Statistical Non-Significance in Empirical Economics

Alberto Abadie
MIT

June 2018

Abstract

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 typical and even prevalent 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 a visible reporting and discussion of non-significant results.

Alberto Abadie, Department of Economics, MIT, abadie@mit.edu. I thank Isaiah Andrews, Joshua Angrist, Amy Finkelstein, Guido Imbens, Judith Lok, Ben Olken, and especially Gary Chamberlain and Max Kasy for comments and discussions. Aubrey Grimshaw provided expert research assistance.

> "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)

1. Introduction

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 (see, e.g., Andrews and Kasy, 2017; Ziliak and McCloskey, 2008). 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 that are common, even prevalent, 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 appropriate measure) on Θ. While agents are Bayesian, we will consider a setting where journals report frequentist results, in particular, statistical significance. Agents construct limited information Bayes posteriors based on the reported results of significance 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.

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, 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 statistical significance of some parameters of interest, and non-significant results may be under-reported or unpublished.

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; Sims and Uhlig, 1991; Gelman and Stern, 2006; Ziliak and McCloskey, 2008; Gelman, 2015; McShane et al., 2017). 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.

The rest of the article is organized as follows. Section 2 provides a simple example, with normal priors and data, that clarifies the informational content of significance tests. Section 3 provides finite-sample and large-sample results

for a general setting, where priors or data many not be normal. In this section, we also consider the case when the prior exhibits probability mass at the point null. Point nulls are often seen as proxies for approximate nullness. In section 4 we formalize this notion by considering the problem of testing an interval null. In section 5 we consider posteriors that condition the result of a significance test and the sign of the estimate of interest. Section 6 provides a calibration using data from experimental economics. Section 7 concludes.

2. 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 3 we will consider the 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 measurement 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 $N(\theta, 1)$. Let

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

θ is deemed significant if $\sqrt{n}|\widehat{\theta}| > c$, for some $c > 0$. In empirical practice, c is often equal to 1.96, the 0.975-quantile of the standard normal distribution. Suppose a journal reports on statistical significance. 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 $\sqrt{n}|\widehat{\theta}| > c$ and $\sqrt{n}|\widehat{\theta}| \leq c$. First, notice that

$$\Pr(\sqrt{n}|\widehat{\theta}| > c|\theta) = \Pr(\widehat{\theta} > c/\sqrt{n}|\theta) + \Pr(-\widehat{\theta} > c/\sqrt{n}|\theta)$$

$$= \Phi(\sqrt{n}\theta - c) + \Phi(-\sqrt{n}\theta - c).$$

Therefore,[1]

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

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

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

and

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

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. Notice that failure to reject carries substantial information, even in the rather under-powered setting generated by the values of μ, σ, c, and n adopted for Figure 1, which imply $\Pr(\sqrt{n}|\widehat{\theta}| > c) = 0.7028$.

[1]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 $\widehat{\theta}$ integrated over the prior is normal with mean μ and variance $\sigma^2 + 1/n$.

