Discuss this article at Journaltalk:
https://journaltalk.net/articles/6096/

EJW
ECON JOURNAL WATCH
Scholarly Comments on Academic Economics

ECON JOURNAL WATCH 21(1)
March 2024: 92–112

What Is the False Discovery Rate in Empirical Research?

Tom Engsted[1]

LINK TO ABSTRACT

The false discovery rate in empirical research within economics and other social sciences is much higher than most researchers think. Apart from the well-documented problem of extensive specification searches (actual or potential 'data mining,' or what Gelman and Loken 2014 call the "garden of forking paths") that naturally characterizes analysis of non-experimental, multi-dimensional, and extremely complex social science data, a less recognized reason comes from within the statistical framework itself. The classical, or frequentist, hypothesis testing framework that is a core element in most empirical analyses has some inherent conceptual and interpretational problems that lead to widespread misunderstandings and misuse. These problems are so prevalent that the American Statistical Association in 2016 issued an official statement on the proper use and interpretation of statistical tests (Wasserstein and Lazar 2016).

The subtleties of classical hypothesis testing are so deep that even professional statisticians often have difficulties understanding them correctly (McShane and Gal 2017). A root cause of the problem is the way of conditioning in such tests. All classical tests are based on distributions of test statistics under the null hypothesis, H_0. For example, the p-value is the probability of observing a test value at least as high as in the sample, given that H_0 is true. If the p-value is below some threshold value (typically 5 percent), H_0 is rejected in favor of an alternative hypothesis, H_1. The underlying rationale is that if the probability of observing the actual data (or more extreme data) is low under the null, then the null is probably false and, instead, the alternative hypothesis is probably true. As the

1. University of Aarhus, 8000 Aarhus, Denmark. I thank Kevin Lang and two anonymous referees for useful comments.

British statistician Ronald Fisher, who popularized the *p*-value, argued when seeing a low *p*-value: "either an exceptionally rare chance has occurred, or the theory [stated in H_0] is not true" (Fisher 1959, 39).

However, if one is not careful such reasoning quickly leads to the 'fallacy of the transposed conditional,' by which the conditional probability of an event A, given another event B, i.e., $P(A \mid B)$, is mistaken for the reverse conditional probability, $P(B \mid A)$. All researchers in their first probability and statistics course have learned not to make this mistake; nevertheless, this is exactly the mistake researchers make when they interpret the *p*-value, i.e., $P(D+ \mid H_0)$ where $D+$ stands for 'the data or more extreme data,' as giving a probabilistic assessment of the null hypothesis in light of the data, i.e., as $P(H_0 \mid D)$. It is generally not recognized that the probability that H_0 is true, given the data, can be very high (and even higher than the probability that H_1 is true) when the *p*-value is low.[2]

Similarly, the 'significance level,' α, in classical hypothesis testing is a conditional probability where the conditioning is on H_0. It gives the Type I error probability, i.e., the probability of rejecting a null hypothesis that is true: $\alpha = P(H_0$ is rejected $\mid H_0$ is true). If α is prefixed at 5 percent, then in repeated application in different samples at most 5 percent of true null hypotheses will be erroneously rejected. So it comes stumbling close to mistakenly interpret α as the fraction of false rejections in repeated application of the test, i.e., as $P(H_0$ is true $\mid H_0$ is rejected). However, this latter conditional probability is *not* controlled by the significance level. $P(H_0$ is true $\mid H_0$ is rejected) is denoted the 'false discovery rate' (FDR). The FDR depends on α, but it also depends on the power of the test, and on the unconditional probability of the null, $P(H_0)$; see Equation (1) below.

A scientific empirical finding is often based on rejecting a null of no effect or relationship (e.g., between two variables). If the academic literature within a given field has produced a lot of such effects or relationships, i.e., a lot of null rejections, a natural question is: How many of these rejections are a mistake? In other words: What is the false discovery rate? To assess FDR, we need to assess $P(H_0)$. Economists often implicitly (or explicitly, e.g., Abadie 2020) assume that $P(H_0)$ is low. I will argue, however, that for both statistical and economic reasons,

2. The reliance on tail area probabilities in classical tests is an additional source of confusion. As noted above, $D+$ in the definition of the *p*-value is 'the data or more extreme data.' Thus, the *p*-value depends on hypothetical data that could have been observed but was not observed. The implications of this particular feature of classical tests are far-reaching but generally not recognized by empirical scientists. For example, correct computation of the *p*-value in principle requires knowledge of the researcher's subjective experimental *intentions*; see Berger and Wolpert (1988) and Wagenmakers (2007) for details and illustrations of the paradoxes that this involves. See also Engsted and Schneider (2024) for a detailed discussion and survey of the foundational literature on *p*-values and hypothesis testing in the context of non-experimental data and empirical modeling within the social sciences.

this unconditional probability needs to be set at a high level, in general higher than 50 percent. As a consequence, since FDR depends positively on $P(H_0)$, then the FDR within economics is quite high and, in any case, substantially higher than the 5% Type I error rate implied by the 5% significance level chosen in most studies.

It is worth emphasizing that computing the FDR is not the same as correcting the significance level and *p*-value for multiple testing. The conventional 5% threshold for α applies for a *single* test in each sample. When more than one test is performed in a given sample, the significance level for each individual test needs to be adjusted downwards in order to control the overall α level. For example, if two independent tests are conducted, a significance level of 2.53 percent must be chosen for each test in order for the overall Type I error probability to be 5 percent ($1 - (1 - 0.0253)^2 = 0.05$). In most empirical papers multiple tests are carried out (e.g., several explanatory variables, often in various combinations, are tried), but it is seldom to see correction for multiple testing. Usually the 5% level is applied for each test, whereby the overall Type I error rate will be higher than 5 percent, and often substantially higher than 5 percent. Of course, this will only intensify the problem of a high FDR (as seen in Equation (1) below, an increase in α leads to an increase in FDR). But, as will become clear in the fourth section, even under the assumption that the true overall Type I error rate is 5 percent, the FDR will in general be substantially higher than 5 percent.

The issues discussed in this paper are well-known from the statistics literature and have been much discussed in other fields such as epidemiology and psychology. In economics, however, the issues are less well recognized and, hence, in my view deserve more discussion.

