Statistical Non-Significance in Empirical Economics

By Alberto Abadie*

Statistical significance is often interpreted as providing greater information than non-significance. In this article we show, however, that rejection of a point null often carries very little information, while failure to reject may be highly informative. This is particularly true in empirical contexts that are common in economics, where data sets are large and there are rarely reasons to put substantial prior probability on a point null. Our results challenge the usual practice of conferring point null rejections a higher level of scientific significance than non-rejections. Therefore, we advocate visible reporting and discussion of non-significant results.

> "It is usual and convenient for experimenters to take 5 per cent. as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard ..." R.A. Fisher in *The Design of Experiments* (Fisher, 1935)

Non-significant empirical results (usually in the form of t-statistics smaller than 1.96) relative to some null hypotheses of interest (usually zero coefficients) are notoriously hard to publish in professional/scientific journals (e.g., Andrews and Kasy, 2019; Ziliak and McCloskey, 2008).[1] This state of affairs is in part maintained by the widespread notion that non-significant results are non-informative. After all, lack of statistical significance derives from the absence of extreme or surprising outcomes under the null hypothesis. In this article, we argue that this view of statistical inference is misguided. In particular, we show that non-significant results are informative and argue that they are more informative than significant results in scenarios common in empirical practice in economics.

To discuss the informational content of different statistical procedures, we formally adopt a limited information Bayes perspective. In this setting, agents representing journal readership or the scientific community have priors, $\mathcal{P}$, over some parameters of interest, $\theta \in \Theta$. That is, a member p of $\mathcal{P}$ is a probability density function (with respect to some dominating measure) on Θ. While agents are Bayesian, we will consider a setting where journals (or researchers) report frequentist results, in particular, statistical significance. Agents construct limited information Bayes posteriors based on the reported results of significance

* Department of Economics, MIT, abadie@mit.edu. I thank Isaiah Andrews, Joshua Angrist, Amy Finkelstein, Brigham Frandsen, Andrew Gelman, Guido Imbens, Judith Lok, Ben Olken, and especially Gary Chamberlain and Max Kasy for comments and discussions. Comments by the editor and two referees substantially improved this article. Aubrey Grimshaw provided expert research assistance.

[1] See also Kim and Ji (2015) and Brodeur et al. (2016).

tests. We will deem a statistical result informative when it has the potential to substantially change the beliefs of the agents over a large range of values for θ.

To understand the intuition behind the main result of this article, it is useful to draw attention to two features often present in empirical studies in economics. First, as argued later in this article, for the parameters of interest in empirical studies in economics, there is rarely any reason to put substantial prior belief on a point null hypothesis. Second, decent statistical power for parameter values away from the null is required and often present in applied studies in economics, where data sets are large are becoming larger. Both of these features increase the prior probability of rejection of the null hypothesis. If the probability of rejection is larger than 0.5, then, non-significance is deemed surprising to a Bayesian agent relative to significance, and therefore, more informative than significance in the sense that it induces a larger change in beliefs.

Notice that, like Ioannidis (2005) and others, we restrict our attention to the effect of statistical significance on beliefs. We adopt this framework not because we believe it is (always) representative of empirical practice (in fact, researchers and journals typically report additional statistics, beyond statistical significance), but because isolating the informational content of statistical significance has immediate implications for how we should interpret its occurrence or lack of it. Correct interpretation of statistical significance is important because, while many other statistics are reported in practice, the scientific discussion of empirical results is often framed in terms of the statistical significance of some parameters of interest, and non-significant results may be under-reported or unpublished. Moreover, by considering significance tests in isolation of other statistics, we restrict ourselves to the best possible scenario for the informational content of significance testing. As we demonstrate in Section I, once we start conditioning on the value of other statistics, the results of significance tests soon become completely uninformative.

Previous studies have described the important limitations of significance testing as an inferential tool in the social sciences and other disciplines (see, in particular, Leamer, 1978; Berger, 1985; Berger and Selke, 1999; Sims and Uhlig, 1991; Gelman and Stern, 2006; Ziliak and McCloskey, 2008; Greenland and Poole, 2013; Gelman, 2015; Wasserstein and Lazar, 2016; Amrhein, Korner-Nievergelt and Roth, 2017; McShane et al., 2019).[2] We, too, advise against the use of statistical significance as the primary marker of scientific discovery in empirical studies in the social sciences. However, the pervasiveness of significance testing in social science research suggests that significance tests will remain part of the empirical toolkit, at least for the foreseeable future. If so, it is important to confer an appropriate interpretation to the results of significance tests.

A large literature has focused on the dangers of a larger-than-size prevalence of "false positives" resulting from selective reporting based on significance (p-

[2]Also related to this article are Frankel and Kasy (2018) and Furukawa (2019), which adopt a policy choice perspective to study the question of which research results should be published.

hacking and "filedrawer").[3] This article switches the focus to the informational value of statistical non-significance and argues that in many empirical scenarios non-significant results carry more information than significant ones. Moreover, the informational value of non-significance relative to significance *does not necessarily emanate from the estimation of "precise zeros"*. In Section II.A we propose a metric to measure the informational value of significance and non-significance, and formally show that, under that metric, the informational value of non-significance exceeds the informational value of significance as long as the prior probability of rejection of the null is greater than 0.5. As a consequence, even in empirical settings where power/precision is low relative to conventional requirements (e.g., power exceeding 0.80 is a usual benchmark in the design of experiments), non-significance may substantially outperform significance in terms of the informational value of the result, especially when the prior assigns substantial probability to subsets of Θ away from the null. To illustrate the empirical relevance of this result we use data from economics laboratory experiments (Camerer et al., 2016; Andrews and Kasy, 2019).

The rest of the article is organized as follows. Section I provides a simple example, with normal priors and data, that clarifies the informational content of significance tests. Section II provides finite-sample and large-sample results for a general setting, where priors or data may not be normal. In this section, we also consider the case when the prior exhibits probability mass at the point null. Section III provides a calibration using data from experimental economics. Section IV concludes.

I. A Simple Example

In this section, we consider a simple example with normal priors and data that captures the essence of our argument. In Section II we relax these assumptions and consider the general case where the priors and the distribution of the data are not restricted to be in a particular parametric family.