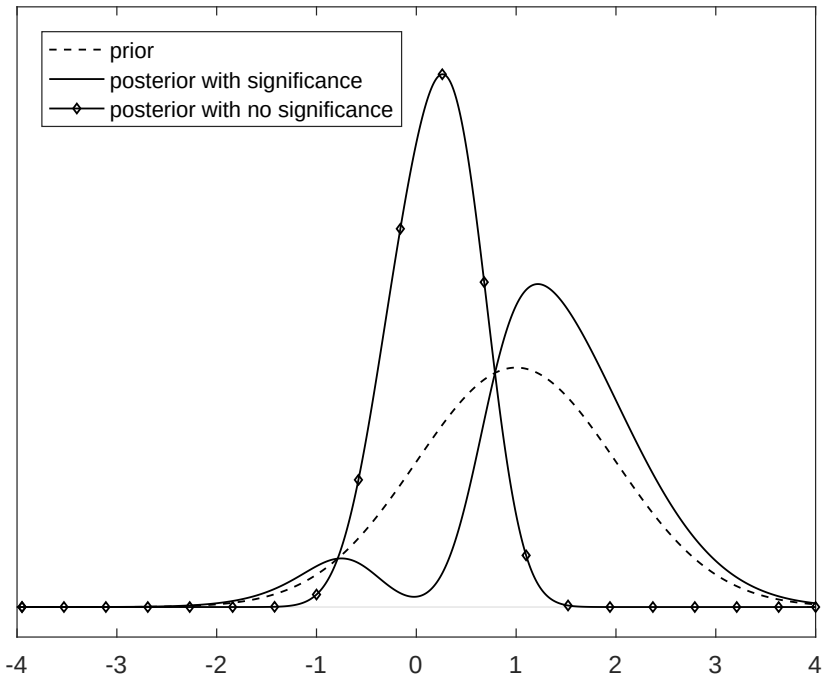

Figure 1: Posterior Distributions After a Significance Test

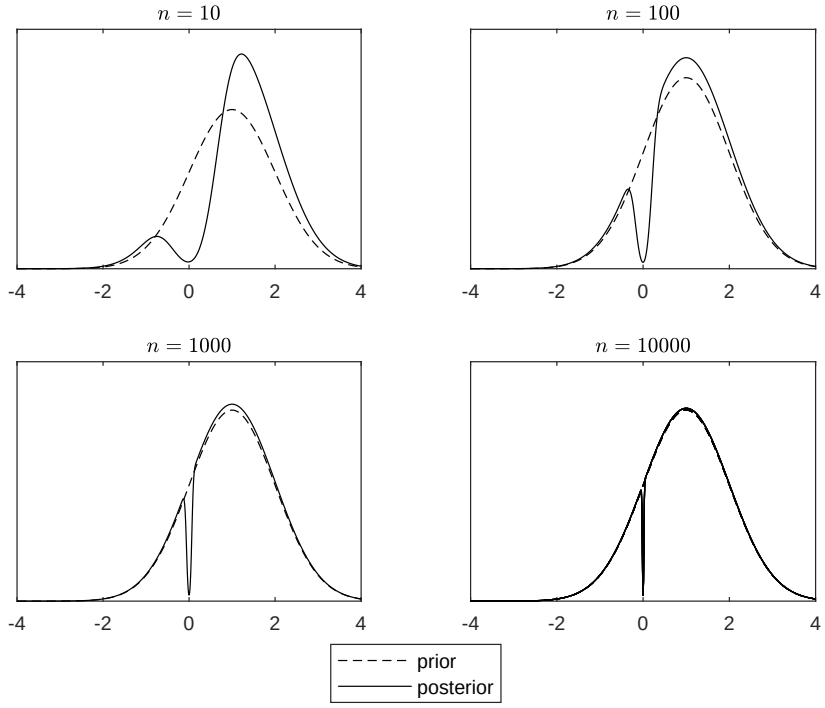

Figure 2: Prior and Posterior with Significance for Different Sample Sizes

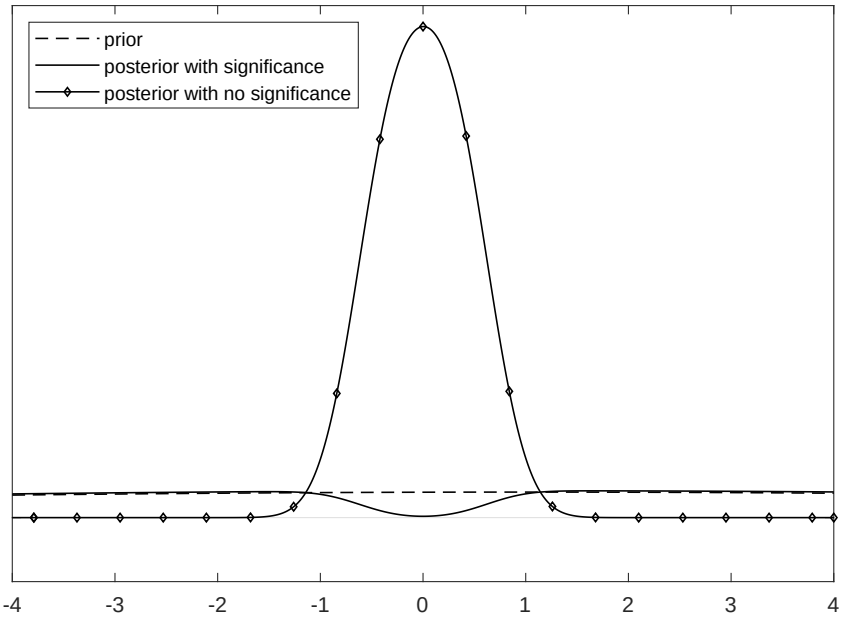

Figure 3: Posterior Distribution with Diffuse Prior After a Significance Test

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

It is worth noticing that the results in Figures 1 and 2 do not derive from

the adoption of an informative prior. Figure 3 reports prior and posterior distributions with $n = 10$ and $\mu = 1$, as in Figure 1, but now with $\sigma = 10$. This choice corresponds to a diffuse prior, i.e., "objective Bayes" analysis. The use of a diffuse prior increases the informativeness of non-significance relative to significance. The adoption of a diffuse prior increases the probability of rejection, $\Pr(\sqrt{n}|\widehat{\theta}| > c)$. As we show in Section 3 below, large values for $\Pr(\sqrt{n}|\widehat{\theta}| > c)$ result in high informativeness of non-significance relative to significance.