In the next section I formally define the FDR and show its dependence on the prior null probability, $P(H_0)$. In the third section I discuss how to assess $P(H_0)$, followed by, in the fourth section, an assessment of the FDR based on α, test power, and $P(H_0)$. In the fifth section I discuss whether *p*-value < 0.05 is a sufficiently strict hurdle rate for declaring evidence against H_0 for a single test in a given sample. The final section contains some concluding remarks on the special problems that non-experimental social science data pose for hypothesis testing.

The false discovery rate

As described above, in a classical hypothesis test the significance level, α, is the Type I error probability defined as $P(H_0 \text{ is rejected} \mid H_0 \text{ is true})$, i.e., the probability of rejecting the null hypothesis H_0, given that H_0 is true. If α is set at 5 percent, then over many repetitions of the test, 5 percent of true null hypotheses will be erroneously rejected. The reverse conditional probability, however, $P(H_0$ is

true $\mid H_0$ is rejected), is not controlled by the significance level. This conditional probability is what defines the false discovery rate, or FDR, and is given as (Storey 2003; see Appendix A for a formal derivation):[3,4]

$$\text{FDR} = P(H_0 \text{ is true} \mid H_0 \text{ is rejected}) = \frac{\alpha}{\alpha + \frac{(1-\beta)(1-P(H_0))}{P(H_0)}}. \quad (1)$$

The FDR depends on:

- α, the significance level
- $1-\beta$, the power of the test, where $\beta = P(H_0 \text{ is not rejected} \mid H_0 \text{ is not true})$ is the Type II error probability, and
- $P(H_0)$, the prior probability that H_0 is true.

In the fourth section I will assess the FDR for tests at a 5% significance level and various levels of power. These concepts (significance level and power) are well-known among empirical researchers applying hypothesis tests. The prior probability, $P(H_0)$, however, is not a concept that is well-known and appreciated by most empirical researchers. I will therefore, in the next section, discuss it in detail.

Prior probability of the null hypothesis

As seen in Equation (1), the FDR depends on $P(H_0)$, i.e., the unconditional (prior) probability of the null. How should we assess this prior probability? I will discuss this in the context of a simple regression model,

$$Y_i = a + bX_i + u_i, \quad i = 1, \ldots, n, \quad (2)$$

where Y_i and X_i are economic variables, u_i is the stochastic error term, and a and b are regression coefficients. The sample size is n, and subscript i refers to observation number i. The data can be measured either across time (time series data) or cross-sectionally (e.g., across individuals at a point in time). The researcher

3. Storey (2003) calls it the "positive false discovery rate" because it is conditional on at least one 'positive' finding, i.e., null rejection. Equation (1) holds exactly for independent tests and asymptotically for weakly dependent tests.

4. The Type I error rate is sometimes called the 'false positive rate,' which is not to be confused with the FDR. Unfortunately, the very informative and widely cited study by Benjamin et al. 2018, denotes the conditional probability in Equation (1) as the "false positive rate".

is interested in examining whether there is a relationship between X_i and Y_i, and therefore conducts a classical test of the point null hypothesis H_0: $b = 0$, against the composite alternative H_1: $b \neq 0$. Assume the researcher applies the conventional 5% significance level such that H_0 is rejected in favor of H_1 if the t-statistic > 2.0 (or, equivalently, p-value < 0.05).

Assume that H_0 is rejected such that a statistically significant relationship between X_i and Y_i has been established. What is now the probability that this is a false discovery? To compute this probability we need to assess the prior probability $P(H_0)$ which is most surely context dependent and ultimately a subjective (personal) matter. But what can be said about $P(H_0)$ more generally?

It seems to be a widely held view in academic economics that point nulls should not be attached strong prior probability weight. After all, economists typically develop models of relationships between economic variables, not models of no relationships. Alberto Abadie states that the common situation in economics is one where "there are rarely reasons to put substantial prior probability on a point null" (2020, 193). In a Bayesian setting where researchers report classical tests while journal readers have priors over the parameter of interest, Abadie (2020) shows that, provided the prior probability of the null is low, a statistical non-rejection is generally much more informative than a rejection at the 5% level. Informativeness is measured by the discrepancy between the prior and the posterior distribution. When the null value $b = 0$ is assigned a low prior probability, a statistical rejection does not substantially alter this prior belief. By contrast, a non-rejection generally leads to a substantial change in beliefs. Intuitively, if $P(H_0)$ is low, you expect H_0 to be rejected. So, a non-rejection is surprising and, hence, informative. Abadie (2020) thus argues that the usual practice of conferring point null rejections a higher level of scientific significance than non-rejections is unwarranted.[5]

Nowadays, statistically significant findings certainly get more attention than findings that are not statistically significant, cf. the discussions of 'p-hacking,' the 'file drawer problem,' and 'publication bias,' that occupy much of the current literature on the replication crisis in empirical research.[6] This is a natural consequence of the fact that an empirical finding is almost always based on rejection of a null hypothesis, as in Equation (1) where finding a relationship between X_i and Y_i (the working hypothesis) requires rejection of H_0: $b = 0$.

However, not long ago it was not uncommon to have one's working hypothesis (economic model) stated in the null and, so, *failure* to reject the null was

5. If the reader is uncomfortable with the (Bayesian) notion of a prior null probability, $P(H_0)$ may alternatively be thought of as simply the overall proportion of true nulls in the population of null hypotheses within a given field (see the next section).
6. See, e.g., Elliott et al. 2022 and the references therein.

taken as evidence in support of the model. Based on a sample of published articles in economics from the 1980s, Brad DeLong and Kevin Lang (1992) analyzed the fraction of *un*rejected null hypotheses that are true. In their sample, 78 of a total of 276 substantive economic hypotheses, formulated as a statistical null hypothesis, failed to be rejected. The 1980s saw a plethora of tests of, e.g., financial market efficiency (return unpredictability; overidentifying restrictions in rational expectations models), the unimportance of anticipated variables in monetary economics, and unit roots in macroeconomic variables (e.g., a unit root in real GDP was thought to have serious consequences for macroeconomic theory), all of which having the substantial model formulated as the null hypothesis. Here it is naturally interesting to analyze the fraction of false (or true) *non*-rejections. Using the property of the *p*-value that it is uniformly distributed under the null, DeLong and Lang (1992) derived an upper bound to the fraction of unrejected null hypotheses that are true, and they found that none of the unrejected nulls in their sample was true. They concluded that "π [the fraction of null hypotheses that are true] is essentially zero: that only a very small fraction of the null hypotheses in published articles are true. Failures to reject nulls are therefore almost always due to lack of power in the test, and not to the truth of the null hypothesis tested" (DeLong and Lang 1992, 1261).