Assume an agent has a prior $\theta \sim N(\mu, \sigma^2)$ on θ, with $\sigma^2 > 0$. A researcher observes n independent measurements of θ with normal errors mutually independent and independent of θ, and with variance normalized to one. That is, $x_1, \ldots, x_n$ are independent $N(\theta, 1)$, conditional on θ. Then, conditional on θ, we obtain

$$\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \sim N(\theta, 1/n).$$

θ is deemed significant if $\sqrt{n}|\widehat{\theta}_n| > c$, for some $c > 0$. In empirical practice, c is often equal to 1.96, the 0.975-quantile of the standard normal distribution. Suppose

[3] See, e.g., Rosenthal (1979); Ioannidis (2005); Simmons, Nelson and Simonsohn (2011); Brodeur et al. (2016); Wasserstein and Lazar (2016); Amrhein, Korner-Nievergelt and Roth (2017).

the researcher reports on statistical significance. Suppose also that n is common knowledge, so agents know the precision of the estimates. We will calculate the limited information posteriors of the agents conditional on significance and lack thereof. These posteriors are the distributions of θ conditional on $\sqrt{n}|\widehat{\theta}_n| > c$ and $\sqrt{n}|\widehat{\theta}_n| \leq c$. First, notice that

$$\begin{aligned}\Pr(\sqrt{n}|\widehat{\theta}_n| > c|\theta) &= \Pr(\widehat{\theta}_n > c/\sqrt{n}|\theta) + \Pr(-\widehat{\theta}_n > c/\sqrt{n}|\theta) \\ &= \Phi(\sqrt{n}\theta - c) + \Phi(-\sqrt{n}\theta - c). \qquad (1)\end{aligned}$$

Integrating over the prior, we obtain the prior probability of rejection (henceforth, probability of rejection),[4]

$$\Pr(\sqrt{n}|\widehat{\theta}_n| > c) = \Phi\left(\frac{\sqrt{n}\mu - c}{\sqrt{1+n\sigma^2}}\right) + \Phi\left(\frac{-\sqrt{n}\mu - c}{\sqrt{1+n\sigma^2}}\right). \qquad (2)$$

The limited information posteriors given significance and non-significance are:

$$p\big(\theta\big|\sqrt{n}|\widehat{\theta}_n| > c\big) = \frac{\frac{1}{\sigma}\phi\left(\frac{\theta-\mu}{\sigma}\right)\Big(\Phi(\sqrt{n}\theta - c) + \Phi(-\sqrt{n}\theta - c)\Big)}{\Phi\left(\frac{\sqrt{n}\mu - c}{\sqrt{1+n\sigma^2}}\right) + \Phi\left(\frac{-\sqrt{n}\mu - c}{\sqrt{1+n\sigma^2}}\right)}, \qquad (3)$$

and

$$p\big(\theta\big|\sqrt{n}|\widehat{\theta}_n| \leq c\big) = \frac{\frac{1}{\sigma}\phi\left(\frac{\theta-\mu}{\sigma}\right)\Big(1 - \Phi(\sqrt{n}\theta - c) - \Phi(-\sqrt{n}\theta - c)\Big)}{1 - \Phi\left(\frac{\sqrt{n}\mu - c}{\sqrt{1+n\sigma^2}}\right) - \Phi\left(\frac{-\sqrt{n}\mu - c}{\sqrt{1+n\sigma^2}}\right)}. \qquad (4)$$

The two posteriors, along with the normal prior, are plotted in Figure 1 for $\mu = 1$, $\sigma = 1$, $c = 1.96$, and $n = 10$. This figure illustrates the informational value of a significance test. Rejection of the null carves probability mass around zero in the limited information posterior, while failure to reject concentrates probability mass around zero.

[4]This calculation uses the following fact of integration

$$\int \Phi\left(\frac{\lambda-\theta}{\xi}\right)\frac{1}{\sigma}\phi\left(\frac{\theta-\mu}{\sigma}\right)d\theta = \Phi\left(\frac{\lambda-\mu}{\sqrt{\sigma^2+\xi^2}}\right)$$

for arbitrary real λ and μ and positive σ and ξ. Alternatively, the result can be easily derived after noticing that the distribution of $\widehat{\theta}_n$ integrated over the prior is normal with mean μ and variance $\sigma^2 + 1/n$.

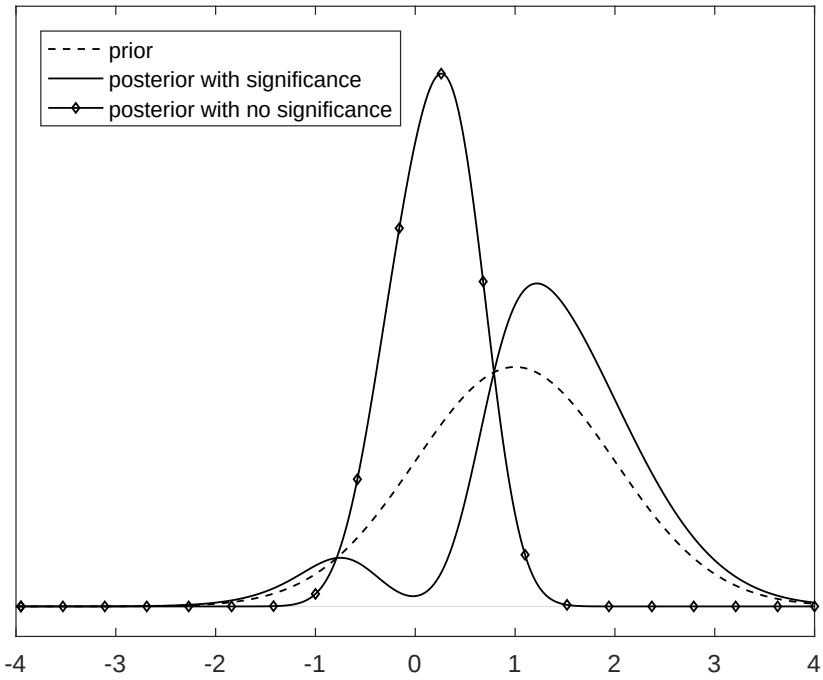

Figure 1. Posterior Distributions After a Significance Test

As noted earlier, the discrepancy between a prior and a posterior distribution provides a basic measure of informativeness of a statistical result. It tells us the extent to which the additional information contained in the results changes beliefs on the distribution of the parameter of interest. In Section II.A we will present a specific metric of informativeness based on such a discrepancy. However, in the absence of a specific metric of informativeness, a mere visual inspection of Figure 1 indicates that failure to reject carries substantial information.