Equations (2) and (3) 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

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

and

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

So, in this very particular context, knowledge of the t-ratio ($\sqrt{n}\widehat{\theta}$) is sufficient to go back to the full information posterior. The same is true for the combined information given by the P-value, $2\Phi(-\sqrt{n}|\widehat{\theta}|)$, and the sign of $\widehat{\theta}$. This 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.

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.

3. General Case

3.1. 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 we obtain

$$\left|1 - \frac{p(\theta|\widehat{T}_n \leq c)}{p(\theta)}\right| = \left(\frac{\Pr(\widehat{T}_n > c)}{\Pr(\widehat{T}_n \leq c)}\right) \left|1 - \frac{p(\theta|\widehat{T}_n > c)}{p(\theta)}\right| \tag{4}$$

for $\Pr(\widehat{T}_n \leq c) > 0$ and θ such that $p(\theta) > 0$. The absolute value expressions on both sides of Equation (4) 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.

Equation (4) implies that the local informativeness of non-significance relative to significance at θ is solely determined the ratio $\Pr(\widehat{T}_n > c)/\Pr(\widehat{T}_n \leq c)$, which (remarkably) does not depend on θ. For example, for $\Pr(\widehat{T}_n > c) = 1/2$, which typically indicates a rather underpowered setting or a large prior probability mass at the point null, Equation (4) implies that non-significance is exactly as informative as significance. Moreover, the relative informativeness of non-significance increases with the statistical power of the test. Next section provides large sample calculations.

3.2. Large Sample Analysis

To extend the large sample results of the previous section beyond normal priors and data, we will consider a test statistic, $\widehat{T}_n$, such that

$$\Pr\left(\widehat{T}_n > c \middle| \theta, \theta \neq 0\right) \to 1,$$

and

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

That is, we consider significance tests that are consistent under fixed alternatives and have asymptotic size equal to α.

3.2.1. 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,

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

and

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

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

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

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 $p(\theta|\widehat{T}_n \leq c)$ is complicated by the fact that both $\Pr(\widehat{T}_n \leq c|\theta)$ and $\Pr(\widehat{T}_n \leq c)$ converge to zero. We will assume that the testing procedure is such that $\Pr(\widehat{T}_n \leq c|\theta)$ decays to zero at an exponential rate as a function of n. This is a weak requirement, which typically follows from large deviations arguments. For the next calculation, it is convenient to adopt a short-hand notation for the probability of Type II error,

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

Suppose that

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

This is also a weak assumption as it merely rules out perfect local asymptotic power. Then, by change of variable $z = n^{1/2}\theta$ and Fatou's lemma, we obtain[2]

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

[2]For the second to last equality, notice that if $a_n \geq 0$ and $b_n \to b > 0$ as $n \to \infty$, then $\liminf_{n\to\infty}(a_n b_n) = \liminf_{n\to\infty} a_n \lim_{n\to\infty} b_n$.

It follows that $\Pr(\widehat{T}_n \leq c)$ converges to zero at a polynomial rate. 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 2, 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.

