The Journal of Finance®

Campbell R. Harvey
The President of the American Finance Association 2016

THE JOURNAL OF FINANCE • VOL. LXXII, NO. 4 • AUGUST 2017

Presidential Address: The Scientific Outlook in Financial Economics

CAMPBELL R. HARVEY*

ABSTRACT

Given the competition for top journal space, there is an incentive to produce "significant" results. With the combination of unreported tests, lack of adjustment for multiple tests, and direct and indirect *p*-hacking, many of the results being published will fail to hold up in the future. In addition, there are basic issues with the interpretation of statistical significance. Increasing thresholds may be necessary, but still may not be sufficient: if the effect being studied is rare, even $t > 3$ will produce a large number of false positives. Here I explore the meaning and limitations of a *p*-value. I offer a simple alternative (the minimum Bayes factor). I present guidelines for a robust, transparent research culture in financial economics. Finally, I offer some thoughts on the importance of risk-taking (from the perspective of authors and editors) to advance our field.

SUMMARY

- Empirical research in financial economics relies too much on *p*-values, which are poorly understood in the first place.
- Journals want to publish papers with positive results and this incentivizes researchers to engage in data mining and "*p*-hacking."
- The outcome will likely be an embarrassing number of false positives—effects that will not be repeated in the future.
- The minimum Bayes factor (which is a function of the *p*-value) combined with prior odds provides a simple solution that can be reported alongside the usual *p*-value.
- The Bayesianized *p*-value answers the question: What is the probability that the null is true?
- The same technique can be used to answer: What threshold of *t*-statistic do I need so that there is only a 5% chance that the null is true?
- The threshold depends on the economic plausibility of the hypothesis.

LET ME START WITH AN ORIGINAL EMPIRICAL EXAMPLE. Some of you are familiar with my work on identifying factors. Consider a new factor that is based on equity returns from CRSP data. The *t*-statistic of the long-short portfolio is 3.23,

*Campbell R. Harvey is with Duke University and the National Bureau of Economic Research. I appreciate the comments of Manuel Adelino, Geert Bekaert, Jim Berger, Alon Brav, Anna Cieslak, Wayne Ferson, David Hirshleifer, Jay Im, Ben McCartney, Stefan Nagel, Brenda Priebe, Emmanuel Roman, Stephan Siegel, Michael Stutzer, Joy Tong, and Ivo Welch. I am especially grateful to Yan Liu for comments on many versions of this paper. Jay Im and Jane Day provided research assistance.

DOI: 10.1111/jofi.12530

which exceeds the Harvey, Liu, and Zhu (HLZ; 2016) recommended threshold of 3.00. The factor has a near-zero beta with the market and other well-known factors, yet an average return that comfortably exceeds the average market excess return. What is this factor?

Here are the instructions that I gave my research assistant: (1) form portfolios based on the first, second, and third letters of the ticker symbol; (2) show results for 1926 to present and 1963 to present; (3) use a monthly, not daily, frequency; (4) rebalance portfolios monthly and once a year; (5) value weight and equally weight portfolios; (6) make a choice on delisting returns; and (7) find me the best long-short portfolio based on the maximum t-statistic.

There are 3,160 possible long-short portfolios based on the first three letters of the tickers. With two sample periods, there are 6,320 possible portfolio choices, equal and value weights bring this number to 12,640, and two choices for reconstituting the portfolio doubles this number again. In short, there are a huge number of choices.

Many would argue that we should increase the choice space further because there are other possible choices that I did not give to my research assistant. Suppose, for instance, there are three ways to handle delisting returns. Ex ante, one was chosen. The argument is that we should consider the fact that, hypothetically, we could have had three choices, not just the one chosen (see Gelman and Loken (2013)).

It is not surprising that, under a large enough choice set, the long-short strategy has a “significant” t-statistic—indeed, dozens of strategies have “significant” t-statistics. This is an egregious example of what is known as p-hacking.

One might think this is a silly example. But it is not. A paper referenced in the HLZ (2016) factor list shows that a group of companies with meaningful ticker symbols, like Southwest’s LUV, outperform (Head, Smith, and Watson (2009)). Another study, this time in psychology, argues that tickers that are easy to pronounce, like BAL as opposed to BDL, outperform in IPOs (Alter and Oppenheimer (2006)). Yet another study, in marketing, suggests that tickers that are congruent with the company’s name outperform (Srinivasan and Umashankar (2014)). Indeed, some have quipped that ETF providers such as Vanguard might introduce a new family of ETFs called “AlphaBet” with each ETF investing in stocks with the same first letter of a ticker symbol.

Many interpret HLZ (2016) as suggesting that we “raise the threshold for discovering a new finding to $t > 3$.” However, the point of that paper is that many in our profession (including myself) have been making an error in not adjusting thresholds for multiple tests. *In this address, I emphasize that making a decision based on $t > 3$ is not sufficient either*. In particular, raising the threshold for significance may have the unintended consequence of increasing the amount of data mining and, in turn, publication bias. Journals contribute to data mining through their focus on publishing papers with the most “significant” results. The reason is that journal editors are often competing for citation-based impact numbers. Indeed, if you go to the American Finance Association’s homepage, you will see the *Journal of Finance*’s impact factor prominently displayed. Because papers that do not report “significant” results generate fewer citations

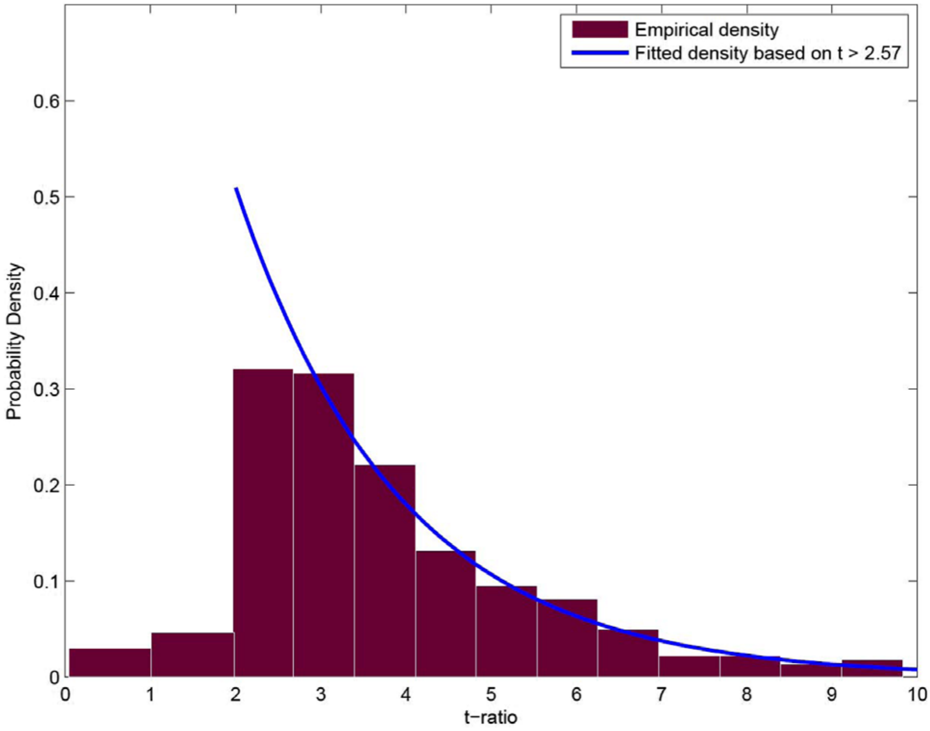


Figure 1. Distribution of reported *t*-statistics from factor studies, 1963 to 2012. Evidence from Harvey, Liu, and Zhu (2016).

(Fanelli (2013)), these papers are less likely to be published. This leads to publication bias, whereby readers see only a select sample of the actual research conducted.

Researchers also contribute to publication bias. Knowing journals' preference for papers with "significant" results, authors may not submit papers with "marginal" results. Such bias is known in other sciences as the "file drawer effect"—research is filed away rather than submitted to a journal (Rosenthal (1979)). Publication bias may also be induced by authors cherry-picking the most significant results (p-hacking) to submit to a journal. And even if journals were willing to publish papers with marginal or negative results, many authors may file away a research project because they do not want to spend their valuable research time on a paper they are not excited about.

Evidence of publication bias is starkly presented in HLZ (2016), who conduct a meta-analysis of factor studies published in top journals over the period 1963 to 2012. Based on their evidence, Figure 1 shows that the number of studies reporting t-statistics in the range of 2.0 to 2.57 is almost the same as the number reporting in the range of 2.57 to 3.14—which only makes sense if there is publication bias. Also, notice that very few papers with negative results (t-statistic less than 2.00) are published.

To summarize, researchers in our field face a complex agency problem. Ideally, our goal is to advance knowledge in our field of financial economics. As I show below, editors overwhelmingly publish papers with positive results. Realizing this, many authors believe that a necessary condition for publication is to obtain a positive "significant" result. This belief may lead authors to avoid certain projects, for example, those that involve time-intensive data collection, because they judge the risk to be too high. But these risky projects are exactly the type of initiatives that should be encouraged because they are often the ones that advance our knowledge the most.

In this address, I take a step back and examine how we conduct our research. Unfortunately, our standard testing methods are often ill-equipped to answer the questions that we pose. Other fields have thought deeply about testing. I share some of their insights in Section I. In Section II, I trace the history of the p-value and detail the American Statistical Association's six principles on the use of p-values. In Section III, I describe p-hacking and its detection. In Section IV, I discuss the problem of rare effects and how they impact our inference, and in Sections V and VI, I present a Bayesian perspective on hypothesis testing. I propose a simple Bayesian alternative that essentially involves a transformation of the usual p-value in Section VII, and I detail the limitations of this approach in Section VIII. In Section IX, I outline a set of best practices as well as offer recommendations designed to strengthen our research culture and encourage more risk-taking when making research and publication decisions. I offer concluding remarks in Section X.

I. Evidence from Other Fields

In contrast to financial economics, active research programs in other fields analyze how research is conducted. Many of these studies look across disciplines. For example, Fanelli (2010) studies how likely it is to get published in different fields with results that do not support the main hypothesis developed in the paper (see Figure 2).

Notice that journals in the Space Science field have little problem publishing articles that do not support the main hypothesis, while at the bottom of the list is the field of Psychology. The Economics and Business field is not that far away from Psychology.

The problem is more serious, however, than this snapshot in time suggests. Across all sciences, there is a distinct time trend as also detailed by Fanelli (2012). For example, in 1990, only 70% of papers in social sciences reported positive results but this proportion grows to 90% by 2007. Figure 3 shows that similar trends are evident in the physical and biological sciences.

So why do we see this pattern? Fanelli (2010) links it to Auguste Comte's (1856) famous Hierarchy of Science, with mathematics and astronomy on top and sociology on the bottom. As you move down the hierarchy, complexity increases and the ability to generalize results decreases. Fanelli uses the "hard" and "soft" classification. In the hard sciences, the results speak for themselves and are highly generalizable. In soft sciences, findings could depend

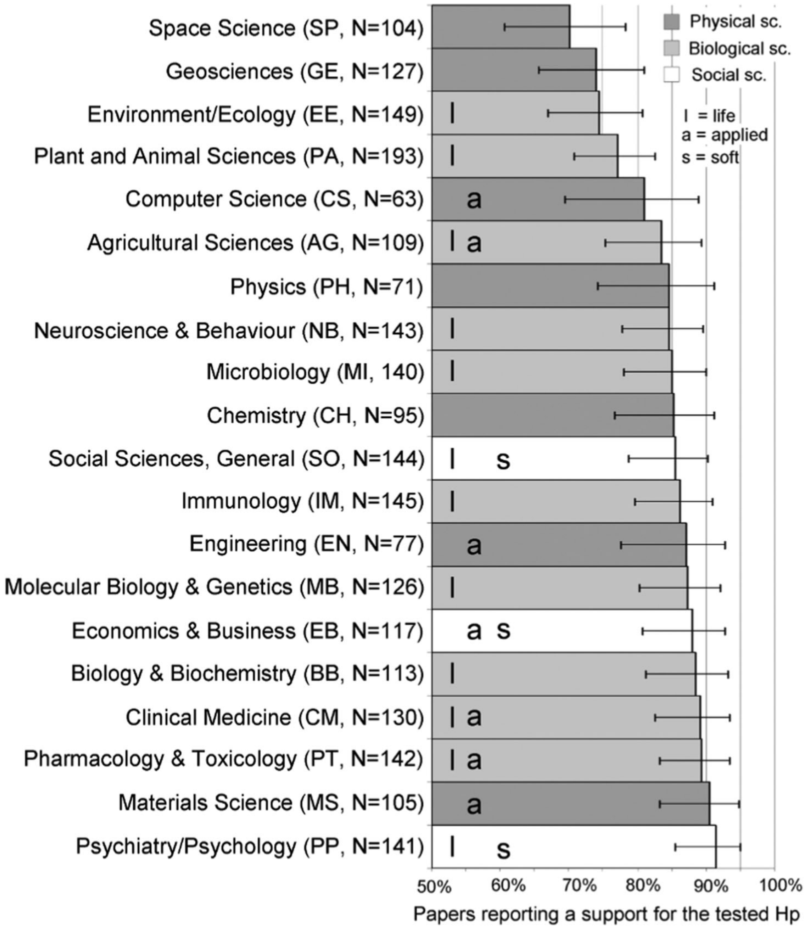
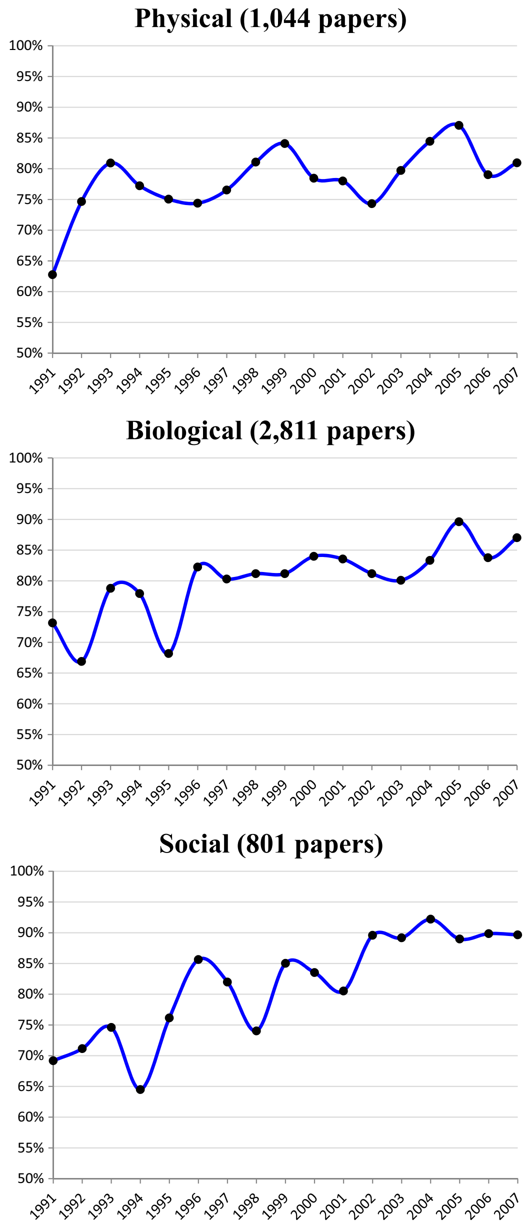