Figure 2 shows how prior and posteriors after significance compare as a function of the sample size. When n is small, significance affects the posterior over a large range of values. When n is large, significance provides only local to zero information. That is, significance is not informative in large samples. This is explained by the fact that the probability of rejection in Equation (2) converges to one as the sample size increases. Intuitively, the occurrence of an event (rejection of the null) that has large probability under the prior should not have a substantial effect on beliefs. In contrast, by the law of total probability, it follows that conditional on non-significance probability mass concentrates around zero as n increases, so the prior and the posterior differ substantially in this case.[5] That is,

[5]To preserve a visually informative scale and because they concentrate around zero as n increases, we omit the posteriors without significance from Figure 2. For the values of μ, σ, c, and n adopted

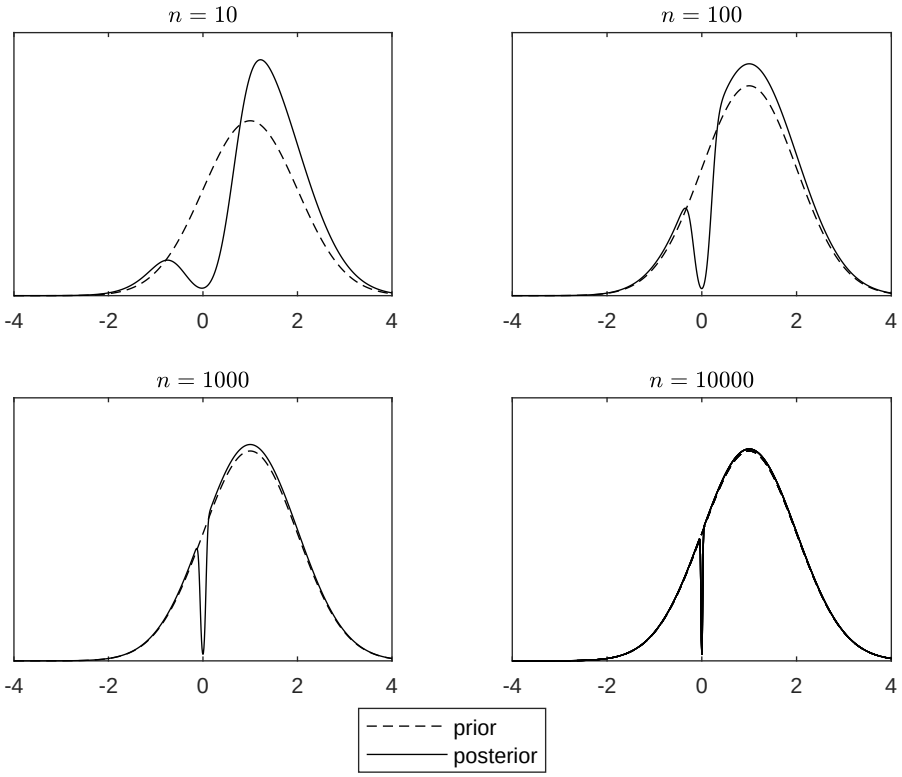

FIGURE 2. PRIOR AND POSTERIOR WITH SIGNIFICANCE FOR DIFFERENT SAMPLE SIZES

the occurrence of an event (non-rejection of the null) that is very unlikely given the prior has a large effect on beliefs.

Equations (3) and (4) report limited information posteriors. The full information posterior is

$$p\big(\theta|x_1,\ldots,x_n\big) = \frac{1}{\sigma_n}\phi\left(\frac{\theta-\mu_n}{\sigma_n}\right),$$

where

$$\mu_n = \frac{\mu + n\sigma^2\widehat{\theta}_n}{1+n\sigma^2},$$

and

$$\sigma_n^2 = \frac{\sigma^2}{1+n\sigma^2}.$$

So, in this very particular context, knowledge of $\widehat{\theta}_n$ is sufficient to go back to the full information posterior. The same is true for the combined information given

for Figures 1 and 2, the implied probabilities of rejection, $\Pr(\sqrt{n}|\widehat{\theta}_n| > c)$, are 0.7028 ($n = 10$), 0.9052 ($n = 100$), 0.9700 ($n = 1000$), and 0.9905 ($n = 10000$).

by the two-sided p-value, $2\Phi(-\sqrt{n}|\widehat{\theta}_n|)$ and the sign of $\widehat{\theta}_n$ (or a one-sided p-value on its own). More generally, under regularity conditions, there exist asymptotically sufficient statistics of the same dimension as the number of unknown parameters in a model. In particular, under regularity conditions, the maximum likelihood estimators of the unknown parameters of a model are asymptotically sufficient. Conditional on those statistics, *statistical significance carries no information*. This result has two important implications. First, it underscores, in contrast to the R.A. Fisher quote in the preamble of this article, the importance of *not ignoring* results that fail to attain statistical significance.[6] Instead, statistical inference should rely on statistics that may better describe the full-information posterior distribution of the parameters of interest. Confidence intervals, which have an interpretation as Bayes credible regions in large samples, seem a natural choice. Second, it demonstrates that at least for the notion of informativeness adopted in this article, the study of the informational content of significance tests requires a limited information perspective. Under full information (e.g., after conditioning on sufficient statistics) statistical significance is irrelevant in terms of its informational content.[7]

The results of this section have immediate counterparts in large sample settings with asymptotically normal distributions. They can also be generalized to non-parametric settings, as we demonstrate next.