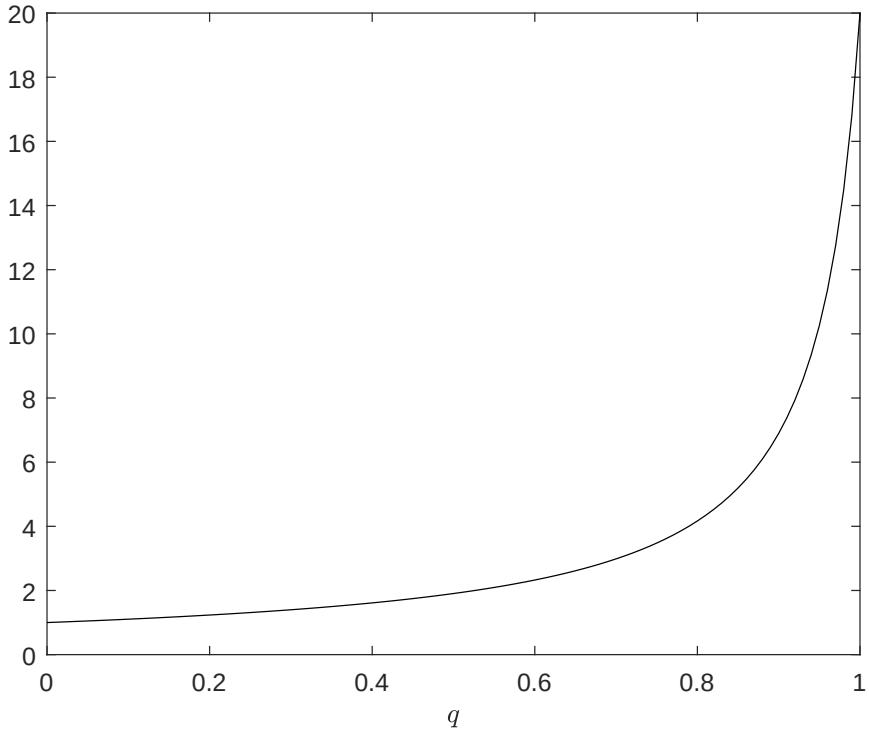

Figure 4: Limit of $p(\theta|\widehat{T}_n > c)/p(\theta)$ as a function of q ($\theta \neq 0$, $\alpha = 0.05$)

3.3. 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. Figure 4 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. For example, with $\alpha = 0.005$ (as advocated in Benjamin et al., 2017), for significance to at least double the probability of $\theta \neq 0$ we need $q \geq 1/(2(1-\alpha)) = 0.5025$. In fact, regardless of the size of the test, q needs to be bigger 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 on the causal effect of a policy intervention may sometimes concentrate probability 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.[3,4] 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

[3]See McShane et al. (2017) for a related discussion.

[4]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.

null hypothesis when $\widehat{T}_n > c$. Bayesians, however, would typically find that the posterior probability of the point null far exceeds the posterior probability of the alternative. Lindley's paradox can be explained by the fact that, as n increases, the distribution of the test statistic under the alternative diverges. 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.

4. Testing an Interval Null

In view of the lack of informativeness of non-significance in large samples (under a point null), one could instead try to reinterpret significance tests as tests of the implicit null "θ is close to zero".

To accommodate this possibility, we will now concentrate in the problem of testing the null that the parameter θ is in some interval around zero. Under the null hypothesis, $\theta \in [-\delta, \delta]$, where δ is some positive number. Under the alternative hypothesis, $\theta \notin [-\delta, \delta]$. Consider the normal model of section 2. To obtain a test of size α we control the supremum of the probability of Type I error:

$$\Pr(\sqrt{n}|\widehat{\theta}| > c \,|\, |\theta| = \delta) = \Phi(\sqrt{n}\delta - c) + \Phi(-\sqrt{n}\delta - c).$$

Therefore, we choose c such that $\Phi(\sqrt{n}\delta - c) + \Phi(-\sqrt{n}\delta - c) = \alpha$. While there is no closed-form solution for c, its value can be calculated numerically for any given value of $\sqrt{n}\delta$, and a very accurate approximation for large $\sqrt{n}\delta$ is given

by

$$c = \Phi^{-1}(1-\alpha) + \sqrt{n}\delta.$$

That is, controlling size in this setting implies that the critical value has to increase with the sample size at a root-n rate, with the constant given by δ. In turn, this implies that the probability of rejection, $\Pr(\sqrt{n}|\widehat{\theta}| > c|\theta) = \Phi(\sqrt{n}\theta - c) + \Phi(-\sqrt{n}\theta - c)$ converges to one if $\theta \notin [-\delta, \delta]$, and converges to zero if $\theta \in (-\delta, \delta)$. As a result, the large sample posterior distributions with and without significance are truncated versions of the prior, with the prior truncated at $(-\delta, \delta)$ under significance, and at $(-\infty, -\delta) \cup (\delta, \infty)$ under no significance. If δ is large, both significance and non-significance are informative. If, however, δ is small, we go back to the setting where significance carries only local-to-zero information. Figure 5 reports exact prior and posterior distributions for the same prior as in Figure 1, and with $\delta = \{0.5, 1, 2\}$, $\alpha = 0.05$ and $n = 10000$.

5. Conditioning on the Sign of the Estimated Coefficient

In previous sections we have shown that statistical significance may carry very little information in large samples. As a result, the values of other statistics should be taken into account along with significance when the null is rejected in a significance test. As discussed above, in a normal (or asymptotically normal) setting it does not take much to go back to full information (e.g., P-value and the sign of $\widehat{\theta}$). Here we consider the question of whether minimally augmenting the information on significance with the sign of $\widehat{\theta}$ results in informativeness when the null is rejected. This exercise is motivated by the possibility that the sign of the estimated coefficient is implicitly taken into account in many discussions of results from significance tests.