Figure 2. Percent of research papers reporting support for the tested hypothesis by field. Source: Fanelli (2010).

on the researcher's prestige in the community, political beliefs, and personal preferences—all of which could be important in selecting the hypotheses to be tested, the data to be collected, and the methods used in analyzing and interpreting the data—and are often too specific to generalize. A high-profile example of a hard scientific endeavor was the search for the Higgs boson. Importantly, "soft" does not mean "bad" and "hard" does not mean "good." Indeed, Comte (1856) considered sociology, which today would encompass economics and psychology, as "the principal band of the scientific sheaf."[1] Part of the complexity stems from the challenge in interpreting results when the human researcher interacts with the results. This interaction does not play much of a role in the hard sciences but can be impossible to avoid in the soft sciences.

[1] See Comte (1856, p. 457).



Physical (1,044 papers)

Biological (2,811 papers)

Social (801 papers)

Figure 3. “Positive” results increase over time in the science literature. Data from Fanelli (2012).

Returning to the Hierarchy of Science, why is it that most empirical articles published in economics and finance "support" the idea being tested? Here, I consider possible explanations proposed by Fanelli (2010) in the context of our field.

(1) *We have better theories in financial economics than, say, in particle physics*. This seems unlikely.

(2) *By observing phenomena, we have better prior beliefs in financial economics than in other scientific fields where priors are diffuse because there may be no direct or indirect observation*. This is a credible explanation given that researchers in our field accumulate knowledge over time about how markets work, and this knowledge often forms the basis for new theories or hypothesis tests. However, despite the importance of prior beliefs in financial economics research, they are often not taken into account in standard hypothesis tests, which impacts the way we interpret statistical evidence. I elaborate on this point below.

(3) *In financial economics, the hypotheses tested are often narrow and focused*. This also is a credible explanation. Compare, for example, the conclusion that the Higgs boson is a "central part of the Standard Model of particle physics that describes how the world is constructed" (Royal Swedish Academy of Sciences (2013)) to the hypothesis that companies with smaller boards of directors outperform or the hypothesis that stock returns are higher in January than in other months of the year. A narrower hypothesis may have a better chance of being supported when taken to the data.

(4) *The connection between theories, hypotheses, and empirical findings is more flexible in financial economics*. We have all seen theories tested where choices are made. Perhaps the most fundamental theory in finance is the Sharpe (1964), Lintner (1965), and Mossin (1966) Capital Asset Pricing Model. Many choices are made for testing. For example, what is included in the market portfolio? Do we test on U.S. or international assets? Does the theory apply to equities or all assets? What is the sample period? Should we test using portfolios or individual assets? Should risks be constrained to be constant? Are risk premiums constant? This is much different from flipping a 4 TeV switch, collecting trillions of new observations from the Large Hadron Collider, and searching for a specific particle decay signature as in the quest for the Higgs boson.

(5) *It is more likely that there are interaction effects between the researcher and the effect being researched*. This is a serious problem in experimental and survey-based research in economics and finance as well as other social sciences like psychology. For example, experimental subjects might be aware of the researchers' hypotheses and change their behavior. Another version of this problem involves confirmation bias in how the researcher interprets noisy data. While confirmation bias has been documented in many fields, it is most prevalent in behavioral research (Marsh and Hanlon (2007)).

(6) *Manipulation of the data and results*. Outright fraud is likely minimal in our field as most studies are conducted using data such as CRSP and Compustat that are readily available to other researchers (in contrast to, say, experimental fields, where the researcher creates the data for the study). However, the growth in the number of papers that use proprietary data decreases the probability of being caught and, in turn, the effective cost of fraud. In addition, many empirical design choices may be crucial for the results. For example, what should we include in the sample? Should the data be winsorized and, if so, at what level? Should certain outliers be discarded? Should the data be scaled by another variable? How many control variables should be included? What instruments should be used? Which estimation method should be used? These and similar choices are all tools of the p-hacker.

(7) *A lack of a replication culture*. In other fields, replication studies are published in top journals and often invalidate previous results (i.e., they present a negative result with a new experiment or new data set). In financial economics, in contrast, because much of the data are widely available to researchers (e.g., CRSP and Compustat), replication studies are less interesting and thus are rarely featured in top journals.

(8) *It is hard to publish findings that do not "support" the hypothesis being tested*. It is well-documented that editors and reviewers are more likely to reject negative findings that do not support the hypothesis tested in the paper (Song et al. (2000)). Research shows that papers with "significant" results receive more citations (Fanelli (2013)). This pattern can influence the way people conduct research. Above I discuss a striking truncation of marginally significant or insignificant results in top journals. One danger is HARKing (Hypothesizing After the Results are Known) (Kerr (1998)). Essentially, you have a theory that Y should be correlated with $X1$. You test that theory and include controls $X2$ to $X20$. You find no correlation between Y and $X1$, but you find a correlation between Y and $X7$. You then change the theory and throw $X1$ in with the control variables. HARKing is most likely to occur when it is relatively easy to run an experiment, for example, a regression that uses Computstat data as opposed to an experiment that requires over $5 billion in funding to construct the Large Hadron Collider (CERN (2017)).[2]

In short, the vast majority of papers published in financial economics provide "support" for the theory or hypothesis being tested for a variety of reasons. Some of these reasons are unavoidable, such as narrow hypotheses and the influence of prior information. However, it may be possible to minimize some of the other causes, such as p-hacking.

In an effort to address this issue, I begin by taking a step back. There is a reason I put the word "support" in quotations when referring to papers that present results that "support" the proposed theory or hypothesis: the traditional

[2] See CERN (2017, p. 17). See also https://press.cern/backgrounders/facts-figures.

hypothesis testing framework is unable to tell researchers the probability the hypothesis is true.

II. Understanding the *P*-value

The idea of using a *p*-value for hypothesis testing was introduced by Fisher (1925). His idea is to objectively separate findings of interest from noise. The null hypothesis is usually a statement of no relation between variables or no effect of an experimental manipulation. The *p*-value is the probability of observing an outcome or a more extreme outcome if the null hypothesis is true.

Goodman (1993) provides an excellent review of the history of the *p*-value, which I draw upon here. Fisher's approach focuses on measuring the strength of the evidence against the null hypothesis. He refers to the *p*-value as a "significance probability," that is, the probability of observing data as least as extreme as the actual outcome when the null hypothesis is true: if the *p*-value is less than a threshold of say 5%, the test rejects the null hypothesis. In Fisher's original framework, the *p*-value is not interpreted as the frequency of error if the experiment is repeated. Rather, it applies only to the data used in the test. This is a crucial philosophical point—because this approach applies only to the data under examination, inferences based on this approach may not generalize. Importantly, Fisher argues that the *p*-value should be used in conjunction with other types of evidence when available.

As I explain in more detail below, the most basic mistake in using *p*-values is to assume that a test with a *p*-value of 5% implies that there is only a 5% chance that the null hypothesis is true. This is a mistake because a *p*-value is calculated under the assumption that the null hypothesis is correct.

Neyman and Pearson (NP 1933) provide a different (and incompatible) framework that compares two hypotheses, the null and the alternative. Notice that in the original Fisher framework, there is no alternative hypothesis. NP introduce the idea of a Type I error (false positive rate, or rejecting the null when the null is true) and a Type II error (false negative rate, or failing to reject the null when the alternative is true). Their method focuses on a behavior or a decision-making rule, for example, rejecting the null and accepting the alternative, based on the observed data. In this framework, the (false positive) error rate, α, is set based on the particular experimental situation before the test is conducted (e.g., calibration of an instrument designed to find defects in the manufacturing process).[3] The researcher then reports whether the test falls

[3] The error rate is formally known as the "size" of the test. However, the term "size" has at least two ambiguities: it can be confused with the sample size and it can be confused with the magnitude of the effect. Hence, I prefer not to use the term. The error rate is also sometimes referred to as the significance level, but there is a subtle difference between the significance level and the error rate. The significance level is the *p*-value threshold used to reject a null. The error rate is the level of Type I error observed in following a particular testing procedure. For example, in small samples we can still follow the $t = 2.0$ rule. We would reject the null if the *p*-value implied by a normal distribution is below 5%. In this case, the significance level would be 5%. However, the actual error rate for the test would be higher than 5% since in small samples the normal approximation to a *t*-statistic is inaccurate. In this example, the significance level would be different from the error

into the critical region determined by the error rate α, for instance, p-value $<$ 0.05—not the particular magnitude of the p-value. NP's motivation was likely to reject the null and accept the alternative if the p-value $< \alpha$ or to accept the null and reject the alternative if the p-value $> \alpha$. However, aware of the limitations of their approach, NP observe that their framework "tells us nothing as to whether a particular case H is true . . . " when the p-value $< \alpha$ (p. 291). The key words here are "particular case."

Consider flipping a coin. The null hypothesis is that the coin is fair. Given that it is fair, we know what the null distribution looks like. Suppose we run a trial and it produces 10 heads in 10 flips. Under the null, the probability of that happening is 0.00097 ($=0.5^{10}$). With a pre-established cutoff of 0.05, we reject the hypothesis that the coin is fair. Notice, however, that the magnitude of the p-value itself is not informative—the key is whether we are in the rejection region.

However, I have described only one experiment. The result could be different in a second experiment. The argument in NP is that if you follow a certain behavior or rule and reject if p-value $<$ 0.05, then over the long run you will have a 5% error rate. To see what this means, consider an example that comes from product quality testing. A company uses a machine to manufacture a part and employs a separate instrument for quality control. The quality control instrument has a tolerance that can be physically established. With a cutoff or alpha of 5%, we know that over many trials 5% of the parts will be identified as defective when they are, in fact, not defective.[4] So on any given day, a part that has p-value $<$ 0.05 is thrown out. It might be the case that, on this particular day every single rejection decision is wrong. However, if this rule is followed over the long run, the company will make the right decision 95% of the time. Another example comes from a legal setting (Goodman (1999)). On any given day all innocent defendants may be convicted, but if the decision rule is followed over the long run, on average innocent people will be convicted only 5% of the time. These examples illustrate what NP meant by "particular case."

As I mention above, NP suggest that the error rate α be set according to the particular situation. As α is decreased, the Type II error rate will usually increase. Consider, for example, a household water heater failing because it is defective versus a jet engine failing because it is defective. In the case of the jet engine, we are willing to accept a lot of false positives (incorrectly label a part defective) to minimize chances of false negatives (miss detecting a defective part), so α is set to a higher level. The particular situation therefore dictates not only how low α will be set but also the Type II error rate.

In sum, the NP approach is deductive and can best be thought of as a decision rule that is useful if the decision is made many times. We assume that we know the true error rate, and if our observations fall into a critical region, we

rate. However, in most situations, we would like the significance level to be exactly the same as the error rate. Another way to think about the difference is that the significance level is the *desired* level of Type I error whereas the error rate is the *actual* level of Type I error.

[4] I do not mention the power of the test, that is, the probability of correctly rejecting the null hypothesis, here because I am focusing on the false positive error rate.

reject the null. Importantly, this approach is not informative about a "particular case"—it is valid over the long run. In contrast, the Fisher approach is inductive. We examine the evidence, and this evidence leads to an increased probability of a conclusion. Again, Fisher thought of the p-value as only one input to be used in the decision-making process—but this input can change beliefs.

There was a fierce debate between Fisher and NP. Over time, Fisher's p-value has been embedded in the NP framework in a way that often obscures its original meaning. NP introduced the idea of the Type I error rate, which is the false positive rate in repetitive experiments. The p-value for a test statistic is compared to the Type I error threshold to determine the test outcome, creating a subtle link between the p-value and the Type I error rate. As a result, people often mistakenly describe the p-value as the Type I error rate.

Interestingly, both approaches have the same goal: to provide an alternative to Bayesian decision making, which many considered too subjective because a prior belief has to be imposed before a test is conducted. While the two approaches are incompatible, they are usually lumped together under the label of null hypothesis statistical testing (NHST). With years of confusion, the difference between p-values, error rates, and significance levels has become blurred. Indeed, the very definition of p-value is now subject to confusion. For example, many incorrectly believe that p-values give the probability that the result could have occurred by chance alone.