II. General Case

A. Finite Sample Results

Results like that in Figure 1 are rather general and do not depend on normal priors or data. Consider a test statistic, $\widehat{T}_n$, such that rejection of the null is given by $\widehat{T}_n > c$. Let $p(\cdot)$ be a prior on θ, and $p(\cdot|\widehat{T}_n > c)$ and $p(\cdot|\widehat{T}_n \leq c)$ be the limited information posteriors under significance and non-significance, respectively. Regardless of the shape of the prior and/or the distribution of the data, by the law of total probability

$$p(\theta) = p(\theta|\widehat{T}_n \leq c)\Pr(\widehat{T}_n \leq c) + p(\theta|\widehat{T}_n > c)\Pr(\widehat{T}_n > c).$$

[6] It is worth noticing that Fisher's stance on the value of statistical significance seemed to soften in his late writings. In particular, Fisher (1958) (quoted in Amrhein, Korner-Nievergelt and Roth (2017)) states: "tests of statistical significance are used as an aid to judgment, and should not be confused with automatic acceptance tests, or 'decision functions'."

[7] It is, however, possible to increase the amount of information of the agents without making statistical significance irrelevant. In a supplemental appendix we show that results similar to those in this section hold when the information sets of the agents are as in this section but additionally include the sign of $\widehat{\theta}_n$.

Rearranging terms, we obtain

$$\left|1-\frac{p(\theta|\widehat{T}_n\leq c)}{p(\theta)}\right| = \left(\frac{\Pr(\widehat{T}_n>c)}{\Pr(\widehat{T}_n\leq c)}\right)\left|1-\frac{p(\theta|\widehat{T}_n> c)}{p(\theta)}\right| \tag{5}$$

for $\Pr(\widehat{T}_n \leq c) > 0$ and θ such that $p(\theta) > 0$. The absolute value expressions on both sides of Equation (5) measure the local (at θ) informativeness of significance (right) and non-significance (left). They are zero when the posterior densities with significance (right) / non-significance (left) are equal to the prior density. Otherwise, they measure the discrepancy between prior and posterior densities at θ as a fraction of the prior density at θ.

Equation (5) implies that the local informativeness of non-significance relative to significance at θ is solely determined by the ratio $\Pr(\widehat{T}_n > c)/\Pr(\widehat{T}_n \leq c)$, which (remarkably) does not depend on θ. That is, the ratio $\Pr(\widehat{T}_n > c)/\Pr(\widehat{T}_n \leq c)$ provides also a global measure of the informativeness of non-significance relative to significance.[8]

The connection between the probability of rejection, the power of the test, and the prior distribution of θ is given by

$$\Pr(\widehat{T}_n > c) = \int \Pr(\widehat{T}_n > c|\theta)\, p(\theta)\, d\nu,$$

where, for $\theta \neq 0$, $\Pr(\widehat{T}_n > c|\theta)$ is the power of the test. Now, for $\Pr(\widehat{T}_n > c) = 0.5$, which typically indicates a rather underpowered setting or a large prior probability mass at the point null, Equation (5) implies that non-significance is exactly as informative as significance. Moreover, the relative informativeness of non-significance increases with the probability of rejecting the null. In turn, the probability of rejection depends on the central tendency of the prior, but also crucially on its dispersion. That is, the probability of rejection could be substantial even if the power of the test evaluated at the center of the prior is close to zero.[9] In particular, the adoption of a diffuse prior ("objective Bayes")

[8]Notice that the result in Equation (5) can easily be expressed in terms of the total variation distance, to obtain a global measure of the discrepancy between the prior and posterior distributions,

$$\left(\frac{1}{2}\int \left|p(\theta|\widehat{T}_n\leq c) - p(\theta)\right| d\nu\right) = \left(\Pr(\widehat{T}_n > c)/\Pr(\widehat{T}_n\leq c)\right)\left(\frac{1}{2}\int \left|p(\theta|\widehat{T}_n> c) - p(\theta)\right| d\nu\right),$$

for some dominating measure, ν. See, e.g., DasGupta (2008) section 2.1 for the definition and properties of the total variation metric. Total variation is by no means the only possible measure of discrepancy between two distributions, but is particularly well-suited for our purposes, because the relative informativeness of significance *vs.* non-significance in terms of total variation follows directly from the law of total probability.

[9]To illustrate this point, consider the setting of Section I, with prior $\theta \sim N(0.1, 1)$ and $n = 100$. Then, the power evaluated at $\mu = 0.1$ is equal to 0.17 (Equation (1)). Notice, however, that the probability of rejection, integrated over the prior, is equal to 0.85 (Equation (2)). This distinction is important

centered near the null value of θ increases the probability of rejection and, as a result, reduces the relative informativeness of significance. In Section III, we calibrate a prior using data from economics laboratory experiments, obtaining a probability of rejection greater than 0.5. Finally, it is worth mentioning that the result in Equation (5) is general and does not rely on the nature of the null hypothesis (point null or composite).

B. Large Sample Analysis

To extend the large sample results of Section I beyond normal priors and data, we will consider testing procedures with certain basic properties. First, we will consider tests that have asymptotic size equal to some constant $\alpha \in (0, 1)$,

$$\Pr\left(\widehat{T}_n > c \middle| \theta = 0\right) \to \alpha. \tag{6}$$

Let $\beta_n(\theta) = \Pr(\widehat{T}_n \leq c|\theta)$. We will require that the probability of type II error converges to zero exponentially. That is,

$$\limsup_{n\to\infty} \frac{1}{n} \log \beta_n(\theta) < 0 \tag{7}$$

for every $\theta \neq 0$. Equation (7) typically follows from large deviations results and implies that the test is consistent against fixed alternatives. Finally, we will rule out perfect local asymptotic power

$$\int \liminf_{n\to\infty} \beta_n(z/\sqrt{n})\, dz > 0. \tag{8}$$

These are very weak requirements, as they hold for most testing procedures used in practice.