Such analyses, where the null is the working hypothesis, have become less common. Nowadays, as discussed above, it is more common to have the economic model under consideration stated in the alternative hypothesis, so that rejection of the null is interpreted as support for the model. For example, in financial economics, instead of testing for market efficiency, researchers now develop models for new risk factors in the cross-section of asset returns and then test their empirical relevance by putting them in the statistical model stated in the alternative hypothesis, thus hoping for rejection of the null that the new factors are not relevant (I will elaborate on this example in the next section). In such analyses the (implicit) assumption is that the null has a low prior probability, cf. Abadie's (2020) statement as referenced above.

From a statistical point of view, however, this line of reasoning is problematic. Singling out a point null, H_0: $b = 0$, against a continuum of values in the composite alternative, H_1: $b \neq 0$, must imply that this particular point value should have a substantial prior probability assessment; if not, why single it out? Statisticians James Berger and Thomas Sellke (1987, 115) point out that for such a test it will rarely be justifiable to choose $P(H_0)$ less than 50 percent. Furthermore, if the point null does not have a substantial prior probability, what is then the rationale behind choosing a low significance level, like 5 percent? As it is often stated in statistics textbooks, the null constitutes the maintained hypothesis that we strongly believe in and, hence, there has to be strong evidence against the null in

the data before we are willing to reject it; this requires a low significance level, i.e., a low probability of rejecting a true null. As E. L. Lehmann and Joseph Romano put it in their classic text on testing statistical hypotheses: "If one firmly believes the hypothesis to be true, extremely convincing evidence will be required before one is willing to give up this belief, and the significance level will accordingly be set very low" (2008, 58).[7] The classical null hypothesis significance testing (NHST) paradigm is intended for settings where H_0 is the maintained hypothesis that is held to be true unless there is strong evidence against it in the data, i.e., $P(H_0)$ is taken to be high. It can be argued that in economics point nulls are rarely exactly true. However, H_0: $b = 0$ should not be interpreted as an exact zero effect, but rather as a 'negligible' effect. Berger and Mohan Delampady (1987) show that, unless the sample size is very large, a point null will often be a good approximation to a small interval null.

Thus, for statistical reasons there are solid arguments for putting a high prior probability weight on the null hypothesis. If we beforehand believe that $P(H_0)$ is low, testing a point null at conventionally low significance levels is almost by definition unsuited for the analysis. In the next section we shall see that there are also solid economic reasons for operating with a high prior null probability, although it goes against the (implicit) notion of many economists.

Assessing the false discovery rate empirically

Publication of comprehensive and progressively growing lists of statistically significant effects or relationships is a common phenomenon in many fields, and although several causes for an effect may be present, typically only a few causes contribute substantially. It is well-known that a multitude of risk factors for lung cancer have been identified, but cigarette smoking is *the* substantial factor accounting for 90 percent of lung cancer diagnoses (**link**). Similarly, Jonathan Sterne and George Davey Smith report that "by 1985 nearly 300 risk factors for coronary heart disease had been identified, and it is unlikely that more than a small fraction of these actually increase the risk of the disease" (2001, 227–228). For epidemiology they accordingly suggest to set the prior probability that the null hypothesis (no effect) is true to 90 percent. For experiments in psychology, Daniel Benjamin et al. (2018) state that the prior odds of H_1 relative to H_0 may be

7. In pure Neyman-Pearson hypothesis testing (where prior beliefs and the p -value are absent) the choice of a low significance level is due to Type I errors having more serious consequences than Type II errors. However, in the social sciences, meaningfully assigning 'costs' to these errors is very difficult, if not impossible, and is hardly ever done in practice.

about 1:10, corresponding to $P(H_0) \approx 90\%$, similar to Sterne and Smith's (2001) suggestion.

Economics is no different. As was already realized by Edward Leamer (1978; 1983), scientific findings in empirical economics are typically based on passively observed, non-experimental data and extensive specification searches with the purpose of finding statistically significant effects. A given economic phenomenon naturally has several causes, but often the bulk of variation in the 'dependent' variable is explained by only a few of these. As alluded to in the previous section, in financial economics, several hundreds of risk factors for the stock market have been identified—the so-called 'factor zoo'—and statistical significance at the 5% level seems to have been an important and necessary condition for publication of a new risk factor (Harvey et al. 2016; Harvey 2017; Harvey and Liu 2019).[8] According to the Capital Asset Pricing Model (CAPM), only one systematic risk factor explains variation in expected returns in the cross-section of assets, and although CAPM is known to be empirically inadequate, it is generally accepted among financial economists that only a few additional factors are needed to account for the major part of the cross-sectional variation in asset returns. Thus, the vast majority of the hundreds of financial risk factors found statistically significant in the academic literature are not substantially important, just as the vast majority of the hundreds of statistically significant risk factors for lung cancer or coronary heart disease are not substantially important.

For predicting asset returns, Campbell Harvey asks the question: "Among the many variables that researchers have explored, how many do we believe have 1:1 odds of being true return predictors before we look at the data?" and he answers "Very few" (2017, 1419). For empirical analyses in financial economics, Harvey (2017) generally suggests to (explicitly or implicitly) operate with a prior null probability, $P(H_0)$, substantially above 50 percent. That seems to be a sober recommendation for empirical analyses in economics more generally.

Table 1 shows the false discovery rate (FDR), computed as in Equation (1), for testing at the 5% level and for various levels of statistical power and prior null probabilities. With a neutral a priori assessment of the hypotheses, i.e., $P(H_0) = P(H_1) = 0.50$, a test with maximum power (100 percent) gives an FDR = 4.8%, close to the Type I error rate of 5 percent. But for a low-powered test FDR is somewhat higher than 5 percent. Empirical researchers often pay scant attention to the power properties of their tests and tend to ignore (or are not aware) that many empirical studies are under-powered. In a recent meta-study, John Ioannidis et al. (2017) find that the typical power in empirical economic research is just 18 percent.