To illustrate the extent of the confusion, suppose you have an experimental hypothesis that U.S. public companies with small boards of directors outperform companies with large boards. You create two value-weighted portfolios and test for differences in mean returns (with a host of controls such as industry effects).[5] The key parameter of interest, the mean performance difference, is significant with $t = 2.7$ (p-value $= 0.01$). Consider the following six statements (true/false):

(i) You have disproved the null hypothesis (no difference in mean performance).
(ii) You have found the probability of the null hypothesis being true.
(iii) You have proved the hypothesis that firms with small boards outperform firms with large boards.
(iv) You can deduce the probability of your hypothesis (small better than large) being true.
(v) If you reject the null hypothesis (no difference), you know the probability that you are making a mistake.
(vi) You have a reliable finding in the sense that if, hypothetically, the experiment were repeated a large number of times, you would obtain a significant result 99% of the time.

All six of these statements are "false."[6] The main issues are as follows:

[5] For this illustration, I simplify the empirical tests. See Yermack (1996) and Coles, Daniel, and Naveen (2008).

[6] I adapt this example from Gigerenzer (2004).

(1) The p-value does not indicate whether the null hypothesis or the underlying experimental hypothesis is "true." It is also incorrect to interpret the test as providing ($1 - p$-value)% confidence that the effect being tested is true. Hence, both (i) and (iii) are false.

(2) The p-value indicates the probability of observing an effect, D, (or greater) given the null hypothesis H_0 is true, that is, $p(D \mid H_0)$. It does not tell us $p(H_0 \mid D)$, and hence (ii) is false.

(3) The p-value says nothing about the experimental hypothesis being true or false, and hence (iv) is false. Question (v) also refers to the probability of a hypothesis that the p-value does not address, and hence (v) is false.

(4) The complement of the p-value does not tell us the probability that a similar effect will hold up in the future unless we know the null is true—and we do not. Hence, (vi) is false.

There are also a number of additional issues:

(5) The p-value is routinely used to choose among specifications, for example, specification A has a lower p-value than specification B and hence we choose specification A. However, comparing p-values across specifications has no statistical meaning. The p-value for one specification does not tell us that the other specification is true.

(6) A low p-value, while rejecting the null hypothesis, tells us little about the ability of the hypothesis to explain the data. For example, you might observe a low p-value but the model has a low R^2.

(7) Low p-values could result from failing to control for multiple testing.

(8) Low p-values could result from selection and/or p-hacking.

(9) Low p-values could result from a misspecified test.

(10) P-values crucially depend on the amount of data. It has been well-known since Berkson (1938, 1942) that, with enough data, you can reject almost any null hypothesis.

(11) P-values do not tell us about the size of the economic effect.[7]

Let me emphasize here the second point above: it is fundamentally important for researchers to answer the right question. A p-value tells us $p(D \mid H)$. However, it is often interpreted incorrectly as indicating $p(H \mid D)$. Carver (1978, pp. 384–385) colorfully illustrates how serious this mistake is as follows:

> What is the probability of obtaining a dead person (label this part D) given that the person was hanged (label this part H); this is, in symbol form, what is p(D | H)? Obviously, it will be very high, perhaps 0.97 or higher. Now, let us reverse the question. What is the probability that a person has been hanged (H), given that the person is dead (D); that is, what is p(H | D)? No one would be likely to make the mistake of substituting the first estimate (0.97) for the second (0.01); that is, to accept 0.97 as the

[7] A related issue is that an economically and statistically small change in a parameter may lead the parameter to change from "insignificant" to "significant." See Gelman and Stern (2006).

> probability that a person has been hanged given that the person is dead. Even though this seems to be an unlikely mistake, it is exactly the kind of mistake that is made with interpretations of statistical significance testing—by analogy, calculated estimates of p(H | D) are interpreted as if they were estimates of p(D | H), when they are clearly not the same.

The confusion surrounding the use of *p*-values has led to considerable discussion among scientists (Nuzzo (2014)). Indeed, in an extraordinary move in 2015, a social psychology journal banned the publication of *p*-values in any of their papers (Trafimow and Marks (2015)). Other scientific fields such as epidemiology have also demonstrated an aversion to publishing *p*-values (Lang, Rothman, and Cann (1998)). Related concerns recently prompted an intervention by the American Statistical Association.[8]

A. The American Statistical Association's Six Principles

Concerned about the growing crisis in the conduct of scientific experiments,[9] and the declining public confidence in the integrity of scientific experiments, the Board of Directors of the American Statistical Association tasked Ronald Wasserstein to assemble senior experts in the field with the goal of developing a statement that would help correct the course. In March 2016, the Association released "Statement on Statistical Significance and *P*-Values" (American Statistical Association (2016)). The statement and associated paper (Wasserstein and Lazar (2016)) is a reaction to the perception that much of the scientific field misunderstands statistical significance, and that in many cases researchers are substituting *p*-values for scientific reasoning. The statement decries the fact that "the *p*-value has become a gatekeeper for whether work is publishable" and makes specific references to both *p*-hacking and data dredging (p. 1).

Below I reproduce the key principles outlined in the statement and comment on each one in the context of our field.

(1) *P-values can indicate how incompatible the data are with a specified statistical model*. The specified statistical model might be the null hypothesis of no effect. For example, consider the null hypothesis that expected stock returns are constant. The alternative is that past returns predict future returns. A low *p*-value tells us that the observed data appear to be incompatible with the null hypothesis.

(2) *P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone*. Returning to the stock autocorrelation test, a low *p*-value is inconsistent with the hypothesis that expected returns are constant. A low *p*-value

[8] In a famous paper published in *PLoS Medicine*, Ioannidis (2005) argues that "most published research findings [in his and related fields] are false." It is the most read and shared article in *PLoS Medicine*'s history with 2.2 million views. Of course, as some argue, in medicine the stakes are higher than in financial economics: life or death.

[9] See also Gelman and Loken (2014).

does not imply that the autocorrelation model is “true.” In addition, a high p-value does not mean that the hypothesis that expected returns are constant (i.e., stock returns are pure noise) is “true.” Many other tests (e.g., introducing other information like dividend yields) may show that the data are not consistent with the null hypothesis.

(3) *Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold*. It is rare that there is a clean dividing line between “true” and “false,” and a p-value should not be used as such a dividing line. Rather, it is crucial to account for other factors such as the plausibility of the theory and its assumptions, the quality of the data, and any other evidence that might be relevant for the study. Indeed, this point is explicitly recognized by Fisher (1925), who notes that p-values do not substitute for “critical examinations of general scientific questions, which would require the examination of much more extended data, and of other evidence . . . .”

(4) *Proper inference requires full reporting and transparency*. Reporting select results that have low p-values makes it impossible to interpret the analysis. Such practice amounts to p-hacking and borders on academic fraud. Indeed, the American Finance Association’s *Code of Professional Conduct and Ethics* states in Section 6d(4) that “Financial economists should not selectively report findings in ways that would mislead or deceive readers” (2016). For interpretation, it is essential that researchers reveal the full extent of all hypotheses tested. Internet appendices have made it easier to provide all test results.

(5) *A p-value, or statistical significance, does not measure the size of an effect or the importance of a result*. On the size of an effect, I believe that financial economists do a much better job than researchers in other fields. Articles in finance journals routinely report both economic and statistical significance. Economic significance is often described as the impact of moving from the first quartile to the third quartile of the distribution of the variable in question (it should not be the impact of moving from the 1st to the 99th percentile). The size of an effect is often ignored or misconstrued in other fields. For example, a study published in the prestigious *Proceedings of the National Academy of Science* (*PNAS*) that involved a large sample of over 19,000 married couples concluded that couples that met online were happier than those that met offline, with p-value < 0.0001, but a careful reading of the paper reveals that the size of the effect is tiny: the couples that met online scored 5.64/7.00 while those that met offline scored 5.48/7.00 (Cacioppo et al. (2013)).[10] And misconstruing the size of an effect is routine in medicine (Hutton

[10] However, it is crucial to know the variation in happiness across couples. Cacioppo et al. (2013, p. 10,136) report that $M = 5.64$, $SE = 0.02$, and $n = 5{,}349$ for one group while $M = 5.48$, $SE = 0.01$, $n = 12{,}253$ for the other group. If the mean is 5.64 and the standard error for the mean is 0.02, then the individual standard error is about $0.02 \times \sqrt{5{,}349} = 1.4$. So there is a lot of variation across individuals.

(2010)). For instance, as a result of a large-scale study on the impact of statin drugs on nonfatal coronary events, an advertisement proclaimed that these drugs "reduce the risk of heart attack by 36%."[11] Yet the incidence of heart attacks reported in the study (Sever et al. (2003, table 3)) was 2.7 per 100 among the placebo and 1.7 per 100 among the patients taking the statin. Thus, while the relative risk decreases by 36%, the absolute risk decreases by only 1%.

While the reporting of economic significance is fairly routine in finance journals, it is common to see the effect of variables with a low *p*-value described using the word "strong." It is possible, however, to observe a low *p*-value for an effect that explains only a modest part of the variation in the measure of interest.

(6) *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis*. A low *p*-value offers only weak evidence against the null hypothesis, and such evidence is especially weak if you believe that the incidence of the effect in question is rare (see Section VI below). Similarly, a high *p*-value does not imply that the null hypothesis is true. It is therefore important not to stop the analysis once a *p*-value is obtained. Potentially, additional data can be gathered to bolster confidence in the conclusions.

III. *P*-hacking

As I mention in the introduction, editors' focus on publishing the most significant results can induce authors not to submit articles with weak results or to cherry-pick the results that they submit to a journal, that is, to engage in *p*-hacking. Focusing here on the latter, there are many forms of *p*-hacking, with some more egregious than others. For example, running thousands of correlations and reporting only the most significant one constitutes academic fraud. Such a practice is red flagged by the American Statistical Association (2016) and is contrary to the American Finance Association's *Code of Professional Conduct and Ethics*. A more subtle version of this type of *p*-hacking is studying correlations among, say, 11 variables and choosing to report only 10 of the correlations. Unfortunately, not reporting all variables examined in empirical work is commonplace in financial economics.

P-hacking is not limited to the reporting of individual explanatory variables. Often researchers investigate aggregation schemes. For example, suppose 11 variables are tried and none pass the usual test for "significance." Suppose further that various aggregations are tried and the sum of three of the 11 variables passes the usual hurdle for significance. Only reporting this particular aggregation amounts to *p*-hacking.

P-hacking also occurs when the researcher tries a number of statistical approaches (e.g., linear probability vs. Logit or Probit, panel regression vs. Fama-MacBeth (1973), Newey-West (1987) lag 1 vs. lag 4, different clustering

[11] A full-page ad appeared in the *Wall Street Journal*, November 6, 2007, p. A13.

choices, different choices of instrumental variables) and reports the one with the most impressive "significance." Indeed, self-serving reasoning often leads researchers to convince themselves that the most significant result is the one the researcher originally intended to examine.

Data manipulation and exclusion can also lead to *p*-hacking. Researchers make many choices in terms of standardization, log or other transformations, winsorization, and outlier exclusion. If these choices lead to the most significant results being presented, this is *p*-hacking. Similarly, the choice of data set can lead to *p*-hacking. For example, if the researcher reports a significant result using the 1970 to 2017 period and does not reveal that the same result is weaker in the 1960 to 2017 period, this is *p*-hacking. *P*-hacking is more difficult to detect with proprietary data where replication is costly. If a researcher intends to hand-collect data on say 1,000 companies, starts to conduct analysis midway through the data collection, and stops the data collection upon finding a significant result, this too is *p*-hacking (Simonsohn, Nelson, and Simmons (2014), Head et al. (2015)).

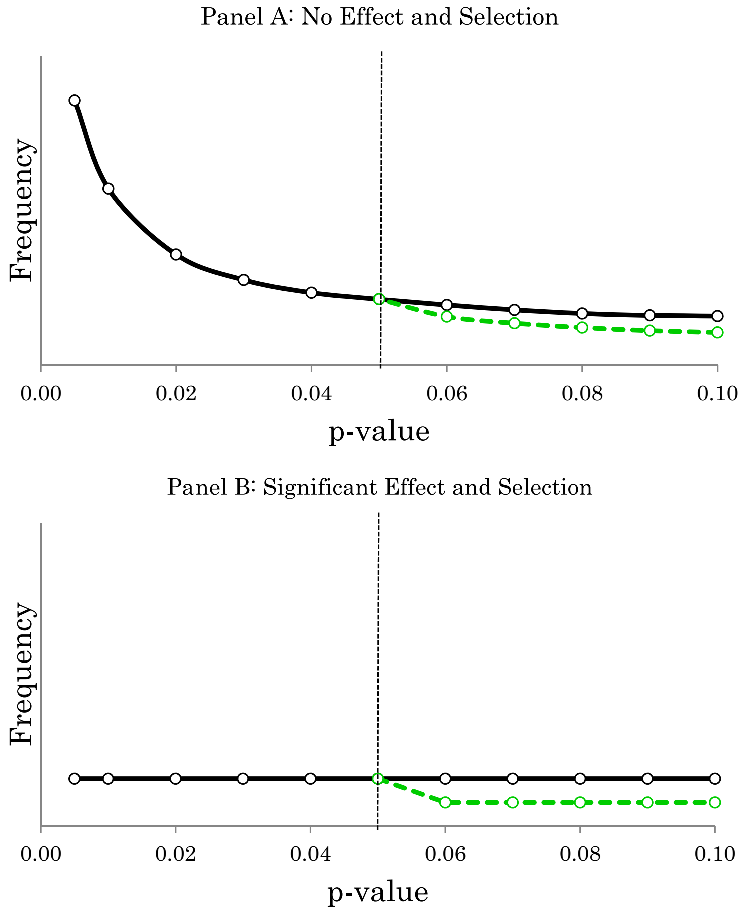
Indeed, researchers might not even know that their results reflect *p*-hacking. For example, delegating analysis to a research assistant may lead to *p*-hacking if the assistant searches for a result that they think the researcher will be "pleased" with.