Continuous Prior. — We will first assume a prior that is absolutely continuous with respect to the Lebesgue measure, with a version of the density that is positive and continuous at zero. By dominated convergence, we obtain:

$$\Pr\left(\widehat{T}_n > c\right) \to 1.$$

We first derive the posterior densities under significance,

$$p(0|\widehat{T}_n > c) = \frac{\Pr\left(\widehat{T}_n > c \middle| \theta = 0\right)}{\Pr\left(\widehat{T}_n > c\right)} p(0) \to \alpha\, p(0),$$

because, as reported in Ioannidis, Stanley and Doucouliagos (2017), statistical power in empirical studies in economics may be low for values of θ close to an average of estimates across different studies.

and

$$p(\theta|\widehat{T}_n > c) = \frac{\Pr\left(\widehat{T}_n > c|\theta\right)}{\Pr\left(\widehat{T}_n > c\right)} p(\theta) \to p(\theta),$$

for $\theta \neq 0$. So, in large samples significance only changes beliefs locally around zero. The posterior density at $\theta = 0$ after non-significance is

$$p(0|\widehat{T}_n \leq c) = \frac{\Pr\left(\widehat{T}_n \leq c|\theta = 0\right)}{\Pr\left(\widehat{T}_n \leq c\right)} p(0) \to \infty.$$

For $\theta \neq 0$, the posterior density after non-significance is

$$p(\theta|\widehat{T}_n \leq c) = \frac{\Pr\left(\widehat{T}_n \leq c|\theta\right)}{\Pr\left(\widehat{T}_n \leq c\right)} p(\theta).$$

Calculating the limit of $p(\theta|\widehat{T}_n \leq c)$ when $\theta \neq 0$ is complicated by the fact that both $\Pr(\widehat{T}_n \leq c|\theta)$ and $\Pr(\widehat{T}_n \leq c)$ converge to zero. By Equation (7), for any fixed $\theta \neq 0$, $\Pr(\widehat{T}_n \leq c|\theta)$ converges to zero at an exponential rate as a function of n. We will now show that Equation (8) implies that the rate of decay of $\Pr(\widehat{T}_n \leq c)$ is not faster than polynomial. By change of variable $z = n^{1/2}\theta$ and Fatou's lemma, we obtain

$$\begin{aligned}
\liminf_{n\to\infty} n^{1/2} \Pr(\widehat{T}_n \leq c) &= \liminf_{n\to\infty} n^{1/2} \int \beta_n(\theta)\, p(\theta)\, d\theta \\
&= \liminf_{n\to\infty} \int \beta_n(z/\sqrt{n})\, p(z/\sqrt{n})\, dz \\
&\geq \int \liminf_{n\to\infty} (\beta_n(z/\sqrt{n})\, p(z/\sqrt{n}))\, dz \\
&= \int \liminf_{n\to\infty} \beta_n(z/\sqrt{n}) \lim_{n\to\infty} p(z/\sqrt{n})\, dz \\
&= p(0) \int \liminf_{n\to\infty} \beta_n(z/\sqrt{n})\, dz > 0.
\end{aligned}$$

As a result,

$$p(\theta|\widehat{T}_n \leq c) \to 0,$$

for $\theta \neq 0$. That is, like in the normal case of Section I, conditional on non-significance the posterior converges to a degenerate distribution at zero.

To sum up, we have shown that, in a large sample non-parametric setting without prior probability mass at the point null, non-significance can be extremely informative while significance carries no information. We will next consider the case where the prior exhibits a probability mass at the point null.

PRIOR WITH PROBABILITY MASS AT ZERO. — We now consider the case when the prior has probability mass q at zero, with $0 < q < 1$. Then

$$\Pr\left(\widehat{T}_n > c\right) \to q\alpha + (1-q) \in (\alpha, 1).$$

Now, the posterior after significance is,

$$p(0|\widehat{T}_n > c) = \frac{\Pr\left(\widehat{T}_n > c|\theta = 0\right)}{\Pr\left(\widehat{T}_n > c\right)} p(0) \to \left(\frac{\alpha}{q\alpha + (1-q)}\right) q < q,$$

and

$$p(\theta|\widehat{T}_n > c) = \frac{\Pr\left(\widehat{T}_n > c|\theta\right)}{\Pr\left(\widehat{T}_n > c\right)} p(\theta) \to \left(\frac{1}{q\alpha + (1-q)}\right) p(\theta) > p(\theta),$$

for $\theta \neq 0$. In contrast to the continuous prior case, significance changes beliefs away from zero in large samples. If we start with a prior that assigns a large probability to $\theta = 0$, significance may greatly affect beliefs over regions for θ that are away from zero. Notice, however, that for moderate values of q the effect of significance on beliefs may be negligible in large samples. Figure 3 shows the limit of $p(\theta|\widehat{T}_n > c)/p(\theta)$ as a function of q, for $\theta \neq 0$ and $\alpha = 0.05$. This limit is close to one for modest values of q. In order for significance to at least double the value of the probability density function at values θ such that $\theta \neq 0$ we need $q \geq 1/(2(1-\alpha)) = 0.5263$. Notice that reducing the size of the test, α, does not substantially change the value of the limit of $p(\theta|\widehat{T}_n > c)/p(\theta)$, except for very large values of q. Regardless of the size of the test, q needs to be larger than 0.5 in order for significance to double the probability density function of beliefs at non-zero values of θ.

The posterior after non-significance is,

$$p(0|\widehat{T}_n \leq c) = \frac{\Pr\left(\widehat{T}_n \leq c|\theta = 0\right)}{\Pr\left(\widehat{T}_n \leq c\right)} p(0) \to \frac{1-\alpha}{q(1-\alpha)} q = 1,$$

and for $\theta \neq 0$,

$$p(\theta|\widehat{T}_n \leq c) = \frac{\Pr\left(\widehat{T}_n \leq c|\theta\right)}{\Pr\left(\widehat{T}_n \leq c\right)} p(\theta) \to 0.$$

Like in the case of a continuous prior, non-significance seems to have a stronger effect on beliefs than significance in settings that seem most relevant for empirical practice in economics (i.e., settings with moderate values for a prior probability mass at the point null.)