8. Statistical significance at the 5% level has been and continues to be the standard hurdle rate for publication in economics more generally (Andrews and Kasy 2019).

Table 1 shows that with neutral prior odds and such low power, 21.7 percent of all significant findings are false.

TABLE 1. False discovery rate (FDR) computed using Equation (1) with $a = 0.05$

	Power $(1 - \beta)$			
	0.18	0.50	0.75	1.00
$P(H_0)$ = 0.10	0.030	0.011	0.007	0.006
= 0.25	0.085	0.032	0.022	0.016
= 0.50	0.217	0.091	0.062	0.048
= 0.75	0.455	0.231	0.167	0.130
= 0.90	0.714	0.474	0.375	0.310

The FDR naturally increases with increasing $P(H_0)$. As suggested above, $P(H_0) = 90\%$ is not an unreasonable choice for many areas within epidemiology and the social sciences. If, in addition, the typical test has power as low as 18 percent (as shown by Ioannidis et al. 2017), Table 1 shows that the FDR is 71.4 percent. No wonder that there is a replication crisis in empirical research! The intuition behind the 71.4% false discovery rate is straightforward: Assume that 10,000 tests are conducted. In 9,000 of these tests H_0 is true ($P(H_0) = 0.90$). With power equal to 18 percent, $0.18 \cdot 1{,}000 = 180$ of the false nulls are correctly rejected. With a 5% significance level, $0.05 \cdot 9{,}000 = 450$ of the true nulls are incorrectly rejected, and so among the total of $180 + 450 = 630$ rejections, $450/630 = 71.4$ percent are false discoveries.[9]

In some fields within economics, data samples are very large with thousands or even millions of observations (e.g., high-frequency financial data; register-based micro-data). In such analyses lack of power is not a problem. Instead, the problem in very large samples is that tiny and economically unimportant effects become statistically significant at conventional significance levels.[10] Statistics and econometrics textbooks sometimes suggest to lower α when the sample becomes (very) large, but standard hypothesis testing theory generally contains no formal procedures for how to optimally relate α to n, and reliance on conventional significance thresholds like 5 percent continues to dominate empirical research (as shown by, e.g., Harvey 2017; Andrews and Kasy 2019; Brodeur et al. 2020). There is no indication that empirical researchers generally lower the significance level when

9. Equation (1) also shows the danger of the intuitive notion that rejection of the null with a low-powered test is a strong result. If the null has a high prior probability assessment, such a rejection has a high probability of being wrong.

10. A t-statistic, for example, for H_0: $b = 0$, is defined as the estimate of b divided by its standard error, where the latter automatically shrinks with the sample size, n. Thus, for an arbitrarily small deviation from $b = 0$, the t-statistic automatically increases with n.

the sample size becomes big. Note also that, as seen in Table 1, even with maximum power, FDR is somewhat higher than the Type I error rate when the prior null probability is higher than 50 percent.

Thus, while low-powered studies naturally lead to a high FDR, in studies with large sample sizes and high power, tiny and economically insignificant findings that, as a consequence, have a high risk of being unreplicable in new samples, will often be statistically significant at conventional significance levels. Both cases thus contribute to the replication crisis.

A few decades ago, as noted in the previous section, it was not uncommon to have the proposed economic model stated in the null hypothesis, and not rejecting the null was evidence in support for the model, i.e., the 'discovery' was associated with non-rejection. In this case, specifying a low prior null probability would often be more reasonable. For example, the non-linear cross-equation parameter restrictions of highly stylized rational expectations models stated as the null hypothesis (see, e.g., Hansen and Sargent 1980) could reasonably be viewed as not likely to be true a priori (probably the implicit prior view of DeLong and Lang 1992). If we condition on H_0 *not* being rejected and compute the probability that the null is true, we obtain (see Appendix B):

$$P(H_0 \text{ is true} \mid H_0 \text{ is } \textit{not} \text{ rejected}) = \frac{(1-\alpha)P(H_0)}{\beta + (1-\alpha-\beta)P(H_0)}. \tag{3}$$

Table 2 reports this conditional probability for the same α (5 percent), power levels, and prior null probabilities as in Table 1. Naturally, with maximum power (100 percent, i.e., $\beta = 0$), the null is certainly true if it is not rejected, independent of $P(H_0)$ and α.[11] However, if both power and $P(H_0)$ are low, the probability that H_0 is true if it is not rejected, is low. Table 2 is an alternative way of expressing DeLong and Lang's (1992) insight that when none or only very few non-rejected nulls in a given sample of published articles are true (as they found), $P(H_0)$ must be low and the non-rejection must be due to low power of the test. Note, however, as argued above, this situation (low $P(H_0)$) is rarely relevant when we look at the more contemporary cases where an empirical finding is associated with rejection of the null. In such cases, the FDR in Table 1 with high values of $P(H_0)$ is more relevant. And, as argued in the previous section, from a statistical point of view, meaningfully

11. With 100 percent power, all false nulls will be rejected. Thus, if H_0 is not rejected it must be true, since if H_0 was not true it would have been rejected. More generally, one minus the conditional probability in Equation (3) is $P(H_0$ is *not* true $\mid H_0$ is *not* rejected) which can be denoted the False Non-Discovery Rate, FNDR. When the power of tests is low, this rate can be quite high. Thus, power has a substantial effect on both FDR and FNDR.

testing a point null against a composite alternative with a low α level in any case requires a high $P(H_0)$.

TABLE 2. $P(H_0$ **is true |** H_0 **is** *not* **rejected) computed using Equation (3) with** $\alpha = 0.05$

	Power $(1-\beta)$			
	0.18	0.50	0.75	1.00
$P(H_0) = 0.10$	0.114	0.174	0.297	1.000
$= 0.25$	0.279	0.388	0.559	1.000
$= 0.50$	0.537	0.655	0.792	1.000
$= 0.75$	0.777	0.851	0.919	1.000
$= 0.90$	0.912	0.945	0.972	1.000

I have discussed $P(H_0)$ in the language of a (Bayesian) prior probability. I believe this is the most natural way to consider it (see Engsted and Schneider 2024 for elaboration). However, $P(H_0)$ can alternatively be given a purely classical interpretation as the overall proportion of true nulls in the population of null hypotheses within a given field, and there exist classical ways of estimating this proportion. Interestingly, such estimates often indicate that the proportion is high, thus lending support to the recommendation of setting a high value of $P(H_0)$. For example, Laurent Barras et al. (2010), in separating skill from luck in assessing mutual fund performance across more than 2,000 mutual funds, estimate the proportion of truly neutral funds in the population (i.e., funds that deliver zero risk-adjusted excess returns, so-called 'zero alpha' funds) to be 75 percent, corresponding to $P(H_0) = 0.75$ associated with the null of zero alpha.[12]

Is *p*-value < 0.05 a sufficiently strict hurdle rate?