To gauge the degree of publication bias, one can look at the distribution of *p*-values. Both selection (file drawer effect) and *p*-hacking induce patterns in the distribution of *p*-values. If there is no effect (e.g., there is no reason for *Y* to be correlated with *X*), the *p*-value distribution should be flat (e.g., a 10% probability of obtaining a *p*-value of 0.10 or less and a 2% probability of obtaining a *p*-value of 0.02 or less when the null hypothesis is true), as in the solid line in Panel A of Figure 4. For example, in my ticker symbol exercise, the distribution of *p*-values is flat. However, if there is an effect, the distribution is not flat (there are substantially higher probabilities of getting a *p*-value of 0.05 than a *p*-value of 0.10 or 0.20), as shown in the solid curve in Panel B of Figure 4.

The dashed lines in Figure 4 show the effect of selection. If papers are not submitted to journals with marginal *p*-values (say 0.05 to 0.10), then the distribution shifts downward to the right of 0.05 (dashed lines).

Using this same type of analysis, Figure 5 shows the impact of *p*-hacking. The figure plots the same initial *p*-curves as in Figure 4 using solid lines and the impact of *p*-hacking using dashed lines. The area of interest is below a *p*-value of 0.05. In both panels, there are more results than expected just below the 0.05 threshold. Thus, as can be seen, *p*-hacking has a much different effect from selection: while selection decreases the number of papers published, *p*-hacking increases the number of "significant" results that are published.

Using data from HLZ (2016), Figure 6 plots the distribution of *p*-values in factor studies. There is good news and bad news. The good news is that the curve is not flat, which is what you would expect to see if the null hypothesis of no factors were true. The bad news is that the selection effect is obvious, which is also evident in Figure 1, where a surprisingly small number of studies are


Panel A: No Effect and Selection

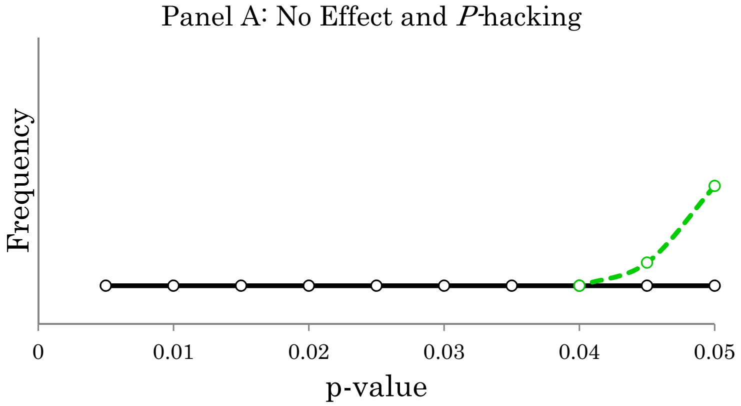
Panel B: Significant Effect and Selection

Figure 4. Impact of selection (file drawer effect) on *p*-values. Panels from Head et al. (2015). The solid lines represent the distribution of *p*-values. The dashed lines show the impact of selection.

published with $2.0 < t < 2.57$. In effect, the *p*-value threshold for publication in factor studies is not 0.05 but 0.01 for top finance journals.

IV. The Problem of Rare Incidence and Improbable Effects

A rare occurrence is an event that occurs very infrequently, such as the onset of a rare disease. An improbable occurrence is different; in this case, the effect is implausible. Both pose the same challenge to inference using traditional statistical tools. In most areas of finance research, the likelihood that empirical researchers uncover a true causal relation is small. When guided by theory, the probability could be higher, but not by much given the multiplicity of theories available for each area of research and the fact that some theories are constructed to fit known facts.



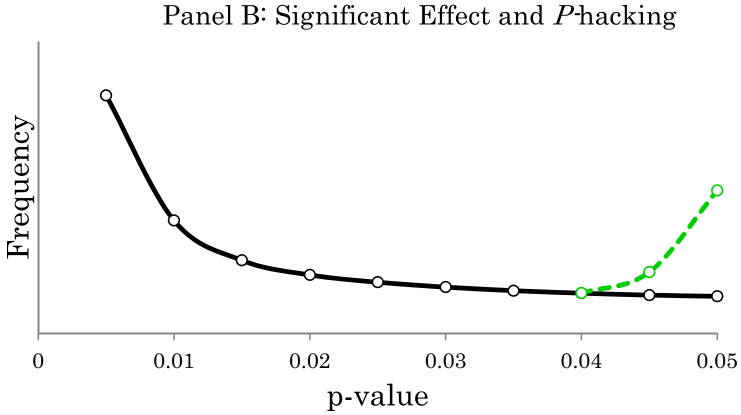


Figure 5. Impact of *p*-hacking on *p*-values. Panels from Head et al. (2015). The solid lines represent the distribution of *p*-values. The dashed lines show the impact of *p*-hacking.

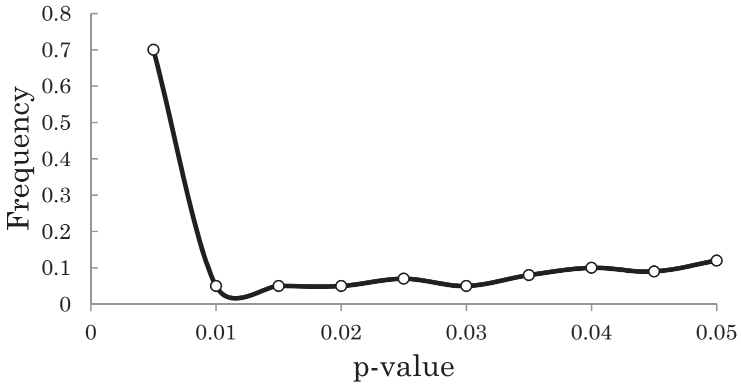


Figure 6. The distribution of *p*-values in factor studies. Data from Harvey, Liu, and Zhu (2016).

When a particular effect is unlikely, classical hypothesis testing often leads to severely biased testing outcomes. Unfortunately, such bias is generally not mentioned in empirical research. This may be changing after Bem's (2011) positive finding on extrasensory perception published in a top journal drew many top statisticians' critiques (e.g., Francis (2012), Ritchie, Wiseman, and French (2012)).

I illustrate the problem with a rare incidence example from medical science. For women between the age of 40 and 50, breast cancer is a relatively rare event that impacts only 1% of this age group. The first-line test, a mammogram, is 90% accurate in diagnosing cancer. This is the "power" of the test, that is, the probability of correctly rejecting the null hypothesis of no cancer. Let us assume that the rate of false diagnosis, or the error rate, is 10%. Suppose a woman is told that the test is positive. What is the probability that this woman has breast cancer?

Suppose that among 1,000 cases, 10 (=1,000 × 1%) have breast cancer and the remaining 990 do not. For those with breast cancer, the rate of successful diagnosis is 90%, so 9 of 10 will be correctly diagnosed as having breast cancer. For those without breast cancer, 10% or 99 (=990 × 10%) will be incorrectly identified as having breast cancer. The rate of incorrectly rejecting the null (or Type I error) is often referred to as the "error rate" of the test or the significance level.

In this example, the error rate is a combination of the accuracy of the mammogram machine and our tolerance for false negatives (missed diagnoses). Thus, if we raise the significance level from, say, $t > 1.6$ to $t > 2$ (say 5%), this might mean that more people with cancer are misdiagnosed as free of cancer, while if we set the threshold so low that all 10 women with cancer are correctly diagnosed, there will be a huge number of false positives.[12]

If the mammogram machine has been running a long time, we can learn about its performance at various test thresholds. It might be the case that, over the long run, setting the threshold at $t > 1.6$ produces a 10% false positive rate. However, in most applications we do not have this luxury—we are considering a new data set or running a new test on a well-known data set.

Turning back to the example, a total of 108 (=9 + 99) patients test positive for cancer and the rate of false positives is 92% (=99/(9 + 99)), even though individually the rate of false positives is only 10%. The high rate of false positives in breast cancer diagnoses is a well-known issue in medical research that has spurred the development of enhanced mammogram tests that reduce this rate by increasing the power of the test.

Consider another example. Nosek, Spies, and Motyl (2012) examine the hypothesis that political extremists see only black and white, literally. In an Internet-based experiment, they presented 1,979 participants with

[12] In general, there is no universal relation between a test's power and the error rate. In most situations, power increases as the error rate increases. Even in the simplest testing scenario, where we test for the mean difference, the exact relationship between the error rate and the test's power depends on the sample size.

noncontroversial words that appeared in different shades of gray. The participants were asked to identify the shades of gray from a color gradient. Determining the accuracy of the choice was straightforward because the color of each word had an exact match on the gradient.

Following the experiment, the participants were classified into two groups—political moderates and political extremists. Moderates saw shades of gray more accurately than left- or right-wing extremists, with a p-value of 0.001 providing strong evidence against the null hypothesis of no difference. Obviously, this is a striking finding that potentially links physiology to political beliefs.

Before submitting the paper for publication, however, the authors replicated the experiment using 1,300 additional participants. The replication p-value was vastly different, at 0.59. As Nuzzo (2014, p. 151) writes, the "more implausible the hypothesis—telepathy, aliens, homeopathy—the greater the chance that the exciting finding is a false alarm, no matter what the p-value is."

What insight does the above discussion have for research in financial economics? The key takeaway is that the more improbable the effect, the more careful we have to be because there will be many false positives. Let π be the probability of encountering a true causal relationship. In the context of hypothesis testing, α denotes the significance level when the null is true and β denotes the power of the test (the probability that the test will reject the null when the alternative is true). If tests are independent, the expected fraction of false discoveries is given by

$$\text{Expected fraction of false discoveries} = \frac{\alpha}{\frac{\pi}{1-\pi}\beta + \alpha},$$

where $\frac{\pi}{1-\pi}$ is the odds ratio of true versus false hypotheses.

We can use the above formula to calibrate the false discovery rate for research in financial economics. As we often do in empirical research, we fix the test threshold α at a prespecified level (e.g., 5%). Holding α constant, the minimum level of the expected fraction of false discoveries is $\frac{\alpha}{\frac{\pi}{1-\pi}+\alpha}$, which is achieved when the test's power (i.e., β) is 100%. However, if π is much smaller than α (i.e., the effect is rare), then this minimum level is approximately $1/(1 + \pi/\alpha)$, which is close to one. Hence, if a true discovery is unlikely, then no matter how powerful the test is, the probability of a false discovery is high. This aggregate view of false discoveries in finance research, which is distinct from the usual concern for single-hypothesis tests, casts doubt on the credibility of many empirical findings.

Table I reports false discovery rates for various degrees of unlikeliness for the effect in question and various significance levels for three levels of test power. Even when the test power is 100%, which means that all women with cancer are correctly diagnosed, the false positive rate is an alarming 83% at the 0.05 level of significance.

Among the many specifications in Table I, the significance level is nominal only when the odds of the null hypothesis being true are 1:1. How often do

Table I

The Rate of False Discovery Depends on Test Power and Error Rate

This table shows the probability that the null hypothesis is true for three levels of power.

	Prior Beliefs (Odds, Null:Alternative)					
	Long Shot					Even Odds
	99-to-1	49-to-1	24-to-1	19-to-1	4-to-1	1-to-1
	0.01	0.02	0.04	0.05	0.20	0.50
Panel A: Power = 1.00						
Significance level (α)						
0.100	0.91	0.83	0.71	0.66	0.29	0.09
0.050	0.83	0.71	0.55	0.49	0.17	0.05
0.010	0.50	0.33	0.19	0.16	0.04	0.01
0.001	0.09	0.05	0.02	0.02	0.004	0.001
Panel B: Power = 0.9						
Significance level (α)						
0.100	0.92	0.84	0.73	0.68	0.31	0.10
0.050	0.85	0.73	0.57	0.51	0.18	0.05
0.010	0.52	0.35	0.21	0.17	0.04	0.01
0.001	0.10	0.05	0.03	0.02	0.004	0.001
Panel C: Power = 0.8						
Significance level (α)						
0.100	0.93	0.86	0.75	0.70	0.33	0.11
0.050	0.86	0.75	0.60	0.54	0.20	0.06
0.010	0.55	0.38	0.23	0.19	0.05	0.01
0.001	0.11	0.06	0.03	0.02	0.005	0.001

we see such odds in empirical research in our field? Take, for example, the literature on return predictability. 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? Very few. However, we do not take these prior beliefs into account when calculating and interpreting p-values. It is therefore not surprising to see that most of the factors identified as return predictors are dismissed in out-of-sample tests as in Welch and Goyal (2008).

A more interesting exercise is to think about the time evolution of the rate of false discoveries for a particular strand of research. Take, for example, the discovery of anomalies. HLZ (2016) document an explosion of anomaly discovery over the last two decades. They argue that, under traditional single-hypothesis tests, the rate of false discoveries should also increase. The rationale is that the ex ante probability of running into a true anomaly should decline over time.

There are several reasons for a declining rate of anomaly discovery. First, true anomalies should become more scarce over time as the low-hanging fruit has been picked. Second, as we run out of theories based on first principles, we rely more on specialized theories, which inevitably imply a lower rate of true discovery ex ante. Finally, relative to the large number of financial variables and firm characteristics that one can explore, the number of securities is limited.

