econometrics

# Towards a New Paradigm for Statistical Evidence

Edited by

Jae H. (Paul) Kim and Muhammad Ishaq Bhatti

Printed Edition of the Special Issue Published in *Econometrcis*

www.mdpi.com/journal/econometrics

MDPI

# Towards a New Paradigm for Statistical Evidence

# Towards a New Paradigm for Statistical Evidence

Editors

**Jae H. (Paul) Kim**
**Muhammad Ishaq Bhatti**

MDPI • Basel • Beijing • Wuhan • Barcelona • Belgrade • Manchester • Tokyo • Cluj • Tianjin

*Editors*

Jae H. (Paul) Kim
Department of Economics and
Finance, La Trobe Business
School, La Trobe University
Australia

Muhammad Ishaq Bhatti
Department of Economics and
Finance, La Trobe Business
School, La Trobe University
Australia

*Editorial Office*
MDPI
St. Alban-Anlage 66
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *Econometrics* (ISSN 2225-1146) (available at: https://www.mdpi.com/journal/econometrics/special_issues/p_value).

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

LastName, A.A.; LastName, B.B.; LastName, C.C. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

**ISBN 978-3-0365-0882-5 (Hbk)**
**ISBN 978-3-0365-0883-2 (PDF)**

# Contents

# About the Editors

**Jae H. (Paul) Kim** has published widely in the areas of empirical finance, econometrics, and time series forecasting. His current research areas in finance include return predictability, testing for market efficiency, and methodological issues in statistical inference. He is an author of four software packages written in R: two for time series analysis and forecasting and the other two testing for asset market efficiency and return predictability.

The wild bootstrap variance ratio test he has proposed is available in Eviews, accessible to a mass of students and researchers around the world. After completing his PhD at the University of Sydney in 1997, he has worked at James Cook University, La Trobe University, and Monash University, before returning to La Trobe in June 2009. His recent papers are discussed in an article written by an industry commentator with a title ”The trend that is ruining finance research”: https://www.advisorperspectives.com/articles/2017/09/04/the-trend-that-is-ruining-finance-research?channel=Financial%20Planning.

**Muhammad Ishaq Bhatti** is a Professor and the founding director of the Islamic Banking and Finance Programme at Latrobe University (LTU). Currently, he is a graduate research coordinator for the department of Economics, Finance & Marketing. Previously, he has taught at Monash, Griffith, the International Islamic University, the University of Alberta and he has visited Rider, Magberg, Hitotsubashi, Auckland and Middle Eastern Universities.

He is author of more than 120 articles, 7 books, and a member of the editorial board of various journals, including the *European Journal of Finance*. His major areas of research, scholarship and teaching are in Business and data analytics, Quantitative and Islamic finance, Econometrics and Statistics.

He was a member of the team who won HEC of Pakistan, Turkey Central Bank Training 2013, Islamic Development Bank, and the Australian Research Council Discovery Grant jointly with Suren Basov and Saudi capital Market—Islamic Mutual Fund project with Naseem Al Rahahleh.

He is currently editing Routledge’s ’Islamic Business and Finance’ book series detailed in the link below: https://www.routledge.com/Islamic-Business-and-Finance-Series/book-series/ISLAMICFINANCE.

econometrics

MDPI

*Editorial*

# Towards a New Paradigm for Statistical Evidence in the Use of *p*-Value

**Muhammad Ishaq Bhatti * and Jae H. Kim ***

La Trobe Business School, La Trobe University, Melbourne, VIC 3086, Australia
* Correspondence: i.bhatti@latrobe.edu.au (M.I.B.); J.Kim@latrobe.edu.au (J.H.K.)

As the guest editors of this Special Issue, we feel proud and grateful to write the editorial note of this issue, which consists of **seven** high-quality research papers. We are incredibly grateful to the colleagues who submitted their papers and to the referees who provided thoughtful and constructive feedback within the tight deadlines provided by the journal. We acknowledge the hard work and commitments of the authors, who have implemented revisions that addressed all essential comments and suggestions of the referees. As guest editors, it was a pleasure for us to observe this prompt, collegial and constructive reviewing process. In our tasks and efforts, we were assisted by the efficient support provided by the Editor-in-Chief, Marc Paolella, and the assistant editor.