Some remarks about priors with probability mass at a point null are in order. First, it is difficult to think of relevant settings in empirical economics where reasonable prior beliefs assign probability mass to point nulls. For example, beliefs

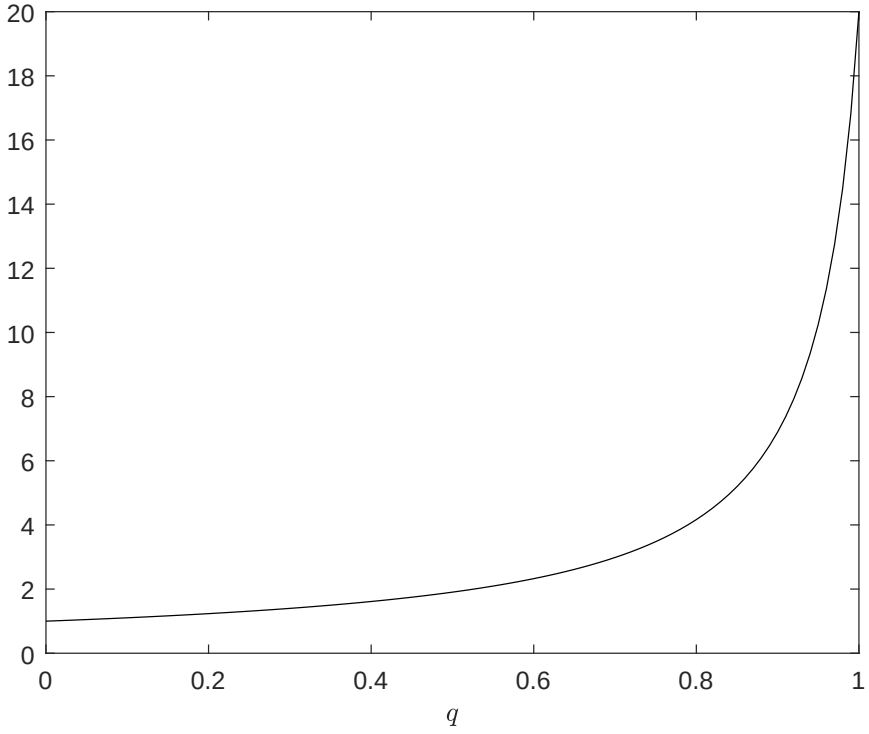

FIGURE 3. LIMIT OF $p(\theta|\widehat{T}_n > c)/p(\theta)$ AS A FUNCTION OF q ($\theta \neq 0$, $\alpha = 0.05$)

on the causal effect of a policy intervention may sometimes concentrate smoothly around zero, but more rarely in such a way that a large probability mass at zero is a good description of a reasonable prior.[10,11] Moreover, priors with probability mass at a point null generate a drastic discrepancy, known as Lindley's paradox, between frequentist and Bayesian testing procedures (see, e.g., Berger, 1985). Lindley's paradox arises in settings with a fixed value of $\widehat{T}_n$ and a large n. In those settings, frequentists would reject the null hypothesis when $\widehat{T}_n > c$. Bayesians, however, may find that the posterior probability of the point null far exceeds the

[10]See Berkson (1938); Berger (1985); De Long and Lang (1992); McShane et al. (2019) for related discussions. One could try to reinterpret significance tests as tests of the implicit null "θ is close to zero". In a supplemental appendix, we study the problem of testing the null that the parameter θ belongs to an interval.

[11]This is not to say that there are not settings where a point null hypothesis could be highly privileged. Fisher (1935) motivated the development of statistical tests using the famous "lady tasting tea" example. The null hypothesis stated that a certain lady could not discern, by tasting only, whether tea or milk had been added first to a cup. It is possible that in this example the null hypothesis was highly privileged. Similarly, statistical testing has been applied to detect extrasensory perception, where the belief in the null hypothesis of no extrasensory perception may be strong. In microarray studies, scientists may be interested in finding genes involved in the development of a medical condition. Efron and Hastie has called these exercises "fishing expeditions", because for each gene the null hypothesis of no effect is highly privileged (Efron and Hastie, 2016). Such settings do not seem common in economics.

posterior probability of the alternative. Lindley's paradox can be explained by the fact that, as n increases, the distribution of the test statistic diverges under the alternative. Therefore, a fixed value of the test statistic as n increases can only be explained by the null hypothesis, provided that the prior assigns probability mass to the null. Notice that conditioning on the event $\{\widehat{T}_n \leq c\}$ (as opposed to conditioning on the value of $\widehat{T}_n$) is not subject to Lindley's paradox and it may be the natural choice to evaluate a testing procedure for which significance depends on the value of $\{\widehat{T}_n \leq c\}$ only.[12]

III. Calibration using data from economics laboratory experiments

In this section, we use data from economics laboratory experiments (Camerer et al., 2016; Andrews and Kasy, 2019) to calibrate the parameters of the prior and the number of available observations in the posterior density formulas of Section I. The goal is to approximate the posterior densities with and without significance in a setting that roughly resembles a realistic scenario. Interestingly, the primary definition of a successful replication in Camerer et al. (2016) is a "significant effect in the same direction as in the original study," without a reference to the magnitude of the coefficients in the original and replication studies. This choice illustrates the extent to which statistical significance is viewed as a primary attribute of scientific discovery in economics. Moreover, for this dataset Andrews and Kasy (2019) estimate a large jump in the probability of publication for studies that attain statistical significance at the 5 percent level.

We make use of the fact that the data in Camerer et al. (2016) and Andrews and Kasy (2019) contain the original values of test statistics for a set of 18 experimental laboratory studies published in two leading economics journals and the corresponding test statistic values for replications of those 18 studies. In particular, we make use of the z-statistics computed in Andrews and Kasy (2019) for the replication studies.