The discussion so far bears a simple Bayesian interpretation. If the prior probability of a true discovery is low, then the posterior probability is also likely to be low across a wide range of choices for error rates and test power. Such posterior probabilities, when compared to the often-reported frequentist p-values that are based on the particular hypotheses being analyzed, should be more helpful for researchers to guard against false discoveries from a population perspective.

V. We Are All Bayesians

If you are rational, then you are a Bayesian. The most rudimentary example of the link to rationality is Bayes' Rule: we have some belief about an effect, we observe the data, and then we revise our belief. The key addition in the Bayesian statistical framework is incorporating prior beliefs. Indeed, if you accept the argument that the false positive rate should be higher for theories that are unlikely, then you have already adopted a fundamentally Bayesian line of reasoning.

Let me motivate the use of prior beliefs with an example from a famous letter written by Leonard Savage in 1962 (see Greenhouse (2012) and Churchill (2014)). Consider three experiments. In the first experiment, a musicologist claims to be able to distinguish between pages of music written by Mozart or Haydn. Using 10 pairs of score pages, Mozart versus Haydn, the musicologist identifies each one correctly. In the second experiment, an elderly woman claims that she can tell if milk was put in a tea cup before or after the hot tea was poured in. Again using 10 different trials, the woman identifies each one correctly. In the third experiment, a patron at a bar claims that alcohol helps him predict the future. You conduct an experiment using 10 coin flips and the patron correctly guesses heads or tails for each flip.

In each of these experiments, the p-value is less than 0.001 ($=1/2^{10}$), which is highly significant under the usual standards. However, what you take away from each test should be different. In the first test, you know that the test subject is a musicologist. This is her area of expertise and there is little reason for her to make a false claim. Indeed, it is not even obvious that you need to run the experiment to verify her claim, but in doing so the experiment confirms what you already believe. In the second case, maybe you are initially skeptical of the claim, but the evidence convinces you. Personally, I remember my grandfather sending his tea back at a restaurant because he always asked for the tea bag to be put in the pot before the water was added; he knew when the server got it wrong. So in the second experiment, there is a shift in your beliefs given the outcome of the experiment. What about the third experiment? The claim that alcohol causes clairvoyance is preposterous. Thus, while the p-value is less than 0.001, you chalk this up to luck—such an occurrence should happen naturally once in every 1,024 trials—and your beliefs are largely unaffected.

These simple examples illustrate two points. First, as I discuss earlier, if the effect is improbable (a drunk foretelling the future), there will be a lot of false positives using the standard NHST framework. Second, and more importantly, what really counts is the ability of the test to change beliefs. In the first and third tests, beliefs do not change as a result of the evidence. In the second test, beliefs are updated.

VI. The Bayesian Critique

Most empirical research in financial economics is conducted in the classical mode. There are many reasons for this. For instance, many researchers are uncomfortable specifying prior beliefs, as the imposition of a prior may be arbitrary in that it impacts the results. The perceived computational burden is also higher under a Bayesian approach, as are the econometrics—not just to the researcher but also to the readers of the research.

The longstanding debate on hypothesis testing between Bayesian and frequentist statisticians likely originated with Berkson's (1938) observation that you can reject almost any null hypothesis with enough data. My goal is to explore a few ideas that are easy to implement and that provide supplementary information to the usual test statistics. Before doing so, however, it is useful to review the objections that Bayesians have to NHST.[13]

(1) *Probability is a long-run concept for a frequentist*. The probability that comes out of the classical test assumes that we have many hypothetical samples and that the probability is the limiting frequency over these many samples. Using a Bayesian approach, no such assumption is necessary. The only data that are relevant are the data involved in the test.

(2) *There are issues with the structure of the hypothesis test*. The null hypothesis is often set to zero. For example, we test mutual fund manager performance against a null of zero excess returns. Why zero? Similarly, we compare the difference between two populations and start with the null that there is no difference—but they will never be exactly equal. Indeed, the null hypothesis is unlikely to ever be true (Cohen (1994)) and we are sure to reject the null hypothesis with enough data. In addition, under NHST, there is no direct inference on the alternative hypothesis. In my earlier example of autocorrelation in stock returns, NHST may reject the null hypothesis of no serial correlation but it cannot tell us whether the alternative hypothesis is true. Finally, what is so special about the significance level of 0.05? This is an arbitrary cutoff.

(3) *Prior information is ignored*. This point is emphasized in Savage's examples involving the musicologist, the tea drinker, and the bar patron. There is no way to formally incorporate prior information into the classical approach.

[13] This list comes in part from http://www.stats.org.uk/statistical-inference/, which also contains an extensive bibliography.

(4) *P-values are subject to manipulation*. Earlier I presented a p-hacking example where a researcher intending to hand-collect information on 1,000 companies stops upon finding "significant" results. Using the Bayesian approach may not necessarily cure this problem—a Bayesian might also stop data collection in the middle of the collection process—but a Bayesian might use one part of the data to inform additional analysis of another part of the data (Jarosz and Wiley (2014)). Of course, the Bayesian chooses the prior, which some might consider the ultimate "hack." However, the prior is transparent and skeptical readers can use whatever prior they think is appropriate.

(5) *P-values do not really answer the right question*. As I emphasized earlier, p-values do not tell us the probability of the null hypothesis being true given the data, $p(H_0 \,|\, D)$. Rather, they simply indicate the probability with which the particular evidence will arise if the null hypothesis is true, $p(D \,|\, H_0)$ (Carver (1978)).

VII. A Compromise

Let us begin with a not-so-modest proposal, namely, the full-blown Bayesian approach. Using this approach, we first take prior information into account by specifying a prior on all possible hypotheses—for each hypothesis, we calculate the data likelihood given the null. Then, using Bayes's theorem, we derive the posterior probability for each hypothesis. The full-blown Bayesian approach results in a probabilistic assessment of all hypotheses given the data and thus involves inductive reasoning.

The downside of this approach is that it requires a prior, which is often hard to come by. This raises the question of whether we can avoid using a prior and still enjoy the advantages of the Bayesian approach.

The original NP framework, which requires that the researcher specify both a null and an alternative hypothesis, partly achieves this. It shares several features of the full Bayesian model. For example, the NP test statistic—the likelihood ratio—weighs the likelihood of the data under the null against the likelihood under the alternative. This is similar to the Bayesian approach where the likelihood ratio is used to adjust the ratio of the prior likelihood under both the null and the alternative. But there are important differences between the NP and the full Bayesian approaches. In particular, the NP approach relies on the distribution of the test statistic under the null hypothesis to make inferences and does not provide a probabilistic assessment of the relative likelihood of the null and the alternative. As such, it inherently involves deductive reasoning.[14]

Unfortunately, we usually do not have well-specified alternatives in financial economics. As a result, use of the NP approach is limited. This begs the question

[14] Under the NP approach, if the null is true, we are unlikely to observe the data so we reject the premise that the null is true. This is deductive reasoning. Under the Bayesian approach, given the data, we assign probabilities to different premises. This is inductive reasoning.

of whether we can adapt the Bayesian approach to obtain a metric that does not depend on the specification of the alternative.

The minimum Bayes factor (MBF) under the null, as developed by Edwards, Lindman, and Savage (1963), is one such metric. Let us start with the definition of the Bayes factor under the null. Under Bayes's theorem, the posterior odds ratio between the null model and the model under alternative hypotheses equals the prior odds ratio—the odds ratio before we see the data—times the Bayes factor, which is the ratio of the data likelihood under the null and the data likelihood under alternative hypotheses.[15] In other words, the Bayes factor measures how far we move away from the prior odds ratio after seeing the data.

In general, the Bayes factor still depends on the prior specification of the alternative hypotheses. But a special version of the Bayes factor, the MBF, does not. The MBF is the lower bound among all Bayes factors. It occurs when the density of the prior distribution of alternative hypotheses concentrates at the maximum likelihood estimate of the data. In other words, if prior information makes you focus on one particular alternative hypothesis that turns out to be the value that is most supported by the data, then you have the MBF. Because the MBF indicates the maximum amount the data can move you away from your priors in terms of the odds ratio, it is the Bayes factor that provides the strongest evidence against the null hypothesis.

Consider the following example. Suppose you have 1,000 observations of returns. Your null hypothesis is that the average return is zero. The average return in the data is 0.04. Your prior odds ratio is 3/7, that is, you believe that the null is true with 30% probability. Which prior specification of the alternatives will deliver the strongest evidence against the null hypothesis? It is obvious that setting all of the 70% prior mass at 0.04 achieves this.

Consider another example. Suppose the MBF is 1/10. This implies that the data suggest that we should multiply our prior odds ratio by a factor of at least 1/10. If the prior odds ratio is 1/1 (50% probability, that is, even odds), then the posterior odds ratio becomes 1/10, which corresponds to a probability of 9% (= 1/(10 + 1)) for the null hypothesis. To achieve a posterior probability of 5%

[15] In particular, we have $\frac{f(\theta_0|data)}{f(alternative|data)} = \frac{f(data|\theta_0)}{\int f(data|\theta)\pi_A(\theta)d\theta} \times \frac{\pi_0}{\pi_1}$, where $\pi_A(\theta)$ is the probability density under the alternative hypothesis, $f(x|y)$ is probability density function of x conditional on y, and π_0 and π_1 are the prior probabilities for the null and alternative hypotheses. Note, for frequentists, $f(data|\theta_0)$ is usually called the probability density function and $f(\theta_0|data)$ is called the likelihood function. But for Bayesians, data and parameters are both random, so we use likelihood and probability for both $f(data|\theta_0)$ and $f(\theta_0|data)$. The posterior odds ratio is $\frac{f(\theta_0|data)}{f(alternative|data)}$, the prior odds ratio is $\frac{\pi_0}{\pi_1}$, and the Bayes factor is $\frac{f(data|\theta_0)}{\int f(data|\theta)\pi_A(\theta)d\theta}$. Notice that, while the null hypothesis is necessarily a single number, we can have a continuous density on alternative hypotheses. In this case, the data likelihood under alternative hypotheses is given by $\int f(data|\theta)\pi_A(\theta)d\theta$, where $\pi_A(\theta)$ is the probability density under the alternative hypothesis. In the special case in which the alternative hypothesis is a single number, $\pi_A(\theta)$ reduces to a point mass at θ_A and the data likelihood under the alternative hypotheses becomes $f(data|\theta_A)$.

for the null hypothesis (posterior odds ratio = 5/95), the evidence needs to be stronger (i.e., the MBF needs to be smaller).[16]

What makes the MBF even more useful is that it is easy to calculate in many situations. If a test is based on a normal approximation and the z-score is given by Z, then the MBF is simply $\exp(-Z^2/2)$ where $\exp(\cdot)$ denotes the natural exponent. Hence, a z-score of 2.0, which has a p-value of 5%, corresponds to an MBF of 0.14.

Consider yet another example. Suppose our null hypothesis is that the variable Y is not predictable. We regress Y on X with 300 observations and find a "significant" coefficient with a t-statistic of 2.6, which has a p-value of 0.014. The MBF derived from this alternative (i.e., the estimated value of the coefficient) is 0.034 ($\exp[-2.6^2/2]$). Suppose further that we think there are even odds that the null is true relative to the alternative. This means that there is a 0.033 chance the null is true (MBF × prior odds = 0.034 × 1.0 = 0.034, probability of null = 0.034/[1 + 0.034] = 0.033). However, if you think there are modest odds against the effect, say 2:1, the probability that the null is true increases to 0.064. Essentially, we are creating "Bayesianized" p-values that tell us whether things are true instead of just rejection probabilities conditional on the null.

The MBF is most useful when we do not have any information about the specification of alternative hypotheses, as it is a "global" MBF across all prior specifications of alternative hypotheses. Of course, sometimes we do have information or priors on plausible alternatives. We can modify the MBF to incorporate such information. One useful adaptation calculates the MBF when we believe that the prior probability density for alternatives should be symmetric and descending around the null hypothesis. Let us call this MBF the SD-MBF, where SD denotes symmetric and descending. This type of MBF is relevant for many applications in finance, for instance, in cases in which we do not have a strong prior belief about whether a signal positively or negatively predicts returns.[17] The SD-MBF is given by $-\exp(1) \times p\text{-value} \times \ln(p\text{-value})$, is the natural exponent. The p-value is calculated under the null.[18]

What is the connection between the MBF and the SD-MBF? The MBF is the minimum across all possible prior specifications of alternative hypotheses, whereas the SD-MBF is the minimum across a specific class of prior specifications for the alternative hypothesis. Hence, the MBF is always smaller and presents stronger evidence against the null than the SD-MBF. This is intuitive as the MBF occurs when the entire prior probability mass on the alternative

[16] Alternatively, if the prior is tilted more toward the alternative (10/19 = (5/95)/(1/10)), a 5% posterior probability can be obtained.

[17] Other types of MBF are derived under alternative assumptions on the prior density for alternatives. See Berger and Sellke (1987) for examples of alternative types of MBF. In general, the prior density for alternatives that achieve a certain type of MBF depends on the data likelihood function under the null hypothesis. It does not need to have the form of well-known probability distributions.

[18] The SD-MBF is first proposed in Bayarii and Berger (1998, p. 81), who refer to this approach as "quick and dirty." Also see Goodman (2001) and Nuzzo (2014).

Table II

Using Minimum Bayes Factors to Calculate the Likelihood of the Null Hypothesis Being True (Bayesianized *P*-values)

This table reports the probability that the null is true. MBF is the minimum Bayes factor, exp(–Z^2/2); SD-MBF is the symmetric-descending minimum Bayes factor, –exp(1) × p-value × ln(p-value).