This Special Issue deals with problems of statistical inference and the use of *p*-values. Recently, the issue of the use of *p*-values in various scientific investigations and data analytics techniques has raised questions regarding the validity of statistical decision-making in social sciences including the business and economics disciplines. It is common practice among practitioners and researchers to make statistical decisions exclusively by using the "$p$-value $< 0.05$" criterion, regardless of sample size, statistical power and/or expected loss function underlying the selected models. Some of the well-known scholars have raised serious concerns about this practice and have warned that the use of "*p*-value" may lead to wrong decisions and give distorted scientific results. As an example, we quote a statement made by the American Statistical Association (Wasserstein and Lazar 2016) and the presidential address given by the American Finance Association (Harvey 2017). A few studies have commented on this issue by presenting empirical evidence, such as the paper by Keuzenkamp and Magnus (1995) and McCloskey and Ziliak (1996) for economics, Fazal et al. (2020) for energy, Kim et al. (2018) for accounting, and Kim and Choi (2017) for finance, among others.

The problem has become more challenging with increasing availability of large data sets. In particular, it is widely recognized that statistical significance (based on the conventional *p*-value criterion) is becoming irrelevant for big data (see, for example, Gandomi and Haider 2015). To this end, Rao and Lovric (2016) maintain that "the 21st century researchers work towards a '*paradigm shift*' in testing statistical hypothesis". There are calls that the researchers conduct more extensive exploratory data analysis before inferential statistics are considered for decision-making (see, for example, Leek and Peng 2015; Soyer and Hogarth 2012). There are even calls that the use of statistical significance based on the *p*-value criterion should be abandoned (Wasserstein et al. 2019). In light of these criticisms and calls for change, the Special Issue has been proposed.

The present Special Issue of *Econometrics* is a collection of seven excellent papers that address some of the following topics:

1. New or alternative methods of hypothesis testing such as estimation-based method (e.g., confidence interval), predictive inference, and equivalence testing.
2. Application of adaptive or optimal level of significance to business decisions.
3. Decision–theoretic approach to hypothesis testing and its applications.
4. Compromise between the classical and Bayesian methods of hypothesis testing.

check for updates

**Citation:** Bhatti, Muhammad Ishaq, and Jae H. Kim. 2021. Towards a New Paradigm for Statistical Evidence in the Use of *p*-Value. *Econometrics* 9: 2. https://doi.org/10.3390/econometrics9010002

Received: 23 December 2020
Accepted: 24 December 2020
Published: 31 December 2020

5. Exploratory data analysis for large or massive data sets.
6. Critical review papers on the current practice of hypothesis testing and future directions in business.

This Special Issue begins with an inaugural article by Richard Startz entitled, **"Not *p*-Values, Said a Little Bit Differently",** which is an important contribution toward the ongoing discussion about the use and/or misuse of *p*-values. Numerical examples are presented which demonstrate that a *p*-value can, as a practical matter, give you a different answer than the one that you want. Further contributions to the topic come from Thomas Dyckman and Stephen A. Zeff on **"Important Issues in Statistical Testing and Recommended Improvements in Accounting Research"**. This paper proposes improvements to both the quality and execution of research related to statistical inference in developing statistical tests which address the limitations in existing literature. They explore the situational effects of "data carpentry", alternatives to winsorizing, and suggest necessary improvements instead of relying on a study's calculated "*p*-values".

One of the many highlights of this Special Issue is the paper titled **"Interval-Based Hypothesis Testing and Its Applications to Economics and Finance"** authored by Jae Kim and Andrew P. Robinson. This paper tackles a long-standing literature review on interval-based hypothesis testing (such as tests for minimum-effect, equivalence, and non-inferiority) widely used in biostatistics, medical science, and psychology. It presents the methods in the contexts of a one-sample *t*-test and a test for linear restrictions in a regression. The paper employs testing for market efficiency, validity of asset-pricing models, and persistence of economic time series. Authors argue that, from the point of view of economics and finance, interval-based hypothesis testing provides more sensible inferential outcomes than those based on point-null hypothesis. It proposes interval-based tests which can be routinely used in empirical research in business, as an alternative to point null hypothesis testing, especially in the new era of big data.