As argued in the previous section, in those areas where a scientific finding is associated with rejection of a statistical null hypothesis, the fraction of false findings across the many rejections is probably quite high, and thus is a contributing factor behind the replication crisis. As a 'quick fix' in order to reduce this fraction, Benjamin et al. (2018) have suggested to redefine statistical significance by lowering the conventional 5% level to 0.5%, corresponding to raising the *t*-statistic threshold from 2.0 to 2.8.

Working with a particular project, i.e., a particular model and a particular

12. As in DeLong and Lang (1992), the estimate of $P(H_0)$ in Barras et al. (2010) is based on the property of the *p*-value that it is uniformly distributed in the interval (0, 1) under the null.

sample of data, many researchers will undoubtedly feel that a t-statistic threshold of 2.8 is too high. Researchers have been trained and accustomed to believing that $t = 2.0$ is a suitable hurdle to pass for having confidence in a discovery. It is generally not recognized that if—as I have argued above should be the case—substantial prior probability weight is put on H_0, the 5% significance level (p-value threshold of 0.05) is in fact a quite *low* hurdle rate.

To see this, let us look at the conditional probability $P(H_0 \mid D)$, i.e., the probability that the null is true, given the data, associated with a t-statistic of 2.0. Define BF as the relative data likelihood under H_0 and H_1, i.e., $BF = P(D \mid H_0)/P(D \mid H_1)$. It can be shown that (see Appendix C):

$$P(H_0 \mid D) = \frac{P(H_0)BF}{1 + [P(H_0)(BF - 1)]}. \qquad (4)$$

In classical statistics, BF is a likelihood ratio statistic given as the data likelihood evaluated at the parameter value under the null divided by the data likelihood evaluated at the maximum likelihood estimate of the parameter. In Bayesian statistics BF stands for the 'Bayes factor,' where the numerator is the same as in the classical likelihood ratio, whereas the denominator is a marginal likelihood computed by positing a density function for the parameter and then averaging (integrating) over the parameter space.

Let the parameter of interest be b (e.g., the regression slope parameter in Equation (2)). If we apply the BF that for a given t-statistic gives maximum evidence against the null, i.e, the lower bound for BF, the result is $BF = \exp(-\frac{1}{2}t^2)$, where t is the usual t-statistic associated with H_0: $b = 0$ (see Berger and Sellke 1987). This particular BF is in fact identical to the classical likelihood ratio since maximum evidence against H_0 is obtained for a parameter density that is completely concentrated at the maximum likelihood estimate.

For $t = 2.0$ we obtain $BF = \exp(-\frac{1}{2} \cdot (2.0)^2) = 0.135$, and with neutral prior odds for the hypotheses $P(H_0) = P(H_1) = 0.50$, we get, by inserting in Equation (4), $P(H_0 \mid D) = 0.50 \cdot 0.135/(1 + [0.50 \cdot (0.135 - 1)] = 0.119$. For $P(H_0) = 0.90$, that I argued in the fourth section would be a reasonable choice in many cases, $P(H_0 \mid D) = 0.549$, that is, although the null would be rejected at the 5% level with a classical test, H_0 is *more* likely true than H_1!

$P(H_0 \mid D)$, computed as in Equation (4), is denoted the "Bayesianized p-value" by Harvey (2017), because it is the Bayesian equivalent to the classical p-value. As I noted in the introduction, the classical p-value is often misinterpreted as giving a probabilistic assessment of the null hypothesis in light of the data. We often informally—and incorrectly—associate 'statistical significance' with a low

probability that the null is true. $P(H_0 \mid D)$ in Equation (4) is the correct statement of this probability and, as seen, it depends on the prior probability assessment of the hypothesis, just as we have seen in the previous sections that the false discovery rate depends on this prior probability.

The above calculation shows that with a neutral prior stance on the null and alternative hypotheses, the Bayesianized *p*-value is an order of magnitude higher than the classical *p*-value (11.9 percent versus 5 percent) and, mind you, for a choice of *BF* that gives absolutely most probabilistic evidence against the null. Of course, the two types of *p*-values are not directly comparable because they measure different things. However, since many empirical researchers tend to associate the classical *p*-value with a probabilistic assessment of H_0, such reseachers will undoubtedly be surprised to learn that what they traditionally consider quite strong evidence against H_0 based on 'objective' statistical criteria, in fact implies a probability assessment of H_0 of 11.9 percent or higher. If researchers, in addition, actually think more deeply about the choice of a 5% significance level and consider it consistent with a prior probability assessment of H_0 even higher than 50 percent (cf. the previous sections), the Bayesianized *p*-value $P(H_0 \mid D)$ associated with a classical *p*-value of 5 percent, will be so high that no one would even consider rejecting the null when seeing a *t*-statistic of 2.0.

Of course, in reality, reseachers may (implicitly) operate with a very low prior probability assessment of the null (as, e.g., Abadie 2020), in which case the 5% level may be regarded a sufficiently strict threshold value. If researchers (implicitly) take a rejection at the 5% level to imply that $P(H_0 \mid D) < 5\%$, then their (implicit) prior null assessment is $P(H_0) < 28\%$ and, thus, $P(H_1) > 72\%$.[13]

Richard Startz (2014) provides a concrete example of the implicit prior probabilities of the hypotheses implied by the choice of a 5% significance level in a classical test, and he concludes: "We usually think that our standards for significance are chosen precisely to point in the direction of the null unless we have strong evidence to the contrary. But as this example illustrates, our usual standards do not accomplish that goal. In other words, in this example the *p*-values we usually regard as providing strong evidence against the null and in favor of the alternative do not in fact provide such evidence unless the econometrician already leaned strongly toward the alternative" (Startz 2014, 141).