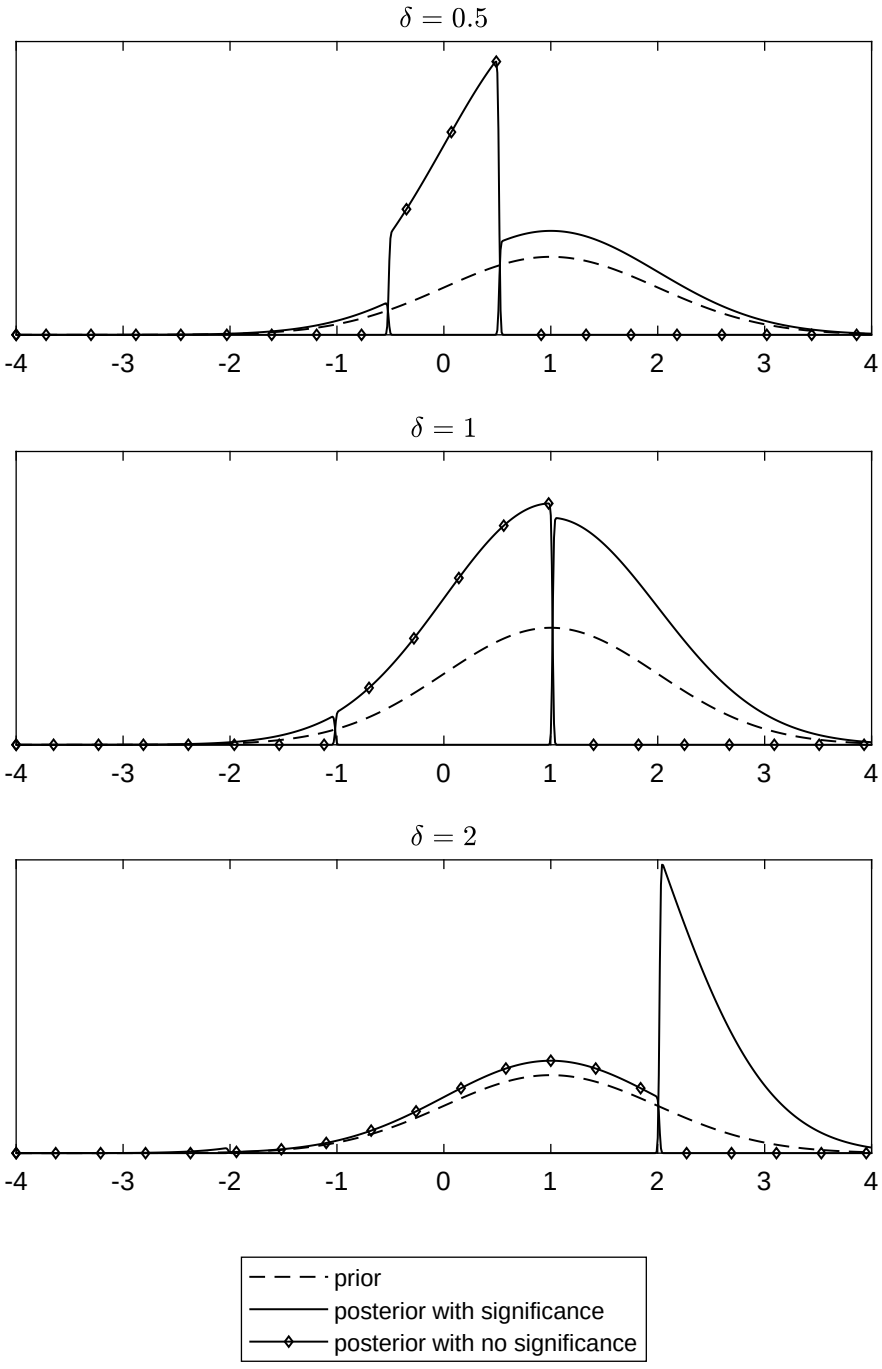

Figure 5: Posterior After a Test of the Null $\theta \in [-\delta, \delta]$ ($n = 10000$, $\alpha = 0.05$)

For concreteness, we will concentrate on the case of a positive coefficient estimate, $\widehat{\theta} > 0$. That is, the limited information posterior under significance and positive $\widehat{\theta}$ conditions on the event $\sqrt{n}\widehat{\theta} > c$. The case with negative $\widehat{\theta}$ is analogous. Using similar calculations as in section 1, we obtain:

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

and

$$p\big(\theta\big|0 < \sqrt{n}\widehat{\theta} \leq c\big) = \frac{\frac{1}{\sigma}\phi\left(\frac{\theta-\mu}{\sigma}\right)\Big(1 - \Phi(\sqrt{n}\theta - c) - \Phi(-\sqrt{n}\theta)\Big)}{1 - \Phi\left(\frac{\sqrt{n}\mu - c}{\sqrt{1+n\sigma^2}}\right) - \Phi\left(\frac{-\sqrt{n}\mu}{\sqrt{1+n\sigma^2}}\right)}.$$

Figure 6 reproduces the setting of Figure 1 but for the case when the posterior is conditional on sign of the estimate in addition to significance. Like in Figure 1, failure to reject carries subtantial information. In fact, both outcomes of the significance test carry additional information, with respect to the setting in Figure 1, which of course is explained by the additional information in the sign of $\widehat{\theta}$.

Notice that, in this case, under significance, the ratio between the posterior and the prior converges to

$$\lim_{n\to\infty} \frac{p(\theta|\sqrt{n}\widehat{\theta} > c)}{p(\theta)} = \begin{cases} 0 & \text{if } \theta < 0, \\ \Phi(-c)/\Phi(\mu/\sigma) & \text{if } \theta = 0, \\ 1/\Phi(\mu/\sigma) & \text{if } \theta > 0. \end{cases}$$

Without significance, the ratio between the posterior and the prior converges to

$$\lim_{n\to\infty} \frac{p(\theta|0 < \sqrt{n}\widehat{\theta} \leq c)}{p(\theta)} = \begin{cases} 0 & \text{if } \theta \neq 0, \\ \infty & \text{if } \theta = 0. \end{cases}$$

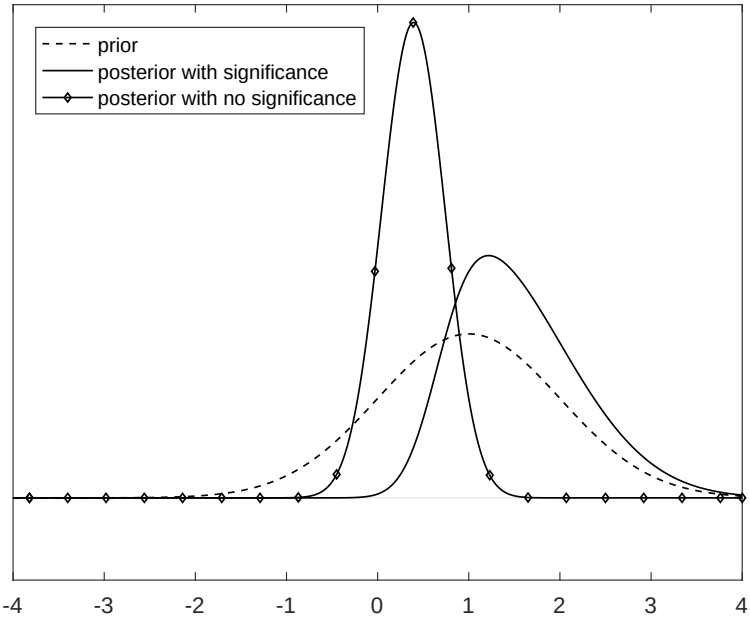

Figure 6: Posterior Distributions Conditional of Significance and Positive Coefficient Sign

That is, as $n \to \infty$ non-significance is highly informative. Under significance, the posterior of θ converges to the prior truncated at zero. As a result, in this case the informational content of significance depends on the value of $\Pr(\theta > 0) = \Phi(\mu/\sigma)$. If this quantity is small, significance with a positive sign is highly informative. Unsurprisingly, when μ/σ is large (that is, in cases where there is little uncertainty about the sign of the parameter of interest), a positive sign of $\widehat{\theta}$ does not add much to the informational content of the test. Moreover, the limit of $p(\theta|\sqrt{n}\widehat{\theta} > c)$ cannot be more than double the value of $p(\theta)$ as long as μ is non-negative. This is relevant to many instances in economics where there are strong beliefs about the sign of the estimated coefficients (e.g., the slope of the demand function, or the effect of schooling on wages) and specifications reporting "wrong" signs for the coefficients of

interest are rarely reported or published.[5]

6. Calibration using data from economics laboratory experiments

In this section we use data from economics laboratory experiments (Camerer et al., 2016; Andrews and Kasy, 2017) to calibrate parameters of the prior and the number of available observations in the posterior density formulas of section 2. The goal is to approximate the posterior densities with and without significance in 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. One of the reasons to adopt this particular setting for our calibration study is that, as explained below, in the context of this data set Andrews and Kasy (2017) 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 (2017) 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 statistics values for replications of those 18 studies. In particular, we use of the z-statistics computed in Andrews and Kasy (2017) for the published and the replication studies.[6]

[5]The notion that empirical estimates might display "wrong" signs is widespread to the point that econometric articles and textbooks discuss this phenomenon and, in some cases, potential remedies. See, e.g., Wooldridge (2016) and Kennedy (2005).