We consider θ equal to the probability limit of the rescaled z-statistic, $2\,\widehat{z}_n/\sqrt{n}$, where n is the number of participants in an experiment. We calibrate a prior for θ using the distribution of the rescaled replication statistics, $2\,\widehat{z}^*_{j,n^*_j}/\sqrt{n^*_j}$, $j = 1, \ldots, 18$. In the previous expression, $\widehat{z}^*_{j,n^*_j}$ is the replication value of a z-statistic for the point null evaluated in study j, and n^*_j is the number of participants in the replication. We make this particular choice for the sake of simplicity and because, for the simple case when $\widehat{z}_n$ is the usual two-sample z-statistic with equal number of observations on the treatment and control arms, θ becomes the

[12] In particular, notice that if $\Pr(\widehat{T}_n > c|\theta = 0) < \Pr(\widehat{T}_n > c)$ then, by Bayes rule, we obtain $\Pr(\theta = 0|\widehat{T}_n > c) < q$. This is in contrast with Lindley's paradox, under which conditioning on a value of $\widehat{T}_n = t$ such that $t > c$ may result in an increase in the probability of the null, $\Pr(\theta = 0|\widehat{T}_n = t, t > c) > q$ (see Berger, 1985).

average treatment effect measured in standard deviations units:

$$\theta = \frac{\tau_1 - \tau_0}{\lambda},$$

where $\lambda = \sqrt{(\lambda_1^2 + \lambda_0^2)/2}$; τ_1 and τ_0 are average outcomes with and without treatment, respectively; and λ_1 and λ_0 are the standard deviations of the outcome with and without treatment, respectively.[13] We calibrate the parameters μ and σ^2 in Section I to be the mean and variance of $2\,\widehat{z}^*_{j,n^*_j}/\sqrt{n^*_j}$, $j = 1, \ldots 18$ ($\mu = 0.3864$ and $\sigma = 0.3680$) and we calibrate the number of observations to be the median number of participants in the original studies ($n = 120$).[14,15] Notice that the distribution of the replication statistics is conditional on publication of the original studies, and may not correspond to a reasonable prior on θ unconditional on publication.[16]

Figure 4 shows the calibrated prior and posteriors with and without significance for the experimental economics data set. In this scenario, there is no indication that significance conveys more information than non-significance. In these data, however, there is substantial evidence of publication bias on the basis of statistical significance (see Andrews and Kasy, 2019).

Finally, notice that the empirical context adopted in this section may be one of low rejection probability, if the effects investigated in laboratory experiments in economics are believed to be small in magnitude relative to sample sizes. Even in this setting, for the calibrated values of μ, σ, and n, we obtain a value of 0.5627 for the probability of rejection, which implies that non-significance is about 29 percent more informative than significance ($0.5627/(1 - 0.5627) = 1.2868$) under the metric of informativeness employed in Equation (5). The informativeness of non-significance relative to significance will be even larger in empirical settings with higher probability of rejection. On the other hand, there are at least two reasons why the calibrated prior that we adopt in this section may artificially inflate the rejection probability and, as a result, the relative informativeness of

[13] $\theta = (\tau_1 - \tau_0)/\lambda$ is the normalized difference in Abadie and Imbens (2011) and Imbens and Rubin (2015).

[14] In a quantile-quantile plot, the distribution of the rescaled replication statistics closely matches a normal distribution.

[15] Because, in the setting of this section, the distribution of the published z-statistics, $\widehat{z}_n$, is approximately normal with mean $(\sqrt{n}/2)\theta$ and variance one, the limited information posterior formulas of Section I apply with n replaced by $n/4$.

[16] Whether or not the distribution of replication statistics approximates a reasonable prior for θ, unconditional on publication of the original studies, depends on the specific nature of the publication process and on the process that generates θ across experiments/replications. For example, if the nature of the publication/file-drawer process is such that experiments on the same realization of θ are repeated (perhaps by independent research teams) until publication, then, the distribution of θ conditional on publication is the same as the unconditional distribution, and the calibrated prior is a noised-up version of the distribution of θ. The same applies if θ is not fixed from experiment to replication, but resampled from the same distribution as in the original study. In general, however, it is not possible to rigorously calibrate a prior for the unconditional distribution of θ without restrictions on the θ-generating process and on the nature of the publication process.

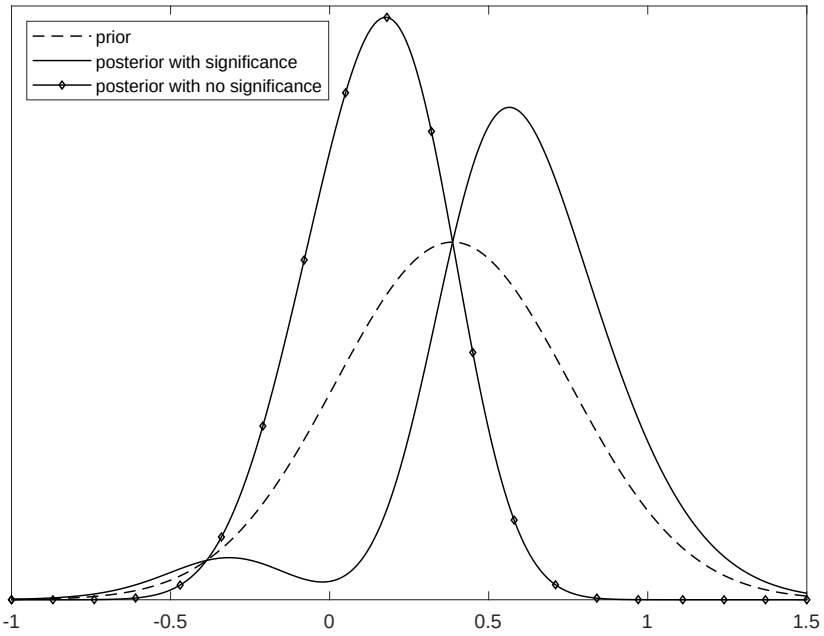

Figure 4. Prior and Posterior Densities Calibrated to Experimental Economics Data

non-significance. First, depending on the nature of the publication process, a prior distribution on θ conditional on publication of the original study may feature smaller probability mass in the vicinity of the null, relative to the unconditional distribution, which increases the probability of rejection. In addition, the calibrated distribution is based on noised-up versions of the values of θ in the replication studies, which may increase the dispersion of the prior and, therefore, the probability of rejection.