---

13. From Equation (4) we obtain $P(H_0) = \frac{P(H_0 \mid D)}{P(H_0 \mid D)(1 - BF) + BF} = \frac{0.05}{0.05(1 - 0.135) + 0.135} = 0.28$, for the *BF* that gives maximum evidence against H_0 ($BF = \exp\left(-\frac{1}{2} \cdot (2.0)^2\right) = 0.135$). Thus, 28 percent is the highest possible prior null probability consistent with associating rejection at the usual 5% level with $P(H_0 \mid D) < 0.05$.

As argued in the previous two sections, there are both statistical and economic arguments against a low prior probability assessment of the null. Thus, classical hypothesis testing involves a paradox: the choice of a low significance level like 5 percent is due to the null being our maintained hypothesis that requires strong evidence against it in the data to be rejected, but, in fact, that same significance level has an implicit prior probability assessment that favors the *alternative* hypothesis! To avoid the paradox, an even stricter significance threshold is needed. Explicit new thresholds are presented below.

The above calculation of $P(H_0 \mid D) = 0.119$ for a t-statistic of 2.0 is based on conditions that resemble as much as possible the classical ideal of 'objectivity' (neutral prior odds: $P(H_0) = P(H_1) = 0.50$) and a Bayes factor (BF) which is identical to the classical likelihood ratio. As seen, under these conditions the Bayesianized p-value is substantially higher than the classical p-value. But, in fact, Bayesian statisticians generally consider this particular BF too extreme and instead recommend Bayes factors that do not so strongly favor the alternative hypothesis.

One choice of BF that has gained general acceptance is based on the Bayesian Information Criterion (BIC) suggested originally by Gideon Schwarz (1978). BIC is widely used, also among classical statisticians and econometricians, as a model selection tool. BIC implies a distribution for the parameter under H_1 that can be approximated by a normal distribution centered around the maximum likelihood estimate and with a relatively large variance (see Raftery 1995; Kass and Raftery 1995). Thus, BIC is appealing if we have some, but not a very precise, idea of the range of variation for the parameter. With, again, t being the usual t-statistic and n the sample size, $\text{BIC} = \log(n) - t^2$, and the Bayes factor becomes $BF = \exp(\frac{1}{2}\text{BIC})$.

Table 3 reports BF based on BIC, and the associated Bayesianized p-value, $P(H_0 \mid D)$, for a t-statistic of 2.0, various sample sizes, and the same prior null probabilities as in Tables 1 and 2. For a relatively small sample of $n = 50$ observations, the Bayes factor becomes $BF = 0.957$ which means that the data are slightly less likely under the null than under the alternative hypothesis. Thus, the data leads to only a very small downward revision of the probability that the null is true, and the Bayesianized p-value is substantially higher than the classical p-value of 0.05.

TABLE 3. *BF* **and** $P(H_0 \mid D)$ **based on BIC, computed using Equation (4), for a** t**-statistic of 2.0**

	$n = 50$, $BF = 0.957$	$n = 100$, $BF = 1.353$	$n = 500$, $BF = 3.026$
	$P(H_0 \mid D)$	$P(H_0 \mid D)$	$P(H_0 \mid D)$
$P(H_0) = 0.10$	0.096	0.131	0.252
$= 0.25$	0.242	0.311	0.502
$= 0.50$	0.489	0.575	0.752
$= 0.75$	0.742	0.802	0.901
$= 0.90$	0.896	0.924	0.965

When the sample size increases, the divergence between Bayesianized and classical *p*-values also increases and $BF > 1$ implies that the data lead to an *upward* revision of the probability that the null is true, a dramatically different conclusion than the rejection of H_0 that most researchers will do when seeing a *t*-statistic of 2.0.[14]

If, as above, researchers (implicitly) take a rejection at the 5% level to imply that $P(H_0 \mid D) < 5\%$, we can reverse engineer the BIC based *BF* to derive the implicit prior null probability. For example, for $n = 100$, we obtain (using the equation in footnote 13) $P(H_0) = 0.037$ and $P(H_1) = 0.963$, which are clearly not neutral. Alternatively, with neutral prior odds, $P(H_0) = 0.50$, we can ask what the *t*-statistic threshold needs to be in order for $P(H_0 \mid D) < 0.05$. The result is $t = 3.24$, corresponding to a *p*-value threshold of 0.0012; for $n = 500$, the required *t*-statistic is $t = 3.48$ and the *p*-value threshold is 0.0003. As seen, these thresholds are even stricter than the new 0.005 *p*-value threshold suggested by Benjamin et al. (2018). From a Bayesian perspective, the conclusion is that the conventional 5% significance threshold is a very low hurdle rate to pass in order to declare a significant finding; it needs to be raised markedly. Alternatively, and perhaps even better, statistical testing with fixed thresholds should be abandoned altogether. I discuss this in the concluding remarks.

Concluding remarks

In this paper I have argued that, for both statistical and economic reasons, when economists apply statistical hypothesis tests in their empirical work, and the substantive economic model appears in the alternative hypothesis, they should generally—explicitly or implicitly—put substantial prior probabilistic weight on the null hypothesis, in contrast to what many economists do in practice. A follow-up recommendation is that the conventional significance or *p*-value threshold should be lowered from the current 5% level. A natural consequence of applying a stricter threshold is to reduce the high false discovery rate that currently haunts empirical research.

However, transitioning from a *p*-value threshold of 0.05 to, e.g., 0.005 (as suggested by Benjamin et al. 2018) will—all else equal—lead to fewer null rejections and thereby more real effects not being discovered (lower test power), i.e., increase the false non-discovery rate. To alleviate this problem Benjamin et al.

14. As seen in Table 3, for a given prior null probability, $P(H_0 \mid D)$ increases when *n* increases. In the statistics literature this is called the Jeffreys-Lindley paradox (Jeffreys 1961; Lindley 1957) that implies that for a fixed but arbitrarily high *t*-statistic, $P(H_0 \mid D)$ approaches one when *n* goes to infinity.

(2018) propose—for experimental studies—to generally increase the sample sizes so as to maintain the conventionally required 80% power.

In economics and most other social sciences, however, empirical studies are usually based on non-experimental, passively observed data where increasing the sample size to obtain sufficiently high power will often be impossible. In some fields with an abundance of data, e.g., high-frequency financial data or register-based micro-data, power is in any case close to 100% and setting a stricter significance or *p*-value threshold should be unproblematic.[15] But in analyses with small or moderately sized samples of non-experimental data, just raising the hurdle rate for significance will naturally lead to more missed discoveries.