Another paper addressing a similar issue is written by David Trafimow, entitled **"A Frequentist Alternative to Significance Testing, *p*-Values, and Confidence Intervals"**. In this article David begins his debate about null hypothesis significance testing, *p*-values without null hypothesis significance testing, and confidence intervals. The first major section addresses some of the main reasons these procedures are problematic and concludes that none of them are satisfactory. However, there is a new procedure, termed the a priori procedure (APP), which validly aids researchers in obtaining sample statistics that have acceptable probabilities of being close to their corresponding population parameters. The second major section provides a description and review of APP advances. Not only does the APP avoid the problems that plague other inferential statistical procedures, but it is easy to perform, too. Although the APP can be performed in conjunction with other procedures, the present recommendation is that it be used alone.

The fifth important paper is the contribution of Jan Magnus, who addresses the issue of the use of *t*-ratios. The title of this paper is **"On Using the *t*-Ratio as a Diagnostic"**, in which the author points out that *tests* and *diagnostics* are the two uses of *t*-ratios in econometrics. The paper proposes a new pretesting method *model averaging* over *t*-ratio and pretest estimators.

The sixth paper of the Special Issue is authored by John Quiggin with the title **"The Replication Crisis as Market Failure"**. Adopting a microeconomic approach, John's paper begins with the observation that the constrained maximization central to model estimation and hypothesis testing may be interpreted as a kind of profit maximization. The output of estimation is a model that maximizes some measure of model fit, subject to costs that may be interpreted as the shadow price of constraints imposed on the model. The replication crisis may be regarded as a market failure in which the price of "significant" results is lower than would be socially optimal.

The seventh paper is on the pedagogy of econometrics, entitled "**Teaching Graduate (and Undergraduate) Econometrics: Some Sensible Shifts to Improve Efficiency, Effectiveness, and Usefulness**" by Jeremy Arkes. According to Wasserstein et al. (2019), "Statistics education will require major changes at all levels to move to a post '$p < 0.05$' world". As educators, we will need to rethink the way we train the future decision-makers, especially in the big data era where the $p$-value criterion is no longer relevant. Jeremy proposes a range of critical points on the issue of teaching econometrics, including the problem related to the $p$-value, maintaining that the teaching of graduate (and undergraduate) econometrics needs to be revamped.

We are very thankful to all authors, who have made considerable efforts to meet the standards of the journal. We believe that this Special Issue has been very successful in attracting seven high-quality contributions from established and well-known scholars in their prospective fields. The journal has provided an open access-publishing facility to the contributors, which is a realistic option for our discipline. We hope this Special Issue will stimulate further research and novelty in understanding a new paradigm for statistical evidence of $p$-value insights. We take this opportunity to thank the numerous reviewers who have greatly contributed to the quality of the published papers. Finally, we thank the editor-in-chief, Marc Paolella, and the team of assistant editors without whose devotion this Special Issue would not have been produced in such a smooth and well-managed manner.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

Fazal, Rizwan, Syed Aziz Ur Rehman, Atiq Ur Rehman, Muhammad Ishaq Bhatti, and Anwar Hussain. 2020. Energy-Environment-Economy causal nexus in Pakistan: A Graph Theoretic Approach. *Energy* 214: 118934. [CrossRef]

Gandomi, Amir, and Murtaza Haider. 2015. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management* 35: 137–44. [CrossRef]

Harvey, Campbell R. 2017. Presidential Address: The Scientific Outlook in Financial Economics. *Journal of Finance* 72: 1399–440. [CrossRef]

Keuzenkamp, Hugo A., and Jan R. Magnus. 1995. On tests and significance in econometrics. *Journal of Econometrics* 67: 103–28. [CrossRef]

Kim, Jae H., and In Choi. 2017. Unit Roots in Economic and Financial Time Series: A Re-evaluation at the Decision-based Significance Levels. *Econometrics* 5: 41, This article belongs to the Special Issue *Celebrated Econometricians: Peter Phillips*. [CrossRef]

Kim, Jae H., Kamran Ahmed, and Philip Inyeob Ji. 2018. Significance Testing in Accounting Research: A Critical Evaluation based on Evidence. *Abacus* 54: 524–46. [CrossRef]

Leek, Jeffrey T., and Roger D. Peng. 2015. Statistics: *P* values are just the tip of the iceberg. *Nature* 7549: 520–612. [CrossRef] [PubMed]

McCloskey, Deirdre N., and Stephen T. Ziliak. 1996. The standard error of regressions. *Journal of Economic Literature* 34: 97–114.