			Prior Beliefs (Odds, Null:Alternative)					
			Long Shot					Even Odds
	Usual		99-to-1	49-to-1	24-to-1	19-to-1	4-to-1	1-to-1
z-Score	P-value	MBF	0.01	0.02	0.04	0.05	0.20	0.50
			Panel A: MBF					
1.645	0.10	0.259	0.96	0.93	0.86	0.83	0.51	0.21
1.960	0.05	0.147	0.94	0.88	0.78	0.74	0.37	0.13
2.576	0.01	0.036	0.78	0.64	0.47	0.41	0.13	0.03
3.291	0.001	0.004	0.31	0.18	0.10	0.08	0.02	0.004
			Panel B: SD-MBF					
1.645	0.10	0.626	0.98	0.97	0.94	0.92	0.71	0.38
1.960	0.05	0.407	0.98	0.95	0.91	0.89	0.62	0.29
2.576	0.01	0.125	0.93	0.86	0.75	0.70	0.33	0.11
3.291	0.001	0.019	0.65	0.48	0.31	0.26	0.07	0.02

hypotheses concentrates on the maximum likelihood estimate of the data—so what we believe coincides with what we observe (via the maximum likelihood estimate), which implies the strongest evidence against the null. In contrast, if we do not have a good sense of what the alternatives should be, it may make sense to spread the prior probability mass on the alternative hypotheses across a wide range of values. Additionally, if we further restrict the prior density to be symmetric and descending around the null, then the SD-MBF, which is more lenient than the MBF (i.e., the MBF is more likely to reject the null), will be more informative when adjusting the prior odds ratio.

To summarize, we have a simple formula to transform the frequentist p-value and the prior probability of the null into a Bayesianized p-value:

$$\text{Bayesianized } p\text{-value} = \text{MBF} \times \frac{\text{prior odds}}{1 + \text{MBF} \times \text{prior odds}}.$$

This formula can be used for the MBF or the SD-MBF. Panel A of Table II presents several examples using four common p-values—0.001, 0.01, 0.05, and 0.10—for the MBF. Suppose you are testing an effect that you think is only about 20% likely to be true (prior odds are 4:1; notice odds = $p_0/(1 - p_0)$, where p_0 is the prior probability). Consider the frequentist p-value of 0.05, which would be significant under NHST assuming a 5% threshold. The MBF is 0.147. The formula above suggests that the Bayesianized p-value is 0.37 (= (0.147 × 4)/(1 + 0.147 × 4)). This means there is a 37% chance that the null hypothesis

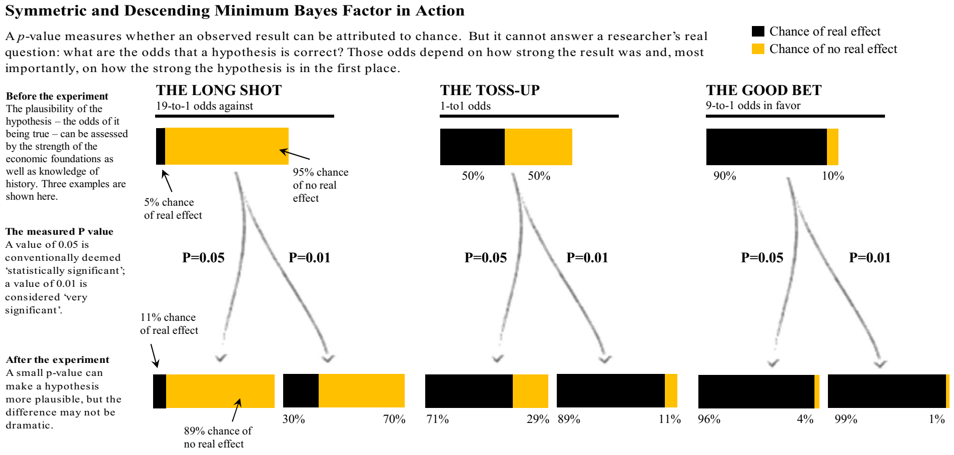


Figure 7. The SD-MBF Illustrated. Figure from Nuzzo (2014).

is true. If you instead think the prior odds are even (1:1), meaning you believe that there is a 50% chance that the effect is true, the Bayesianized p-value drops to 0.13. Notice that both of these values are higher than the usual p-value. I again emphasize the difference in interpretation. We can make *direct* statements about the viability of the null hypothesis under this framework. We cannot do so under the NHST framework because the classical p-value does not tell us the probability that the null hypothesis is true.

Panel B of Table II presents similar results for the SD-MBF. Continuing the example above, the Bayesianized p-value is 0.62 in the 4:1 odds case and 0.29 in the 1:1 odds case. The higher p-values are intuitive. The MBF gives the best possible shot to the alternative by concentrating the prior mass at the alternative while the SD-MBF concentrates the prior mass around the null hypothesis. Hence, the SD-MBF and its Bayesianized p-values are larger. So with even odds, there is a 29% chance the null is true. Again, the inference is much different than under the usual NHST.

The MBF p-value thus gives the smallest Bayesianized p-value among all possible priors that satisfy a given prior odds ratio. It is aggressive in the sense that it presents the strongest evidence against the null. A large MBF p-value implies a total lack of evidence to reject the null. A small MBF p-value implies possible rejection of the null and suggests that the researcher conduct further analysis. The SD-MBF is less aggressive than the MBF and is best used when we have a dispersed belief on plausible values under alternative hypotheses.

Figure 7 illustrates the SD-MBF in action. Notice that the evidence has the greatest impact on the probability of an effect when the prior odds are a toss-up. The values for the Long Shot and the Toss-Up cases can be found in Table II.

There is an additional way to illustrate the use of the MBF. One can set the level of confidence for the effect to be true, combine this with prior odds,

Table III

t-Statistic Thresholds for Minimum Bayes Factors

This table reports the threshold for the *t*-statistic corresponding to the probability that the null is true. MBF is the minimum Bayes factor, $\exp(-Z^2/2)$; SD-MBF is the symnmetric-descending minimum Bayes factor, $-\exp(1) \times p\text{-value} \times \ln(p\text{-value})$. In Panel A, we solve for the statistic *S*, $S = \sqrt{(-2 \times \ln(BPV/((1-BPV) \times PO))}$, where BPV is the Bayesianized *p*-value in column 1 and PO is the prior odds (e.g., 4:1 = 4). In Panel B, we first solve numerically for the *p*-value, $0 = p\text{-value} \times \ln(p\text{-value}) - (-\exp(-1) \times BPV/((1-BPV) \times PO))$, and then find the test statistic, *S*, corresponding to *p*-value/2.

	Prior Beliefs (Odds, Null:Alternative)					
	Long Shot					Even Odds
Prob. Null	99-to-1	49-to-1	24-to-1	19-to-1	4-to-1	1-to-1
Is True	0.01	0.02	0.04	0.05	0.20	0.50
			Panel A: MBF			
0.10	3.69	3.49	3.28	3.21	2.68	2.10
0.05	3.88	3.70	3.50	3.43	2.94	2.43
0.01	4.29	4.12	3.94	3.88	3.46	3.03
0.001	4.80	4.65	4.49	4.44	4.07	3.72
			Panel B: SD-MBF			
0.10	4.10	3.92	3.72	3.65	3.16	2.63
0.05	4.29	4.11	3.93	3.86	3.41	2.93
0.01	4.67	4.51	4.35	4.29	3.89	3.49
0.001	5.16	5.02	4.87	4.82	4.47	4.13

and calculate the test statistic threshold necessary to match that confidence. I present such an example in Table III. Panel A reports the results for the MBF. Suppose you think a variable *X* has a 50-50 chance of being the true explanator of variable *Y*. In addition, you want the level of confidence to be such that there is only a 5% probability that the null is true. The *t*-statistic threshold is 2.43. This threshold is a higher hurdle than the usual *t*-statistic of 2.0. However, again, we are answering a different question here. As the variable becomes less likely in terms of prior odds, the threshold increases. For example, if you think the odds are 4:1 against *X* being the true explanatory variable, the threshold increases to 2.94. Panel B reports the equivalent thresholds for the SD-MBF. These thresholds are higher, which makes sense given that the MBF gives the alternative a better chance than the SD-MBF. For example, with even odds and a 5% probability that the null is true, the SD-MBF threshold is 2.93, which is higher than the 2.43 in Panel A.

Let us now return to the regression prediction example with a slope that has a *t*-statistic of 2.6 and a *p*-value of 0.014. The SD-MBF is 0.162. If the odds are even in terms of the null versus alternative, then the probability that the null is true is only 0.14.[19] Alternatively, if we want to enforce a 5% probability that

[19] SD-MBF = $-\exp(1) \times p\text{-value} \times \ln(p\text{-value}) = 0.162$; probability of null with even odds = $0.162/(0.162 + 1) = 0.14$.

the null is true, we need $t > 2.93$ according to Table III. What assumptions are necessary here? To interpret the SD-MBF for the regression example, we first need to assume that the prior density on alternative values for the slope is symmetric and descending around zero. Under these assumptions, the SD-MBF is the minimum Bayes factor. In other words, in this regression setup, if you believe as a prior that the null (slope $= 0$) is more likely than other values and that a negative slope has the same likelihood as a positive slope with the same magnitude, then the SD-MBF is the minimum Bayes factor among all Bayes factors that can obtain under different prior probability densities for alternatives.[20]

Let us next apply the MBF and SD-MBF to the music, tea, and bar patron examples. All three experiments result in a traditional p-value close to 0.001. In the case of the music test, you already know that the test subject is a musicologist, so a 50% probability (1:1 odds) of the individual not being a musicologist (the null hypothesis) should be an overstatement of your prior. Even under this prior, you have strong evidence (the probability of the null being true is only 0.004 and 0.02 for the MBF and the SD-MBF, respectively) and you can confirm she is an expert. At the other extreme, in the case of the bar patron who claims to predict the future, a 99% probability (99:1 odds) of the patron making a false claim (the null hypothesis) should be an understatement of your prior. It is therefore not surprising to see that, even after 10 successful guesses in a row, the evidence for the alternative is still weak—the probability of the null being true (no clairvoyance) ranges from 0.33 to 0.66.

At the risk of being redundant, notice that our language has changed. Under the classical framework, the p-value is the probability of observing the outcome or a more extreme outcome if the null hypothesis is true. Under the MBF framework, we are able to talk about the probability that the null is true. We are also able to talk about the probability that the alternative is true.

In many cases, the MBF is a nonlinear transformation of the p-value. So does it simply report the p-value on a different scale? In a sense yes because no information besides the p-value is needed to calculate the MBF. However, it is not merely a transformation because it allows us to go from deductive reasoning as under the NP framework to inductive reasoning as under the full Bayesian approach. We are now able to make statements about the odds of the null hypothesis relative to alternative hypotheses. One key element of such statements—the MBF—is independent of the prior specification and

[20] My analysis assumes two-sided tests. This is the assumption under the Fisher framework when we do not specify the alternatives. When we deviate from this assumption, that is, when we employ one-sided tests, we are deviating from the Fisher framework by essentially incorporating prior information (although vague) on alternatives. This changes the calculation of p-values from two-sided to one-sided. This also changes the MBF and the SD-MBF as follows: if the data maximum likelihood estimate is inconsistent with your prior on the alternative (e.g., the null is $\mu = 0$ and you are testing $\mu = 0$ versus $\mu > 0$, in which case a negative data maximum likelihood estimate would be inconsistent with your prior on the alternative), then both the MBF and the SD-MBF should be exactly one, that is, the data do not provide any information to help distinguish the null from the alternative. Otherwise, the MBF and the SD-MBF are the same as in the two-sided case.

hence should carry the same meaning to all researchers. Thus, the MBF can be thought of as the bridge between the p-value, which is not related to the probabilities of different hypotheses, and something that is related to these probabilities.

Some may counter that the MBF is too subjective, since one still needs to specify a prior to calculate the posterior probabilities. However, as emphasized by Fisher (1925), there is no "objective" evaluation of hypotheses. Researchers in financial economics seem to ignore this fact and draw on p-values to make probabilistic assessments of different hypotheses. This leads to misleading interpretations of results and may confound the development of research. My goal here in suggesting that we include MBFs among our test statistics is to encourage researchers to first admit the necessity of subjective probabilities, to next specify a prior or set of priors, and to ultimately answer the question that we are really interested in: what is the probability that the null is true?

So far most of the applications of the MBF above have been illustrative in nature. I now turn to some real examples.

Let us first revisit the cancer example. For breast cancer diagnosis, the application is greatly simplified as there is one possible outcome under the alternative hypothesis, namely, the patient has breast cancer. As a result, we do not need to search among a large number of specifications for alternatives as in a typical MBF calculation. In this case, the MBF collapses to the usual Bayes factor, which equals the ratio between the probability of making a false diagnosis and the probability of making a correct diagnosis, or the error rate (α) divided by test power (β). Given a prior odds ratio, we can use the Bayes factor to derive the posterior likelihoods of both the null and the alternative hypothesis.[21] It is therefore straightforward to apply the MBF method to the breast cancer example, where there is only one possible value for the alternative hypothesis.

Notice how test power enters the Bayes factor and, as a special case of the Bayes factor, the MBF. Bayesian methods for hypothesis testing allow one to weigh the likelihood of the data under the null against the likelihood under the alternative and in doing so take test power into account. This is in contrast to the frequentist approach, where test power plays a less prominent role than the p-value and hence results in confusion when a large-sample result is presented as significant even when economically trivial.

Let us next consider the application of the MBF and the SD-MBF to four published studies. In Table IV, I divide the studies' findings into three categories based on prior beliefs: (1) "A stretch," meaning that the prior probability of the effect being true is 2%, (2) "Perhaps," meaning that the prior probability is in the 20% range, and (3) "Solid footing," meaning that the prior probability has even odds, that is, is 50%.

