988. "The End of the *p*-value*? British Heart Journal* 60 (3): 177–80.

Feinstein, Alvan R. 1998. "P-values and Confidence Intervals: Two Sides of the Same Unsatisfactory Coin." *Journal of Clinical Epidemiology* 51 (4): 355–60.

Fleiss, Joseph L. 1986. "Confidence Intervals vs Significance Tests: Quantitative Interpretation." *American Journal of Public Health* 76 (5): 587–88.

Fricker, Ronald D. Jr, Katherine Burke, Xiaoyan Han, and William H Woodall. 2019. "Assessing the Statistical Analyses Used in Basic and Applied Social Psychology after their *p*-value Ban." *The American Statistician* 73(S1): 374–84.

Gardner, Martin J., and Douglas G. Altman. 1986. "Confidence Intervals Rather than P Values: Estimation Rather than Hypothesis Testing." *Br Med J (Clin Res Ed)* 292: 746–50.

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 Was Posited ahead of Time. http://stat.columbia.edu/~gelman/research/unpublished/forking.pdf.

Gelman, Andrew, and Hal Stern. 2004. "The Difference between "Significant" and "Not Significant" Is Not Itself Statistically Significant." *The American Statistician* 60 (4): 328–31.

Gomez-Uribe, Carlos A., and Neil Hunt. 2015. "The Netflix Recommender System: Algorithms, Business Value, and Innovation." *ACM Transactions on Management Information Systems* 6 (4):1–19.

Goodman, Steven N. 1999a. "Toward Evidence-Based Medical Statistics. 1: The *P* Value Fallacy." *Annals of Internal Medicine* 130 (12): 995–1004.

Goodman, Steven N. 1999b. "Toward Evidence-Based Medical Statistics. 2: The Bayes Factor." *Annals of Internal Medicine* 130 (12): 1005–1013, 1999b.

Goodman, Steven. 2008. "A Dirty Dozen: Twelve *P*-Value Misconceptions." *Seminars in Hematology* 45 (3): 135–140.

Gupta, Somit, Ronny Kohavi, Diane Tang, Ya Xu, Reid Andersen, Eytan Bakshy, Niall Cardin et al. 2019. "Top Challenges from the First Practical Online Controlled Experiments Summit." *ACM SIGKDD Explorations Newsletter* 21 (1): 20–35.

Head, Megan L., Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. 2015. "The Extent and Consequences of P-Hacking in Science." *PLOS Biology* 13 (3): e1002106.

Hirano, Keisuke. 2010. "Decision Theory in Econometrics." In *Microeconometrics,* edited by Steven N. Durlauf and Lawrence E. Blume, 29–35. New York: Springer.

Imbens, Guido W. 2010. "Better LATE than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic Literature* 48 (2): 399–423.

Imbens, Guido. 2018. "Understanding and Misunderstanding Randomized Controlled Trials: A Commentary on Deaton and Cartwright." *Social Science & Medicine (1982)* 210: 50–52.

Ioannidis, John P.A. 2005. "Why Most Published Research Findings Are False." *PLOS Medicine* 2 (8):e124.

Jung, Kiju, Sharon Shavitt, Madhu Viswanathan, and Joseph M. Hilbe. 2014. "Female Hurricanes Are Deadlier than Male Hurricanes." *Proceedings of the National Academy of Sciences* 111 (24): 8782–87.

Kass, Robert E., and Adrian E. Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90: 773–95.

Kohavi, Ron, Randal M. Henne, and Dan Sommerfield. 2007. "Practical Guide to Controlled Experiments on the Web: Listen to Your Customers Not to the Hippo." *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 959–67.

Kohavi, Ron, Diane Tang, and Ya Xu. 2020. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing.* Cambridge, UK: Cambridge University Press.

Krueger, Alan B., and Diane M. Whitmore. 2001. "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project Star." *The Economic Journal* 111 (468): 1–28.