Rao, Calyampudi Radhakrishna, and Miodrag M. Lovric. 2016. Testing Point Null Hypothesis of a Normal Mean and the Truth: 21st Century Perspective. *Journal of Modern Applied Statistical Methods* 15: 2–21. [CrossRef]

Soyer, Emre, and Robin M. Hogarth. 2012. The illusion of predictability: How regression statistics mislead experts. *International Journal of Forecasting* 28: 695–711. [CrossRef]

Wasserstein, Ronald L., and Nicole A. Lazar. 2016. The ASA's statement on $p$-values: Context, process, and purpose. *The American Statistician* 70: 129–33. [CrossRef]

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

econometrics

MDPI

*Article*

# Teaching Graduate (and Undergraduate) Econometrics: Some Sensible Shifts to Improve Efficiency, Effectiveness, and Usefulness

**Jeremy Arkes**

Naval Postgraduate School, Monterey, CA 93943, USA; arkes@nps.edu

Received: 26 November 2019; Accepted: 31 August 2020; Published: 7 September 2020

**Abstract:** Building on arguments by Joshua Angrist and Jörn-Steffen Pischke arguments for how the teaching of undergraduate econometrics could become more effective, I propose a redesign of graduate econometrics that would better serve most students and help make the field of economics more relevant. The primary basis for the redesign is that the conventional methods do not adequately prepare students to recognize biases and to properly interpret significance, insignificance, and p-values; and there is an ethical problem in searching for significance and other matters. Based on these premises, I recommend that some of Angrist and Pischke's recommendations be adopted for graduate econometrics. In addition, I recommend further shifts in emphasis, new pedagogy, and adding important components (e.g., on interpretations and simple ethical lessons) that are largely ignored in current textbooks. An obvious implication of these recommended changes is a confirmation of most of Angrist and Pischke's recommendations for undergraduate econometrics, as well as further reductions in complexity.

**Keywords:** teaching of econometrics; regression analysis; economics pedagogy

## 1. Introduction

On 23 January 2015, basketball player Klay Thompson of the Golden State Warriors hit all 13 of his shot attempts in the 3rd quarter of a game against the Sacramento Kings—this included making 9 of 9 on 3-point shots[1]. These 3-point shots were not all wide-open 3-point shots players typically take (with the team passing the ball around until they find an open player). Rather, several of them were from far beyond the 3-point line or with a defender close enough to him that under normal circumstances, few would dare take such a heavily contested shot.

Everyone knew that Klay Thompson was "in the zone" or "*en fuego*", or that Thompson had the "hot hand" that night. Everyone that is … unless you are a statistician, a psychologist, or an economist (particularly, a Nobel-Prize-winning economist) without adequate training in econometrics or regression analysis. Starting with Gilovich et al. (1985), an entire literature over 25 years found no evidence for the hot hand in basketball. Even the famous evolutionary biologist, Steve Jay Gould, got in on this research (Gould 1989). From the results, these researchers claimed that the hot hand was a "myth" or "cognitive illusion".

This was an incredibly appealing result: that all basketball players and fans were wrong to believe in the hot hand (players achieving a temporary higher playing level) and that they were committing

[1] See https://www.youtube.com/watch?v=BNHjX_08FE0. One extra 3-pointer he made came after a referee whistle, so he was actually (but not officially) 10 of 10. This performance harkens back to a game in which Boston Celtic Larry Bird hit every low-probability shot he put up, as he racked up 60 points against the Atlanta Hawks in 1985—https://www.youtube.com/watch?v=yX61Aurz3VM. (The best part of the video is the reaction by the Hawks' bench to some of Bird's last shots—those opponents knew Bird had the "hot hand").

the cognitive bias of seeing patterns (the hot hand) in data that, the researchers claimed, were actually random and determined by a binomial process. Therefore, the story has shown up in many popular books—e.g., *Nudge* (Thaler and Sunstein 2009) and *Thinking Fast and Slow* (Kahneman 2011). Note that Kahneman and Thaler are the 2002 and 2017 winners of the Nobel Prize in economics, respectively. In addition, this was a story that a recent-Harvard-President-and-almost-Fed-Chairman-nominee gave to the Harvard men's basketball team, as he brought media along in his address to the team (Brooks 2013).

However, it turns out, these researchers and Nobel laureates failed to recognize a few biases to the estimated relationship between making prior shots and the current shot—i.e., alternative explanations for why there was no significant relationship. In addition, they made a major logical error in their interpretation. Both are discussed in a moment.