[6]`https://scholar.harvard.edu/files/kasy/files/publicationbiassupplement.pdf`.

We consider θ equal to the probability limit of the rescaled z-statistic, $2\widehat{z}/\sqrt{n}$, and calibrate a prior for θ using the distribution of the rescaled replication statistics, $2\,\widehat{z}_j^*/\sqrt{n_j^*}$, $j = 1, \ldots, 18$. In the previous expression, $\widehat{z}_j^*$ is the replication value of a z-statistic for the point null evaluated in study j, and n_j^* is the size of the replication sample. We make this particular choice because, for the simple case when $\widehat{z}$ is the usual two-sample z-statistic with equal number of observations on the treatment and control arms, θ becomes the 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.[7] We calibrate the parameters μ and σ^2 in section 2 to be the mean and variance of $2\,\widehat{z}_j^*/\sqrt{n_j^*}$, $j = 1, \ldots 18$ ($\mu = 0.3407$ and $\sigma = 0.2975$) and we calibrate the number of observations to be the median number of observations in the original studies ($n = 120$).[8,9] Although the distribution of the replication statistics may not correspond to a widespread prior on θ, the fact that these values are not affected by publication bias (conditional on publication of the original studies) makes them a reasonable choice to calibrate a prior.

Figure 7 shows the calibrated prior and posteriors with and without significance for the experimental economics data set. In this realistic scenario,

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

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

[9] Because, in the setting of this section, the distribution of the published z-statistics, $\widehat{z}_j$, is approximately normal with mean $(\sqrt{n}/2)\theta$ and variance one, the limited information posterior formulas of section 1 apply with n replaced by $n/4$.

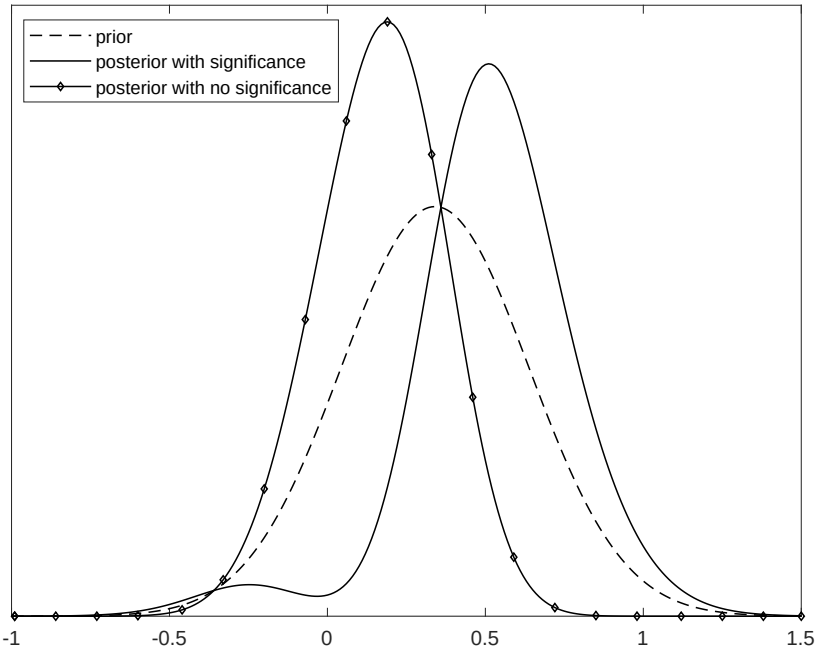

Figure 7: Prior and Posterior Densities Calibrated to Experimental Economics Data

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 (at the 5 percent level). Using the same dataset of published and replicated results in experimental economics, Andrews and Kasy (2017) estimate that the probability of publication with significance is 30 times higher than without significance. Figure 8 reports the published and replication z-statistics for the 18 economics laboratory experiments in Camerer et al. (2016) and Andrews and Kasy (2017). The distribution of published z-statistics shows a large concentration immediately to the right of 1.96. This feature is absent in the distribution of the replication z-statistics.