IV. Conclusions

Significance testing on a point null is the most prevalent form of inference in empirical economics. In this article, we have shown that rejection of a point null often carries very little information, while failure to reject is highly informative. This is especially true in empirical contexts that are common in economics, where data sets are large (and, if anything, are becoming larger) and where there are rarely reasons to put substantial prior probability on a point null. Our results challenge the usual practice of conferring point null rejections a higher level of scientific significance than non-rejections. In consequence, we advocate visible reporting and discussion of non-significant results in empirical practice (e.g., as in Abadie, 2006; Abdulkadiroğlu, Angrist and Pathak, 2014; Angrist et al., 2019; Krueger and Malečková, 2003). More generally, as discussed in Ziliak and Mc-

Closkey (2008), Gelman (2018), McShane et al. (2019) and many others, the weight of statistical evidence should not be primarily assessed on the basis of statistical significance. Other factors, such as the magnitude and precision of the estimates, the plausibility and novelty of the results, and the quality of the data and the research design, should be carefully evaluated alongside discussions of statistical significance or of the magnitude of p-values.

While this article does not directly address the problems of publication and "file drawer" biases (Rosenthal, 1979; De Long and Lang, 1992; Furukawa, 2019), our results imply that in settings where publication depends on statistical significance (see, e.g., Andrews and Kasy, 2019), the process of publication may discard results of high informational value in favor of less informative results. To our knowledge, this additional avenue through which selective publication based on statistical significance distorts inference has not been previously identified in the literature.

REFERENCES

Abadie, Alberto. 2006. "Poverty, Political Freedom, and the Roots of Terrorism." *American Economic Review (Papers and Proceedings)*, 96(2): 50–56.

Abadie, Alberto, and Guido W. Imbens. 2011. "Bias-Corrected Matching Estimators for Average Treatment Effects." *Journal of Business & Economic Statistics*, 29(1): 1–11.

Abdulkadiroğlu, Atila, Joshua Angrist, and Parag Pathak. 2014. "The Elite Illusion: Achievement Effects at Boston and New York Exam Schools." *Econometrica*, 82(1): 137–196.

Amrhein, Valentin, Fränzi Korner-Nievergelt, and Tobias Roth. 2017. "The Earth Is Flat ($p > 0.05$): Significance Thresholds and the Crisis of Unreplicable Research." *PeerJ*, 5: e3544.

Andrews, Isaiah, and Maximilian Kasy. 2019. "Identification of and Correction for Publication Bias." *American Economic Review*, 109(8): 2766–2794.

Angrist, Joshua D., Victor Lavy, Jetson Leder-Luis, and Adi Shany. 2019. "Maimonides Rule Redux." *American Economic Review: Insights*. Forthcoming.

Berger, James O. 1985. *Statistical Decision Theory and Bayesian Analysis.* New York: Springer-Verlag.

Berger, James O., and Thomas Selke. 1999. "Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence." *Journal of the American Statistical Association*, 82: 112–122.

Berkson, Joseph. 1938. "Some Difficulties of Interpretation Encountered in the Application of the Chi-square Test." *Journal of the American Statistical Association*, 33(203): 526–536.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. 2016. "Star Wars: The Empirics Strike Back." *American Economic Journal: Applied Economics*, 8(1): 1–32.

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science*, 351(6280): 1433–1436.

DasGupta, Anirban. 2008. *Asymptotic Theory of Statistics and Probability.* Springer Science & Business Media.

De Long, J. Bradford, and Kevin Lang. 1992. "Are All Economic Hypotheses False?" *Journal of Political Economy*, 100(6): 1257–1272.

Efron, Bradley, and Trevor Hastie. 2016. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science.* New York: Cambridge University Press.

Fisher, Ronald A. 1935. *The Design of Experiments.* Edinburgh: Oliver and Boyd.

Fisher, Ronald A. 1958. *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd. 13th edition.

Frankel, Alexander, and Maximilian Kasy. 2018. "Which findings should be published?" Working Paper.

Furukawa, Chishio. 2019. "Publication Bias under Aggregation Frictions: Theory, Evidence, and a New Correction Method." Working paper.

Gelman, Andrew. 2015. "What Hypothesis Testing Is All About. (Hint: It's Not What You Think.) [Blog post]." `http://andrewgelman.com/2015/03/02/what-hypothesis-testing-is-all-about-hint-its-not-what-you-think/`, Posted on March 2, 2015. Reposted on May 4, 2017. Accessed on June 5, 2018.

Gelman, Andrew. 2018. "The Failure of Null Hypothesis Significance Testing When Studying Incremental Changes, and What to Do About It." *Personality and Social Psychology Bulletin*, 44(1): 16–23.

Gelman, Andrew, and Hal Stern. 2006. "The Difference Between "Significant" and "Not Significant" Is Not Itself Statistically Significant." *The American Statistician*, 60(4): 328–331.

Greenland, Sander, and Charles Poole. 2013. "Living with P values: Resurrecting a Bayesian Perspective on Frequentist Statistics." *Epidemiology*, 24(1): 62–68.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge: Cambridge University Press.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLOS Medicine*, 2(8): 696–701.

Ioannidis, John P. A., T. D. Stanley, and Hristos Doucouliagos. 2017. "The Power of Bias in Economics Research." *The Economic Journal*, 127(605): F236–F265.

Kim, Jae H., and Philip Inyeob Ji. 2015. "Significance Testing in Empirical Finance: A Critical Review and Assessment." *Journal of Empirical Finance*, 34: 1–14.

Krueger, Alan B., and Jitka Malečková. 2003. "Education, Poverty and Terrorism: Is There a Causal Connection?" *Journal of Economic Perspectives*, 17(4): 119–144.

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

McShane, Blakeley B., David Gal, Andrew Gelman, Christian Robert, and Jennifer L. Tackett. 2019. "Abandon Statistical Significance." *The American Statistician*, 73(sup1): 235–245.

Rosenthal, Robert. 1979. "The File Drawer Problem and Tolerance for Null Results." *Psychological Bulletin*, 86(3): 638–641.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science*, 22(11): 1359–1366.

Sims, Christopher A, and Harald Uhlig. 1991. "Understanding Unit Rooters: A Helicopter Tour." *Econometrica*, 59(6): 1591–1599.

Wasserstein, Ronald L., and Nicole A. Lazar. 2016. "The ASA's Statement on p-Values: Context, Process, and Purpose." *The American Statistician*, 70(2): 129–133.

Ziliak, Stephen T., and Deirdre. N. McCloskey. 2008. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives.* Ann Arbor: University of Michigan Press.