The heretical question is whether the strong reliance on the null hypothesis significance testing (NHST) paradigm based on *t*-statistic or *p*-value thresholds, which continues to characterize empirical research, should finally be dropped. An increasing number of statisticians and scientists move in that direction, saying "abandon statistical significance" (McShane et al. 2019) or "retire statistical significance" (Amrhein et al. 2019), and suggesting "moving to a world beyond 'p < 0.05'" (Wasserstein et al. 2019). In addition to the bad empirical practice that strong adherence to the NHST paradigm leads to, there is a growing recognition of the many paradoxes and built-in contradictions involved in this paradigm. Fisher's 'significance testing' based on *p*-values and the Neyman-Pearson 'hypothesis testing' approach based on a cost-benefit analysis of Type I and II errors, are in fact incompatible (Hubbard and Bayarri 2003; Schneider 2015), a fact that is becoming more known but still not generally recognized among empirical researchers, and applied statistics and econometrics textbooks still often present statistical testing as a hybrid Fisher-Neyman-Pearson theory.

For economics and the social sciences in general, where passively observed non-experimental data are prevalent, and where empirical models are to be considered very crude approximations to reality, the NHST paradigm is particularly problematic. In the social sciences the data are often 'convenience samples,' and new samples from the population cannot be drawn. In fact, the underlying population is often not well-defined and researchers then need to rely on imaginary 'super-populations' in order to apply the traditional measures of statistical uncertainty (sampling error) based on the sampling distributions of statistics in (hypothetical) repeated sampling. There are a bunch of problems involved in applying the traditional statistical paradigm in such cases (Berk et al. 1995; Engsted

15. One could, for example, set a new threshold based on BIC and neutral prior odds for the hypotheses, and such that rejection occurs when $P(H_0 \mid D) < 0.05$, cf. the "Is p-value < 0.05 a sufficiently strict hurdle rate?" section above. With, e.g., a sample size of $n = 100,000$ observations, this leads to a *t*-statistic threshold of 4.17, corresponding to a *p*-value threshold of 0.00003.

and Schneider 2024), problems that are either ignored or not recognized by most empirical social science researchers. In essence, the problem with the traditional approach is that it focuses on sampling error and ignores model uncertainty, where the latter is what is important for most social science studies.

In empirical economics, alternative measures of fit not based on statistical significance have been developed since the 1980s (see Engsted 2002 for a survey of such measures). These measures focus on economic significance instead of statistical significance. For example, in financial asset pricing, measures of the economic magnitude of pricing errors of particular asset pricing models exist, but, unfortunately, in this literature many published papers continue to put more emphasis on whether pricing errors are statistically significant.

What is needed is a general recognition among empirical researchers that statistical significance at arbitrary significance levels should no longer serve as a de facto necessary condition for declaring a scientific result because it leads to an unfortunate preoccupation with passing certain statistical thresholds independent on the economic context and the nature of the data. Instead, focus should be on the *economic* magnitude and significance of estimated effects, the robustness of the results to changes in the data and variable definitions, functional form, estimation methods, etc., and on model uncertainty in general.

Appendix A.
Derivation of the FDR in Equation (1)

Using Bayes' formula we can write the probability that H_0 is true, given that H_0 is rejected, as:

$$P(H_0 \text{ is true} \mid H_0 \text{ is rejected}) = \frac{P(H_0 \text{ is rejected} \mid H_0 \text{ is true})P(H_0)}{P(H_0 \text{ is rejected})}$$

and similarly:

$$P(H_1 \text{ is true} \mid H_0 \text{ is rejected}) = \frac{P(H_0 \text{ is rejected} \mid H_1 \text{ is true})P(H_1)}{P(H_0 \text{ is rejected})}.$$

Dividing the first with the second expression gives

$$\frac{P(H_0 \text{ is true} \mid H_0 \text{ is rejected})}{P(H_1 \text{ is true} \mid H_0 \text{ is rejected})} = \frac{P(H_0 \text{ is rejected} \mid H_0 \text{ is true})}{P(H_0 \text{ is rejected} \mid H_1 \text{ is true})}\frac{P(H_0)}{P(H_1)} = \frac{\alpha}{1-\beta}\frac{P(H_0)}{P(H_1)},$$

where α and β are the Type I and II error probabilities, respectively. Then imposing that probabilities add up to one such that $P(H_1) = 1 - P(H_0)$ and $P(H_1 \text{ is true} \mid H_0 \text{ is rejected}) = 1 - P(H_0 \text{ is true} \mid H_0 \text{ is rejected})$, Equation (1) is obtained.

Appendix B. Derivation of Equation (3)

Using Bayes' formula we can write the probability that H_0 is true, given that H_0 is *not* rejected, as:

$$P(H_0 \text{ is true} \mid H_0 \text{ is } not \text{ rejected}) = \frac{P(H_0 \text{ is } not \text{ rejected} \mid H_0 \text{ is true})P(H_0)}{P(H_0 \text{ is } not \text{ rejected})}$$

and similarly:

$$P(H_1 \text{ is true} \mid H_0 \text{ is } not \text{ rejected}) = \frac{P(H_0 \text{ is } not \text{ rejected} \mid H_1 \text{ is true})P(H_1)}{P(H_0 \text{ is } not \text{ rejected})}.$$

Dividing the first with the second expression gives

$$\frac{P(H_0 \text{ is true} \mid H_0 \text{ is } not \text{ rejected})}{P(H_1 \text{ is true} \mid H_0 \text{ is } not \text{ rejected})} = \frac{P(H_0 \text{ is } not \text{ rejected} \mid H_0 \text{ is true})}{P(H_0 \text{ is } not \text{ rejected} \mid H_1 \text{ is true})}\frac{P(H_0)}{P(H_1)} = \frac{1-\alpha}{\beta}\frac{P(H_0)}{P(H_1)},$$

where α and β are the Type I and II error probabilities, respectively. Then imposing that probabilities add up to one such that $P(H_1) = 1 - P(H_0)$ and $P(H_1 \text{ is true} \mid H_0 \text{ is } not \text{ rejected}) = 1 - P(H_0 \text{ is true} \mid H_0 \text{ is } not \text{ rejected})$, Equation (3) is obtained.