From my experience playing basketball and occasionally experiencing the hot hand, I knew the researchers were wrong to conclude that the hot hand was a myth (This, as it turns out, is an example of the fact that sometimes, there are limits to what data can tell us; and, the people engaged in an activity often will understand it better than researchers trying to model the activity with imperfect data or imperfect modeling techniques). Eventually, I developed a more powerful model by pooling all players together in a player-fixed-effects model rather than have players analyzed one at a time, as in the prior studies. In Arkes (2010), I found the first evidence for the hot hand, showing that players were about 3- to 5-percentage points more likely to make a second of two free throws if they had made their first free throw.

Yet, I failed to recognize an obvious bias in past studies and my own study that Stone (2012) noted: measurement error. Measurement error is not just from lying or a coding error. It could also stem from the variable not representing well the concept that it is trying to measure—a point that eluded me, along with the prior researchers. Therefore, whether a player made their first free throw is an imperfect indicator of whether the player was in the hot-hand state, and the misclassification would likely cause a bias towards zero in the estimated hot-hand effect. There was another major problem in these studies from the Gambler's Fallacy, as noted by Miller and Sanjurjo (2018). This leads to a negative bias (not just towards zero, as would bias from measurement error). Both biases make it more difficult to detect the hot hand.

Reading Stone (2012) was a watershed moment for me. I realized that in my graduate econometrics courses, I had learned equation-wise how these biases to coefficient estimates work in econometrics, but I never truly learned how to recognize some of these biases. And, this appears to be a pattern. The conventional methods for teaching econometrics that I was exposed to did not teach me (nor others) how to properly scrutinize a regression. Furthermore, given that such errors were even being committed by some of those we deem to be the best in our field, this appears to be a widespread and systemic problem.

What was also exposed in these studies and writings on the hot hand (beyond the failure to recognize the measurement error) was the authors' incorrect interpretations. They took the insignificant estimate to indicate proof that the hot hand does not exist (A referee at the first journal to which I sent my 2010 hot-hand article wrote that the research had to be wrong because "it's been proven that the hot hand does not exist"). This line of reasoning is akin to taking a not-guilty verdict or a finding of "not enough evidence for a crime" and claiming that it proves innocence. The proper interpretation should have been that the researchers found no evidence for the hot hand. And now, despite the hurdles of negative biases, there is more evidence coming out that the hot hand is real (e.g., Bocskocsky et al. 2014; Miller and Sanjurjo 2018).

This article is my attempt to remedy relatively common deficiencies in the econometric education of scholars and practitioners. I contend that inadequate econometrics education directly drives phenomena such as the errors in the hot-hand research and on other research topics I will discuss below. Although the veracity or falsifiability of the basketball hot hand probably does not materially affect anyone, errors in research can affect public perceptions, which in turn affects how much influence academics can have.

Angrist and Pischke (2017) recently called for a shift in how undergraduate econometrics should be taught. Their main recommended shifts were:

(1) The abstract equations and high-level math should be replaced with real examples;
(2) There should be greater emphasis on choosing the best set of control variables for causal interpretation of some treatment variable;
(3) There should be a shift towards randomized control trials and quasi-experimental methods (e.g., regression-discontinuities and difference-in-difference methods), as these are the methods most often used by economists these days.

Angrist and Pischke's recommendations, particularly (2) and (3), appear to be largely based on earlier arguments they made (Angrist and Pischke 2010) that better data and better study designs have helped economists take the "con" out of econometrics. They cite several random-assignment studies, including studies on the effects of cash transfers on child welfare (e.g., Gertler 2004) and on the effects of housing vouchers (Kling et al. 2007).

In this article, I build on Angrist and Pischke (2017) study to make the argument for a redesign of graduate econometrics. I use several premises, perhaps most notably: (a) a large part of problems in research is from researchers not recognizing potential sources of bias to coefficient estimates, incorrectly interpreting significance, and potential ethical problems; (b) any bias to coefficient estimates has a much greater potential to threaten the validity of a model than bias to standard errors.