Finally, notice that the empirical context adopted in this section is one of

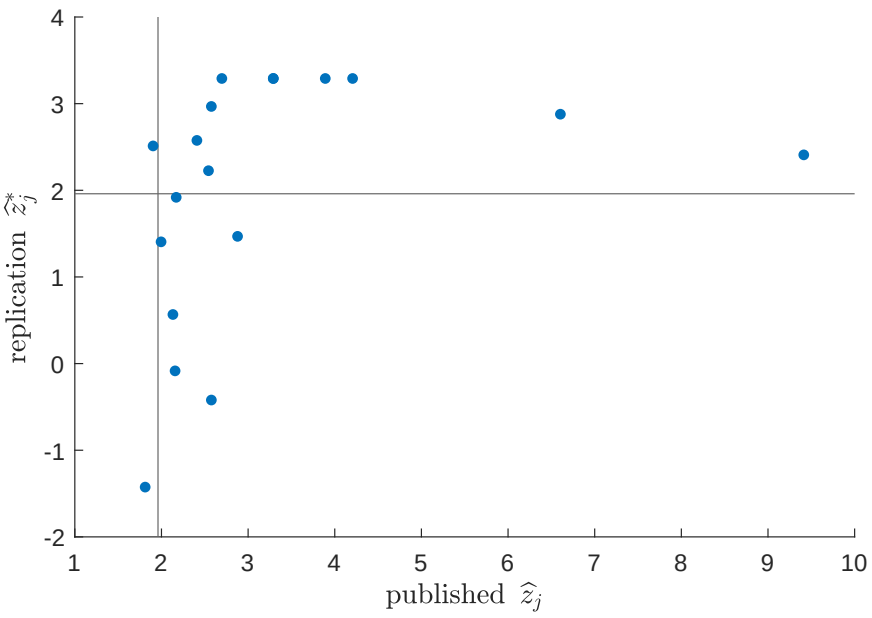

Figure 8: Publication and Replication z-Statistics in the Experimental Economics Data

unusually low statistical power. For the calibrated values of μ, σ, and n, we obtain a value 0.5031 for probability of rejection integrated over the prior. The informativeness of non-significance relative to significance will be even larger in empirical settings with higher statistical power.

7. Conclusions

Significance testing on a point null is the most extended 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 typical 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 a visible reporting and discussion of non-significant results in empirical practice (e.g., as in Angrist et al., 2017; Cantoni, 2018; Krueger and Malečková, 2003). More generally, as discussed in Ziliak and McCloskey (2008), McShane et al. (2017) 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 of the estimates, the plausibility and novelty of the results, and the quality of the research design, should be carefully evaluated alongside discussions of statistical significance or of the magnitude of p-values.

References

Abadie, A. and G. W. Imbens (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics 29*(1), 1–11.

Andrews, I. and M. Kasy (2017). Identification of and correction for publication bias. Working paper.

Angrist, J. D., V. Lavy, J. Leder-Luis, and A. Shany (2017). Maimonides rule redux. NBER Working Paper 23486.

Benjamin, D. J., J. O. Berger, M. Johannesson, B. A. Nosek, E.-J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, et al. (2017). Redefine statistical significance. *Nature Human Behavior*, 6–10.

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

Camerer, C. F., A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, E. Heikensten, F. Holzmeister, T. Imai, S. Isaksson, G. Nave, T. Pfeiffer, M. Razen, and H. Wu (2016). Evaluating replicability of laboratory experiments in economics. *Science*.

Cantoni, E. (2018). Got ID? The zero effects of voter ID laws on county-level turnout, vote shares, and uncounted ballots, 1992-2014. Working paper.

Efron, B. and T. Hastie (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. New York, NY, USA: Cambridge University Press.

Fisher, R. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.

Gelman, A. (2015). What hypothesis testing is all about. (Hint: Its 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, A. and H. Stern (2006). The difference between significant and not significant is not itself statistically significant. *The American Statistician 60*(4), 328–331.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine 2*(8), 696–701.

Kennedy, P. E. (2005). Oh no! I got the wrong sign! What should I do? *The Journal of Economic Education 36*(1), 77–92.

Krueger, A. B. and J. Malečková (2003). Education, poverty and terrorism: Is there a causal connection? *Journal of Economic Perspectives 17*(4), 119–144.

Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data*. New York: Wiley.

McShane, B. B., D. Gal, A. Gelman, C. Robert, and J. L. Tackett (2017). Abandon Statistical Significance. arXiv, 1709.07588.

Sims, C. A. and H. Uhlig (1991). Understanding unit rooters: A helicopter tour. *Econometrica 59*(6), 1591–1599.

Wooldridge, J. M. (2016). *Introductory econometrics : A modern approach.* Boston: Cengage Learning.

Ziliak, S. and D. McCloskey (2008). *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Economics, Cognition, And Society. University of Michigan Press.