Lang, Janet M., Kenneth J. Rothman, and Cristina I Cann. 1998. "That Confounded P-value." *Epidemiology* 9 (1): 7–8.

Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* New York: John Wiley & Sons.

Leamer, Edward E. 1983. "Let's Take the Con out of Econometrics." *The American Economic Review* 73

(1): 31–43.

Makel, Matthew C., Jonathan A. Plucker, and Boyd Hegarty. 2012. "Replications in Psychology Research: How often Do They Really Occur? *Perspectives on Psychological Science* 7 (6):537–42.

Manski, Charles F. 2013. *Public Policy in an Uncertain World: Analysis and Decisions*. Cambridge, MA: Harvard University Press.

Mayo, Deborah. 2020. "P-Values on Trial: Selective Reporting of (Best Practice Guides Against) Selective Reporting." *Harvard Data Science Review* 2 (1).

Meager, Rachael. 2019. "Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments." *American Economic Journal: Applied Economics* 11(1): 57–91.

Morris, Carl N. 1983. "Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association* 78 (381): 47–55.

Nesbø, Jo. 2012. *The Bat*. London: Random House.

Neyman, Jerzey, K. Iwaszkiewicz, and St. Kolodziejczyk.1935. "Statistical Problems in Agricultural Experimentation (with discussion)." *Journal of the Royal Statistal Society* 2 (2): 107–80.

Olken, Benjamin A. 2015. "Promises and Perils of Pre-analysis Plans." *Journal of Economic Perspectives* 29 (3): 61–80.

Ritzwoller, David M., and Joseph P. Romano. 2019. "Uncertainty in the Hot Hand Fallacy: Detecting Streaky Alternatives in Random Bernoulli Sequences." Department of Statistics, Stanford University, 2019.

Romer, David. 2020. "In Praise of Confidence Intervals." *AEA Papers and Proceedings* 110: 55–60.

Schanzenbach, Diane Whitmore. 2006. "What Have Researchers Learned from Project STAR? *Brookings Papers on Education Policy* 9: 205–28.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2013. "Life after p-hacking." Paper presented at Meeting of the Society for Personality and Social Psychology, New Orleans, LA, January 13.

Sims, Christopher A., and Harald Uhlig. 1991. "Understanding Unit Rooters: A Helicopter Tour." *Econometrica: Journal of the Econometric Society* 59 (6): 1591–99.

Sinervo, Pekka K. 2002. "Signal Significance in Particle Physics." arXiv preprint hep-ex/0208005

Stern, Hal S. 2016. "A Test by Any Other Name: P Values, Bayes Factors, and Statistical Inference. Multivariate Behavioral Research." 51 (1): 23–29.

Storey, John D., and Robert Tibshirani. 2003. "Statistical Significance for Genomewide Studies." *Proceedings of the National Academy of Sciences* 10 (16): 9440–45.

Trafimow, David. 2014. "Editorial." *Basic and Applied Social Psychology* 36 (1): 1–2, 2014.

Trafimow, David, and Michael Marks. 2015. "Editorial." *Basic and Applied Social Psychology* 37 (1):1–2.

Van der Vaart, A.W. 2000. *Asymptotic Statistics*. Cambridge, UK: Cambridge University Press.

Wagenmakers, Eric-Jan, Ruud Wetzels, Denny Borsboom, and Han L.J. van der Maas. 2011. "Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi: Comment on Bem (2011)." *Journal of Personality and Social Psychology* 100 (3): 426–32.

Wasserstein, Ronald L., and Nicole A. Lazar. 2016. "The ASA Statement on *p*-Values: Context, Process, and Purpose." *The American Statistician* 70 (2): 129–33.

Wasserstein, Ronald L., Allen L. Schirm, and Nicole A. Lazar. 2019. "Moving to a World beyond 'p<0.05'." *The American Statistician* 73(S1): 1–19.

Ziliak, Stephen T., and Deirdre N. McCloskey. 2011. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor: University of Michigan Press.