[21] Let π be the prior likelihood of the alternative. The prior odds ratio is $(1 - \pi)/\pi$. Multiplying the prior odds ratio by the Bayes factor α/β, the posterior odds ratio is $\frac{1-\pi}{\pi}\frac{\alpha}{\beta}$, which implies a probability of $\frac{\frac{1-\pi}{\pi}\frac{\alpha}{\beta}}{\frac{1-\pi}{\pi}\frac{\alpha}{\beta}+1} = \frac{\alpha}{\frac{\pi}{1-\pi}\beta+\alpha}$ for the null. This is exactly the formula for the expected fraction of false discoveries that I presented earlier.

Table IV

Converting Published Classical *P*-values into Bayesianized *P*-values

This table reexamines the t-statistics and p-value presented in four academic papers. The table reports both MBF and SD-MBFs based on the reported p-values, where MBF = $\exp(-Z^2/2)$ and SD-MBF = $-\exp(1) \times p\text{-value} \times \ln(p\text{-value})$; prior odds ratios; as well as Bayesianized p-values that tell the probability that the null (factor not priced) is true, given the data.

								Bayesianized *P*-values	
Prior Category	Effect	Sample	Reported t-stat	Reported *P*-values	MBF	SD-MBF	Prior Odds Ratio (Null/Alternative)	(MBF)	(SD-MBF)
A stretch	Clever tickers outperform (Head, Smith, and Watson (2009))	1984 to 2005	2.66	0.0079	0.0291	0.1040	49/1	0.588	0.836
Perhaps	Profitability priced (Fama and French (2015))	1963 to 2013	2.92	0.0035	0.0141	0.0538	4/1	0.053	0.117
	Size priced (Fama and French (1992))	1963 to 1990	2.58	0.0099	0.0359	0.1242	4/1	0.125	0.332
Solid footing	Market beta priced (Fama and MacBeth (1973))	1935 to 1968	2.57	0.0100	0.0368	0.1252	1/1	0.035	0.111

The first study in Table IV—the ticker symbol study that I mention in the introduction of this paper—belongs to the "A stretch" category. Most of us would believe that stocks with clever ticker symbols are unlikely to outperform. Hence, the reported p-value in the original study, 0.0079, gives us little information about the likelihood of the effect. Calculating the two MBFs and combining them with the prior belief that the effect is a long shot, the probability that the null (i.e., no effect) is true ranges from 58.8% to 83.6%. The takeaway here is completely different from the standard inference.

Turning to a few findings in the "Perhaps" category, recent work suggests that firm profitability is priced in the cross-section of expected returns. The reported p-value is very low, at 0.0035. However, applying prior information potentially changes the inference. With a 20% prior probability that the effect is true, the SD-MBF suggests that there is a 11.7% probability that the null (i.e., no effect) is true. Similarly, with respect to the size factor, with a 20% prior probability that the effect is true, the SD-MBF suggests that there is a 33.2% probability that the null is true.

Finally, consider results on market beta, which belong to the "Solid footing" category. The reported p-value is 0.01. Under even odds (50% probability that the effect is true), the SD-MBF suggests that there is only an 11.1% probability that the null is true.

There is a simple message here: we cannot simply report the usual p-values. Given the rare effects problem, we somehow need to take prior beliefs into account. Tables III and IV demonstrate how easy it is to employ a Bayesianized approach without having to resort to the computationally challenging full-Bayesian framework.

VIII. Extensions and Limitations

The MBF approach detailed above crucially depends on the frequentist p-value and therefore suffers from many of the same issues that plague the p-value itself. One such issue is that a p-value usually does not tell us how well a model explains the data. For example, an investment factor that has a large t-statistic under cross-sectional tests may not help explain the cross-section of expected returns, that is, there may be a discrepancy between statistical significance and model adequacy. Given the discussion above, this discrepancy is not hard to understand. A classical p-value measures the degree of incompatibility between observed data and the null hypothesis, but cannot tell us what would happen under alternative hypotheses. It is possible that both the null and the alternative hypothesis have poor explanatory power vis-a-vis the data.

In contrast, a full-blown Bayesian approach depends less on the frequentist p-value as it explicitly compares the evidence (i.e., likelihood) under the alternative hypotheses with the evidence under the null hypothesis. Even if we have a small frequentist p-value, which implies a large discrepancy between the model and the data under the null hypothesis, we may still fail to "reject the null" under a full-blown Bayesian approach as this discrepancy could be equally large under alternative hypotheses. In other words, compared to the full-blown Bayesian approach, we may “reject the null” too often under the MBF. This is consistent with the notion that the MBF might be too aggressive, where “aggressive” implies more rejections of the null. However, while the MBF might be too aggressive, it still evaluates the data under alternative hypotheses. Moreover, the usual p-value under NHST is far more aggressive and leads to unacceptable rates of false discoveries.

Compared to the MBF, the SD-MBF is less aggressive in that the maximum likelihood for the alternative hypotheses (i.e., the denominator in the Bayes factor) is calculated under a more constrained distributional family for the prior probability density of the alternative hypotheses, which generates a smaller likelihood under alternatives and hence a larger Bayes factor. As a result, it presents weaker evidence against the null compared to the MBF. In addition, the features of the prior density on the alternative hypotheses that the SD-MBF assumes appear to be appropriate for many hypothesis testing scenarios. We therefore expect SD-MBF to be more useful in empirical applications. Note that if we have more information about the prior probability density of the alternative hypotheses, we can incorporate this information into the calculation of the Bayes factor to derive alternative MBFs that are even more informative than the MBF and SD-MBF.[22]

There are limitations to using MBFs. Suppose we require a level of confidence such that the probability of the null being true must be 5% or less. We conduct an experiment and with even odds we find a Bayesianized p-value of 0.09. Here we fail to “reject the null” hypothesis of no effect because there is a 9% chance that the null is true. However, what happens if the Bayesianized p-value is 0.02 using the MBF? We already know that the MBF gives the alternative the best possible chance. Suppose that the SD-MBF in this case is 0.045. Are we done? That is, can we now confidently “reject the null?” Again, the results are not definitive as there is a lot of simplification going on. This ambiguity is the price that we have to pay for not explicitly specifying the alternative.

Based on the above discussion, I argue that the MBF and SD-MBF be used in an initial step to screen out those effects that are highly unlikely to be true.[23] Assuming that a hypothesized effect passes this first hurdle with a very small Bayesianized p-value, one might next go on to explicitly specify an alternative. I do not expect many researchers will do this due to the challenges of adopting the full Bayesian framework. Fortunately, Bayarri et al. (2016) show that over a wide number of experiments, the SD-MBF is very close to the full Bayes factor when p-values are low.

[22] See Berger and Sellke (1987) and Bayarri et al. (2016).

[23] This is not a trivial step. Many results in financial economics are not able to pass the hurdle using the MBF or SD-MBF once prior odds are taken into account, even if we give the alternative the best possible chance by using MBFs.

IX. Recommendations

This address is not meant to criticize the quality of the research that we have done in the past. But I believe it is essential for the future scientific outlook of financial economics that we embrace some of the guidelines developed in other fields.

Below I offer specific recommendations. These recommendations fall into two categories. The first focuses on how we conduct our research. The second focuses on the incentives in the publication process that impact the research culture in our field.

A. Research Methods[24]

Building on the various points made in this address, below I list recommendations for future work in financial economics.

- *Before looking at the data,* establish a research framework that includes: economic foundation, hypotheses to be tested, data collection methods (e.g., sample period, data exclusions, manipulations, and filters), tests to be executed, statistical methods, plan for reporting results, and robustness checks. The research framework should be transparent to all researchers. There is a growing trend in other sciences to post all these choices online before the experiment is conducted or the data are collected. See Open Science Foundation http://osf.io.
- Employ checklists. Consolidated Standards of Reporting Trials (CONSORT 2010) provides a detailed checklist for researchers in medicine doing clinical trials. Each of the top journals in this field has endorsed these guidelines.
- Recognize that any investigation is unlikely to yield a zero-one (false-true) outcome. Employ statistical tools that take this fact into account.
- Employ statistical methods that control for known research problems like multiple testing and inflation of effect sizes.
- Focus on both the magnitude and the sign of an effect, not just on the level of significance.
- Share nonproprietary data and code in an easily accessible way.
- Report all results, not just a selection of significant results. Internet Appendices make this easy to do.
- If the results relate to one topic, do not carve them up across separate papers—put everything into one paper if feasible.
- If a result is surprising, try to replicate the result perhaps on a new data set before submitting for publication.
- Strengthen evidence by conducting additional tests of the main hypothesis, provided these additional tests are not highly correlated with the initial test.

[24] Some of these recommendations come from Ioannidis (2008).

- Take priors into account. An effect that results from a theoretical prediction should be treated differently from an effect that arises from data mining, that is, economic plausibility must be part of the inference. Of course, care should be taken to avoid theory-hacking, where various assumptions are tried to fit the model to a known empirical phenomenon. Theories are more convincing if they have a new testable implication (in addition to explaining known patterns in the data). This is analogous to an out-of-sample test in empirical work.
- Alongside the usual test statistics, present Bayesianized *p*-values. You might report these measures, which give the probability that the null is true, with more than one prior so readers can make their own assessment as to the plausibility of the hypothesis given the test results.

B. The Agency Problem

Many of the problems that I describe above are endogenous in that agents are simply responding to incentives. At most schools, a single publication in a top journal like the *Journal of Finance* can lead to tenure. Even at many top schools, administrators track the number of publications and rewards are often tied to productivity. Researchers know that journals tend to publish papers with positive results. Editors know that papers with positive results get more citations. Many editors focus on the impact factor of their journal as a measure of success.

As I mention in the introduction, the above factors lead to a complex agency problem. Our goal is to advance scientific knowledge in our field. In doing so, we should care about discoveries—both positive and negative—that are repeatable, that is, discoveries for which the initial findings can be validated both in replication and in out-of-sample tests. In addition, we should not shy away from risky projects that could lead to a substantial leap in knowledge. However, the incentives of editors, peer reviewers,[25] and authors are not necessarily compatible with this goal.

While the main focus of HLZ (2016) is empirical methods, I fear that an unintended consequence of the paper's message is increased data mining in an effort to meet higher thresholds for significance. Adaptation of the Bayesian approach should mitigate this problem because many purely data-mined results suffer from a lack of economic foundation. As such, they are not likely to pass the hurdle if prior beliefs are incorporated.

I believe that another consequence of the path that we have been on in our field is a proliferation of papers that are technically well executed but that advance our knowledge only marginally. Suppose a researcher is choosing between two projects. One involves hand-collecting data over the course of a year while the other involves a standard data source like Compustat. The researcher has even odds on the main hypothesis in both projects.

[25] The peer review process is also part of the problem. See the discussion in Berk, Harvey, and Hirshleifer (2017).

The researcher also knows that it will be difficult to publish either of these papers if the result is negative. As a result, and this is especially true for younger researchers, the less risky project (i.e., the one not requiring time-consuming data collection) is often selected. However, it might be the case that the riskier project is the one that would most advance our knowledge.

I do not have a solution for this agency problem. It impacts not just our field but rather all scientific fields (though some more than others). However, I do have some ideas of steps we might take to mitigate the problem.

To start, we need to select editors that encourage more risk-taking. This means publishing papers that ask interesting economic questions even if the result is negative. There are two motivations here. First, a paper with a Bayesianized p-value of say 0.30 might be addressing an important economic question that provides insights no matter what the result is. Second, the effect might be true. To reduce Type II errors, the editor needs to be willing to take some risk. Indeed, as an extreme example, no editor wants to reject the next Black-Scholes paper.

Moreover, editors should not simply chase impact factors. It is well-known that there are fixed effects across different areas of financial economics (as well as other sciences). This should not lead editors to shun papers in certain areas. The individuals tasked with selecting journal editors should ensure that the prospective editor has a long-term view for the journal.

In addition, top national conferences and journals in financial economics should follow the lead of other fields and accept "registered reports."[26] These reports are detailed proposals that involve, say, the collection of new data or a plan to do research in an unexplored field. Such a proposal is like the front end of the usual paper as it provides economic motivation as well as a plan detailing the methods to be employed once the data are collected. The proposal is peer reviewed, with the editor making a decision as to whether the research question is interesting enough to publish—even if the result is negative. It is understood that the researcher will inevitably need to make certain choices when confronted with issues not anticipated in the original proposal. These choices must be made transparent in the final paper. Registered reports effectively reduce researchers' downside risk, increasing their incentives to undertake otherwise risky projects.

There should also be greater costs associated with p-hacking. In our field, very few replication studies are published, for at least three reasons. First, there is a perception that, given most of the data used in our field are readily available, there is no need to replicate. Second, replication studies involve time-consuming editorial back-and-forth with the original authors but are unlikely to lead to many citations, which reduces editorial support for such studies. Third, replication studies are not treated like regular papers in promotion and tenure decisions, which decrease researchers' incentives to conduct these studies. Additionally, the cost of replication needs to decrease. Currently, only the

[26] The 2017 *Journal of Accounting Research* conference implemented this idea.

Journal of Finance requires that computer code be submitted when a paper is accepted, and none of the top journals in finance require data. However, a relatively new journal, *Critical Finance Review*, has made considerable progress in making replication more mainstream.

If the top journals in our field routinely publish papers with negative results as well as replication studies, the number of papers published in these journals will necessarily increase.[27] Such a step is controversial but increasingly feasible as print journals have only a few years to go before being completely replaced by digital versions. The fact that most journals have gone to a multiple-editor system should also help in this regard.

X. Concluding Remarks

The scientific outlook of financial economics crucially depends on our research and publication culture. As Hubble (1954, p. 41) argues,

> The scientist explores the world of phenomena by successive approximations. He knows that his data are not precise and that his theories must always be tested. It is quite natural that he tends to develop a healthy skepticism, a suspended judgment, and a disciplined imagination.