Appendix C. Derivation of Equation (4)

Bayes' formula implies that

$$\frac{P(H_0 \mid D)}{P(H_1 \mid D)} = \frac{P(D \mid H_0)}{P(D \mid H_1)} \cdot \frac{P(H_0)}{P(H_1)} = BF \cdot \frac{P(H_0)}{P(H_1)}.$$

Imposing that probabilities add up to one such that $P(H_1) = 1 - P(H_0)$ and $P(H_1 \mid D) = 1 - P(H_0 \mid D)$, Equation (4) is obtained.

References

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

Amrhein, Valentin, Sander Greenland, Blake McShane, et al. 2019. Retire Statistical Significance. *Nature* 567: 305–307.

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

Barras, Laurent, Olivier Scaillet, and Russ Wermers. 2010. False Discoveries in Mutual Fund Performance: Measuring Luck in Estimated Alphas. *Journal of Finance* 65(1): 179–216.

Benjamin, Daniel J., James O. Berger, et al. 2018. Redefine Statistical Significance. *Nature Human Behaviour* 2: 6–10.

Berger, James O., and Mohan Delampady. 1987. Testing Precise Hypotheses [with discussion]. *Statistical Science* 2: 317–352.

Berger, James O., and Thomas Sellke. 1987. Testing a Point Null Hypothesis: The Irreconcilability of *P* Values and Evidence [with discussion]. *Journal of the American Statistical Association* 82: 112–139.

Berger, James O., and Robert L. Wolpert. 1988. *The Likelihood Principle*, 2nd ed. Hayward, Cal.: Institute of Mathematical Statistics.

Berk, Richard A., Bruce Western and Robert E. Weiss. 1995. Statistical Inference for Apparent Populations [with discussion]. *Sociological Methodology* 25: 421–485.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020. Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics. *American Economic Review* 110(11): 3634–3660.

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

Elliott, Graham, Nikolay Kudrin, and Kaspar Wüthrich. 2022. Detecting *P*-Hacking. *Econometrica* 90(2): 887–906.

Engsted, Tom. 2002. Measures of Fit for Rational Expectations Models. *Journal of Economic Surveys* 16(3): 301–355.

Engsted, Tom, and Jesper W. Schneider. 2024. Non-Experimental Data, Hypothesis Testing, and the Likelihood Principle: A Social Science Perspective. *Foundations and Trends in Econometrics* (Now Publishers Inc., Boston) 13(1): 1–66.

Fisher, Ronald A. 1959. *Statistical Methods and Scientific Inference*, 2nd ed. Edinburgh: Oliver & Boyd.

Gelman, Andrew, and Eric Loken. 2014. The Statistical Crisis in Science. *American Scientist* 102: 460–465. **Link**

Hansen, Lars Peter, and Thomas J. Sargent. 1980. Formulating and Estimating Dynamic Linear Rational Expectations Models. *Journal of Economic Dynamics and Control* 2: 7–46.

Harvey, Campbell R. 2017. Presidential Address: The Scientific Outlook in Financial Eco-

nomics. *Journal of Finance* 72(4): 1399–1440.

Harvey, Campbell R., and Yan Liu. 2019. A Census of the Factor Zoo. Working paper, March 18. **Link**

Harvey, Campbell R., Yan Liu, and Heqing Zhu. 2016. …And the Cross-Section of Expected Returns. *Review of Financial Studies* 29(1): 5–68.

Hubbard, Raymond, and M. J. Bayarri. 2003. Confusion over Measures of Evidence (p's) versus Errors (α's) in Classical Statistical Testing [with discussion]. *American Statistician* 57(3): 171–182.

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

Jeffreys, Harold. 1961. *Theory of Probability*, 3rd ed. London: Oxford University Press.

Kass, Robert E., and Adrian E. Raftery. 1995. Bayes Factors. *Journal of the American Statistical Association* 90(430): 773–795.

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

Leamer, Edward E. 1983. Let's Take the Con Out of Econometrics. *American Economic Review* 73: 31–43.

Lehmann, E. L., and Joseph P. Romano. 2008. *Testing Statistical Hypotheses*, 3rd ed. New York: Springer.

Lindley, D. V. 1957. A Statistical Paradox. *Biometrica* 44(1–2): 187–192.

McShane, Blakeley B., and David Gal. 2017. Statistical Significance and the Dichotomization of Evidence. *Journal of the American Statistical Association* 112: 885–908.

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

Raftery, Adrian E. 1995. Bayesian Model Selection in Social Research [with discussion]. *Sociological Methodology* 25: 111–163.

Schneider, Jesper W. 2015. Null Hypothesis Significance Tests: A Mix-Up of Two Different Theories: The Basis for Widespread Confusion and Numerous Misinterpretations. *Scientometrics* 102: 411–432.

Schwarz, Gideon. 1978. Estimating the Dimension of a Model. *Annals of Statistics* 6(2): 461–464.

Startz, Richard. 2014. Choosing the More Likely Hypothesis. *Foundations and Trends in Econometrics* (Now Publishers Inc., Boston) 7(2): 119–189.

Sterne, Jonathan A. C., and George Davey Smith. 2001. Sifting the Evidence—What's Wrong with Significance Tests? *British Medical Journal* 322: 226–231.

Storey, John D. 2003. The Positive False Discovery Rate: A Bayesian Interpretation and the Q-Value. *Annals of Statistics* 31(6): 2013–2035.

Wagenmakers, Eric-Jan. 2007. A Practical Solution to the Pervasive Problems of *P* Values. *Psychonomic Bulletin & Review* 14: 779–804.

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

Wasserstein, Ronald L., Allen L. Schirm, and Nicole A. Lazar. 2019. Moving to a World Beyond "$P < 0.05$." *American Statistician* 73(sup1): 1–19. **Link**

About the Author

Tom Engsted is professor of economics at University of Aarhus, Denmark. He has numerous publications within economics, financial economics, and econometrics, in journals such as *Journal of Money, Credit, and Banking*, *Review of Economics and Statistics*, *Journal of Financial and Quantitative Analysis*, and *Review of Financial Studies*. His email address is tengsted@econ.au.dk.

Go to archive of Economics in Practice section
Go to March 2024 issue