And so, the general redesign I propose involves a change from the high-level-math econometric theory to a more practical approach and shifts in emphasis towards new pedagogy for recognizing when coefficient estimates might be biased, proper interpretations, and ethical research practices. I argue that the first two of Angrist and Pischke (2017) arguments should apply to graduate econometrics as well. However, because some of the models in their third argument are based on the rare instances of randomness or having the data to do a more-complicated quasi-experimental method, I recommend a shift in emphasis away from these towards more practical quasi-experimental methods (such as fixed effects). The idea is, rather than teaching people how to find randomness and build a topic around that, it might be more worthwhile for students to learn how to deal with the more prevalent research problem of needing to use less-than-ideal data.

My new recommended changes are:

A. Increase emphasis on some regression basics ("holding other factors constant" and regression objectives)
B. Reduce emphasis on getting the standard errors correct
C. Adopt new approaches for teaching how to recognize biases
D. Shift focus to the more practical quasi-experimental methods
E. Add emphasis on interpretations on statistical significance and $p$-values
F. Advocate less complexity
G. Add a simple ethical component.

Although most of the article makes the case for changes to graduate econometrics, my argument implies that undergraduate econometrics needs a similar redesign. This follows directly from the arguments on graduate econometrics, along with the idea that the common approach, using high-level math, is teaching undergraduates as if they would all become econometric theorists; probably less than one percent of them will.

The ideas and arguments I present come from my experiences in two types of worlds: in research organizations (where I had to develop models to assess policy options) and as an academic (creating my own research and teaching about econometrics).

The article proceeds, in Section 2, with a discussion of the premises behind why I believe changes are needed and demonstrates how much various topics are covered and how little more important topics are covered in the leading textbooks. Section 3 presents some examples of topics with decades of research failing to recognize biases and examples of my own research errors. Section 4 discusses my proposed changes. Section 5 makes the case for changes to undergraduate econometrics. I provide conclusions in Section 6.

## 2. Why a Redesign Is Needed

In this section, I give five reasons why there needs to be a major shift in teaching graduate econometrics, and I show what is emphasized in leading graduate textbooks. By "major shift" or "redesign", I mean that there should be new topics, new pedagogy (for teaching how to scrutinize a regression), and shifts in emphasis for what is taught among existing topics. The five reasons I give also serve as the premises for support of some of Angrist and Pischke (2017) recommendations on redesigning undergraduate econometrics and for the recommended changes I give in Section 4. The five reasons are:

- There are concerns on the validity of much economic research;
- Biases in coefficient estimates threaten a model's validity much more than biases in standard errors;
- The conventional methods for teaching econometrics do not adequately prepare students to recognize biases to coefficient estimates;
- The high-level math and proofs are unnecessary and take valuable time away from more important concepts; and
- There are ethical problems in research, namely on the search for significance and not fully disclosing potential sources of bias.

### *2.1. There Are Concerns on the Validity of Much Economic Research*

There is growing evidence of problems with validity in all academic research, and economics certainly has its problems. In my view, there are three main sources of the concerns. First, there are some topics that have conflicting results in the research—e.g., the research on the effects of minimum-wage increases (see Gill 2018). Second, there are errors in interpretations. For example, akin to the incorrect interpretations in the hot-hand research, Cready et al. (2019) find that 65% of articles in the top Accounting journals with null results misrepresent the true meaning of those null results. I am not aware of a similar study for economics, but as I will discuss below, it is taught incorrectly in several leading econometric textbooks.

Third and (in my view) most importantly, researchers sometimes fail to recognize or fully acknowledge potential biases to the coefficient estimates. Not addressing potential biases could result in the failure of studies to be replicated. This certainly could be the cause of some cases of conflicting results. In Section 3, I give some examples in a few economics topics in which nearly the entire literature failed to recognize likely biases. This highlights the point below that the current methods are not working well for preparing students to develop proper models and to recognize the biases.

### *2.2. Biases in Coefficient Estimates Threaten a Model's Validity More Than Biases in Standard Errors*

From my experience, almost all corrections for clustering or heteroskedasticity result in standard errors being adjusted less than 15%. That said, there can be instances of much larger bias in the standard errors, particularly for panel data sets. For example, Petersen (2009) finds that the bias in standard errors for finance panel data sets is as high as 45% under certain circumstances. However, generally speaking, the bias on coefficient estimates from any of the major pitfalls (e.g., reverse causality, omitted-variables bias, and measurement error) could be significantly larger and, except for measurement error, even produce an estimated effect that has the reverse sign of the true effect (which would mean more than a 100% bias).