This means that, as a researcher, you must be skeptical of both conventional wisdom and your own beliefs.[28] Importantly, you are not a sales person—you are a scientist. Yet how many of our papers include language such as "this result is encouraging?"[29] Such language is unscientific and reveals to the reader that the researcher has a specific agenda to find evidence in support of his or her hypothesis. This is the type of culture that motivates my address.

The title of my address is borrowed from Bertrand Russell's (1931) famous treatise, *The Scientific Outlook,* which I believe should be required reading for all Ph.D. students in financial economics or other sciences. In that work, Russell argues for caution in the presentation of results:

> Who ever heard of a theologian prefacing his creed, or a politician concluding his speeches, with a statement as to the probable error in his opinions? It is an odd fact that subjective certainty is inversely proportional to objective certainty. The less right a man has to suppose himself

[27] *PNAS* publishes more than 3,100 peer-reviewed papers each year in weekly issues. *PLoS ONE* published 28,107 peer-reviewed papers in 2015—more papers on any weekday than the *Journal of Finance* publishes in one year (in 2015, the *Journal of Finance* published 70 articles). Obviously, the number of potential contributors is greater for *PLoS ONE*.

[28] See also Gawande (2016).

[29] I checked. In particular, I searched the top three journals in finance over the period 2000 to 2015 for the word "encourag*." I then examined the context of each instance identified and isolated the papers that use phrases like "encouraging results." I found 29 instances of such phrases in the *Journal of Finance,* 22 in the *Journal of Financial Economics*, and 19 in the *Review of Financial Studies*.

> in the right, the more vehemently he asserts there is no doubt whatever that he is exactly right . . . No man who has the scientific temper asserts that what is now believed in science is exactly right; he asserts that it is a stage on the road towards the exact truth (pp. 64–65).

We are living in an era of increased mistrust about scientific discovery (Gauchat (2012)). Some of this mistrust is well founded. It is hard to interpret the results of pharmaceutical clinical trials, for instance, when those that do not work out are not made public.[30] I previously mentioned the exaggerated published result that couples that meet online are happier. To add to that story, the data were provided by eHarmony, where one of the authors was an employee. In our own field, credibility suffered after the global financial crisis when the documentary *Inside Job* suggested that sponsored research potentially misled investors.[31]

It is hard to undo such damage. Whether it is the widely discredited study that claimed power lines cause cancer or the fraudulent study in *The Lancet* that argued there is a link between the preservative in vaccines and autism, it is difficult to debunk bad science. Indeed, recent work suggests that efforts to debunk bad studies may have the opposite effect (Cook and Lewandowsky (2012)), actually increasing beliefs in the false results.

Most of us would agree that, today, the quality of research in financial economics is high. However, my address is not about today—it is about tomorrow. To avoid stumbling down the same path as some other fields, it is essential that we build and maintain a robust research culture. While there is work to do, the scientific outlook of financial economics is very promising. My modest goal here is to start a conversation about some of the issues that I have raised in this address.

Editor: Stefan Nagel

REFERENCES

Alter, Adam L., and Daniel M. Oppenheimer, 2006, Predicting short-term stock fluctuations by using processing fluency, *Proceedings of the National Academy of Sciences* 103, 9369–9372.

American Finance Association, 2016, Code of Professional Conduct and Ethics. Available at http://www.afajof.org.

American Statistical Association, 2016, Statement on Statistical Significance and P-Values. Available at: https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf.

Bayarri, M. J., Daniel J. Benjamin, James O. Berger, and Thomas M. Sellke, 2016, Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses, *Journal of Mathematical Psychology* 72, 90–103.

Bayarri, M. J., and James O. Berger, 1998, Quantifying surprise in the data and model verification, in José M. Bernardo, James O. Berger, A. Philip Dawid, and Adrian F. M. Smith, eds.: *Bayesian Statistics*. Vol. 6. (Oxford University Press, Oxford).

[30] Organizations such as http://alltrials.net are trying to force companies to reveal results of all scientific trials.

[31] The study in question is titled "Financial Stability in Iceland."

Bem, Daryl J., 2011, Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect, *Journal of Personality and Social Psychology* 100, 407–425.

Berger, James O., and Thomas Sellke, 1987, Testing a point null hypothesis: The irreconcilability of *p* values and evidence, *Journal of the American Statistical Association* 82, 112–122.

Berk, Jonathan, Campbell R. Harvey, and David A. Hirshleifer, 2017, How to write an effective referee report and improve the scientific review process, *Journal of Economic Perspectives* 31, 231–244.

Berkson, Joseph, 1938, Some difficulties of interpretation encountered in the application of the chi-square test, *Journal of the American Statistical Association* 33, 526–536.

Berkson, Joseph, 1942, Tests of significance considered as evidence, *Journal of the American Statistical Association* 37, 325–335.

Cacioppo, John T., Stephanie Cacioppo, Gian C. Gonzaga, Elizabeth L. Ogburn, and Tyler J. VanderWeele, 2013, Marital satisfaction and break-ups differ across on-line and off-line meeting venues, *Proceedings of the National Academy of Sciences* 110, 10135–10140.

Carver, Ronald P., 1978, The case against statistical significance testing, *Harvard Educational Review* 48, 378–399.

CERN, 2017, LHC: The guide, Conseil Européen pour la Recherche Nucléaire, January. Available at http://cds.cern.ch/record/2255762/files/CERN-Brochure-2017-002-Eng.pdf.

Churchill, Gary A., 2014, When are results too good to be true? *Genetics* 198, 447–448.

Cohen, Jacob, 1994, The earth is round (p<.05), *American Psychologist* 49, 997–1003.

Coles, Jeffrey, Naveen Daniel, and Lalitha Naveen, 2008, Boards: Does one size fit all? *Journal of Financial Economics* 87, 329–356.

Comte, Auguste, 1856, *The Positive Philosophy of Auguste Comte*, translated by Harriett Marineau (Calvin Blanchard, New York). Vol. II. Available at: https://doi.org/10.1017/CBO9780511701467.

Consolidated Standards of Reporting Trials, 2010. Available at http://www.consort-statement.org/consort-2010.

Cook, John, and Stephan Lewandowsky, 2012, *The Debunking Handbook* (University of Queensland, St. Lucia, Australia). Available at: http://www.skepticalscience.com/docs/Debunking_Handbook.pdf.

Edwards, Ward, Harold Lindman, and Leonard J. Savage, 1963, Bayesian statistical inference for psychological research, *Psychological Review* 70, 193–242.

Fama, Eugene F., and Kenneth R. French, 1992, The cross-section of expected stock returns, *Journal of Finance* 47, 427–465.

Fama, Eugene F., and Kenneth R. French, 2015, A five-factor asset pricing model, *Journal of Financial Economics* 116, 1–22.

Fama, Eugene F., and James D. MacBeth, 1973, Risk, return, and equilibrium: Empirical tests, *Journal of Political Economy* 81, 607–636.

Fanelli, Daniele, 2010, "Positive" results increase down the Hierarchy of the Sciences, *PLoS ONE* 5, e10068.

Fanelli, Daniele, 2012, Negative results are disappearing from most disciplines and countries, *Scientometrics* 90, 891–904.

Fanelli, Daniele, 2013, Positive results receive more citations, but only in some disciplines, *Scientometrics* 94, 701–709.

Fisher, Ronald A., 1925, *Statistical Methods for Research Workers* (Oliver and Boyd Ltd., Edinburgh).

Francis, G., 2012, Too good to be true: Publication bias in two prominent studies from experimental psychology, *Psychonomic Bulletin and Review* 19, 151–156.

Gauchat, Gordon, 2012, Politicization of science in the public sphere: A study of public trust in the United States, 1974–2010, *American Sociological Review* 77, 167–187.

Gawande, Atul, 2016, The mistrust of science, Commencement speech, California Institute of Technology, reprinted in *The New Yorker*. Available at http://www.newyorker.com/news/news-desk/the-mistrust-of-science.

Gelman, Andrew, and Eric Loken, 2013, The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time, Working paper, Columbia University. Available at: http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.

Gelman, Andrew, and Eric Loken, 2014, The statistical crisis in science, *American Scientist* 102, 460–465.

Gelman, Andrew, and Hal Stern, 2006, The difference between "significant" and "not significant" is not itself statistically significant, *The American Statistician* 60, 328–331.

Gigerenzer, Gerd, 2004, Mindless statistics, *Journal of Socio-Economics* 33, 587–606.

Goodman, Steven N., 1993, P values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate, *American Journal of Epidemiology* 137, 485–496.

Goodman, Steven N., 1999, Toward evidence-based medical statistics. 1: The *p*-value fallacy, *Annals of Internal Medicine* 130, 995–1004.

Goodman, Steven N., 2001, Of *p*-values and Bayes: A modest proposal, *Epidemiology* 12, 295–297.

Greenhouse, J. B., 2012, On becoming a Bayesian: Early correspondences between J. Cornfield and L. J. Savage, *Statistics in Medicine* 31, 2782–2790.

Harvey, Campbell R., Yan Liu, and Heqing Zhu, 2016, . . . and the cross-section of expected returns, *Review of Financial Studies* 29, 5–68.

Head, Alex, Gary Smith, and Julia Watson, 2009, Would a stock by any other ticker smell as sweet? *Quarterly Review of Economics and Finance* 49, 551–561.

Head, Megan L., Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions, 2015, The extent and consequences of *p*-hacking in science. *PLoS Biol* 13, e1002106.

Hubble, Edwin, 1954, *The Nature of Science and Other Lectures* (Huntington Library, San Marino, California).

Hutton, Jane, 2010, Misleading statistics: The problems surrounding number needed to treat and number needed to harm, *Pharmaceutical Medicine* 24, 145–149.

Ioannidis, John P. A., 2005, Why most published research findings are false, *PLoS Medicine* 2, e124.

Ioannidis, John P. A., 2008, Why most discovered true associations are inflated (with discussion), *Epidemiology* 19, 640–658.

Jarosz, Andrew F., and Jennifer Wiley, 2014, What are the odds? A practical guide to computing and reporting Bayes factors, *Journal of Problem Solving* 7, 2–9.

Kerr, Norbert L., 1998, HARKing: Hypothesizing after the results are known, *Personality and Social Psychology Review* 2, 196–217.

Lang, Janet M., Kenneth J. Rothman, and Cristina I. Cann, 1998, That confounded *p*-value (Editorial), *Epidemiology* 9, 7–8.

Lintner, John, 1965, The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets, *Review of Economics and Statistics* 47, 13–37.

Marsh, David M., and Teresa J. Hanlon, 2007, Seeing what we want to see: Confirmation bias in animal behavior research, *Ethology* 113, 1089–1098.

Mossin, Jan, 1966, Equilibrium in a capital asset market, *Econometrica* 34, 768–783.

Newey, Whitney K., and Kenneth D. West, 1987, A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica* 55, 703–708.

Neyman, Jerzy, and Egon S. Pearson, 1933, On the problem of the most efficient tests of statistical hypotheses, *Philosophical Transactions of the Royal Society* Series A 231, 289–337.

Nosek, Brian A., Jeffrey R. Spies, and Matt Motyl, 2012, Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability, *Perspectives on Psychological Science* 7, 615–631.

Nuzzo, Regina, 2014, Statistical errors, *Nature* 506, 150–152.

Ritchie, Stuart J., Richard Wiseman, and Christopher C. French, 2012, Failing the future: Three unsuccessful attempts to replicate Bem's 'retroactive facilitation of recall' effect, *PLoS ONE* 7, e33423.

Rosenthal, Robert, 1979, The "file drawer problem" and tolerance for null results, *Psychological Bulletin* 86, 638–641.

Royal Swedish Academy of Sciences, 2013, Press release, October 8. Available at http://www.nobelprize.org/nobel_prizes/physics/laureates/2013/press.html.

Russell, Bertrand, 1931, *The Scientific Outlook* (The Free Press, Glencoe, Illinois).

Sever, Peter S., Björn Dahlöf, Neil R. Poulter, Hans Wedel, Gareth Beevers, Mark Caulfield, Rory Collins, Sverre E. Kjeldsen, Arni Kristinsson, Gordon T. McInnes, Jesper Mehlsen, Markku Nieminen, Eoin O'Brien, and Jan Östergren, 2003, Prevention of coronary and stroke events with atorvastatin in hypertensive patients who have average or lower-than-average cholesterol concentrations, in the Anglo-Scandinavian Cardiac Outcomes Trial—Lipid Lowering Arm (ASCOT-LLA): A multicentre randomised controlled trial, *The Lancet* 361, 1149–1158.

Sharpe, William F., 1964, Capital asset prices: A theory of market equilibrium under conditions of risk, *Journal of Finance* 19, 425–442.

Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons, 2014, *P*-curve: A key to the file-drawer, *Journal of Experimental Psychology: General* 143, 534–547.

Song, Fujian, A. J. Eastwood, Simon Gilbody, Lelia Duley, and A. J. Sutton, 2000, Publication and related biases, *Health Technology Assessment* 4, 1–115.

Srinivasan, Raji, and Nita Umashankar, 2014, There's something in a name: Value relevance of congruent ticker symbols, *Consumer Needs and Solutions* 1, 241–252.

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

Wasserstein, Ronald L., and Nicole A. Lazar, 2016, The ASA's statement on *p*-values: Context, process, and purpose, *The American Statistician* 70, 129–133.

Welch, Ivo, and Amit Goyal, 2008, A comprehensive look at the empirical performance of equity premium prediction, *Review of Financial Studies* 21, 1455–1508.

Yermack, David, 1996, Higher market valuation of companies with a small board of directors, *Journal of Financial Economics* 40, 185–212.