Supporting this idea is the contention that the major research errors are more likely to come from biased coefficient estimates than biased standard errors. For example, the initial research on estrogen replacement therapy (based on observational data) suggested that it was highly beneficial to women in terms of reduced mortality (e.g., Ettinger et al. 1996). However, a follow-on randomized control trial in 2002 found that taking estrogen could actually lead to a greater risk of death (Rossouw et al. 2002). And, later research after following the participants in the randomized study for longer found that taking estrogen could actually improve health outcomes (Manson et al. 2017), depending on age.

### *2.3. Current Methods Do Not Teach How to Recognize Biases*

This statement is based on several observations. First, as mentioned earlier, there are problems of validity in some academic research. Second, after having received the conventional training in econometrics, I have failed in several instances to recognize pitfalls and biases in my own research. Third, just by common sense, it must be difficult to translate the concept of conditional mean dependence/independence of the error term (the conventional criterion) to recognize whether a coefficient estimate might be biased (from, for example, omitted-variables bias and measurement error). I admittedly have difficulty and must think hard about making this connection. Fourth, to the best of my knowledge, conditional mean dependence of the error term cannot explain the bias from the inclusion in a model of mediating factors, or "bad controls", as Angrist and Pischke (2009) call them. These are variables that are part of the mechanism for why the treatment affects the outcome. (This is different from "collider" variables used in the Directed Acyclic Graph approach, in which a variable is affected by both the treatment and the outcome.)

### *2.4. The High-Level Math and Proofs Are Unnecessary and Take Valuable Time Away from More Important Concepts*

I consider myself to be a "generalist" researcher, with deeper dives into military, labor, health, behavioral, and sports economics. In my dozens of publications and dozens of reports, I never needed the calculus or linear algebra that was used in the econometrics courses I took. Although the necessary math underlying basic probability theory and statistics was important, the calculus and linear algebra used in econometrics never helped me understand the real nuances of what happens when you hold other factors constant nor how to recognize the pitfalls and sources of bias. What contributed to my understanding of these things has been the intuition I have gained from using regressions for many research projects and from the mistakes I made—mistakes due to not adequately grasping how to recognize the pitfalls of regression analysis.

And so, along the lines of Angrist and Pischke (2017) argument, real examples would be much more useful and practical than the math underlying the regressions. The lessons from examples are almost certainly more likely to be retained than abstract equations. Adding visual aids could be even more effective.

This is not to say that the high-level math theory is not important for all students. For those aiming to study econometric theory, they would need that more mathematical approach. However, for improvisation of applying the concepts to new situations, students would likely benefit more from examples than from knowing the high-level math underlying the econometrics.

Let me emphasize that this is my view, based on my experiences described above. As I look back to the errors I have made, what would have helped more than the math for me would have been more practical experience on recognizing pitfalls and understanding the nuances of certain techniques. However, others feel differently and believe the math is essential.

### *2.5. There Is an Ethical Problem in Economic Research*

In the scores of job-market-candidate seminars I have attended in my two decades since graduate school, I do not remember one in which the candidate had an insignificant coefficient estimate on the key explanatory variable. The high percentage of significant results could be due to graduate students giving up on a topic if the results do not support the theory they developed. However, it could also partly stem from some searching for significance (or p-hacking), meaning that some students keep changing the model (by adding or cutting control variables or by changing the method) until they achieve a desirable result. There has been mixed evidence on p-hacking; one study that found evidence for p-hacking is Head et al. (2015), although they argue that the extent of it is relatively minor when compared to effect sizes.

Another issue, mentioned above as a source of validity problems, is that researchers are not always fully honest and forthright about potential limitations of a study. To do so would reduce their chance of being published. Or, for those producing reports for sponsors (e.g., at research organizations), I suspect that many do not want to express any lack of confidence in their results.

These ethical problems are certainly not universal, as most research is probably done objectively and honestly. However, likely due to the pressures to publish and raise research funds, there is certainly a portion of research that could be conducted more responsibly. Simple ethical lessons might be able to help.

### *2.6. What the Textbooks Teach*

Table 1 shows my estimates on the number of pages devoted to various topics in the six textbooks I believe are the most widely used for graduate econometrics. This is not a scientific assessment, as it is based on my judgment of the number of pages having the discussion centered around the topic and does not include other mentions of the topic. One pattern is that other than for "simultaneity", there appears to be greater emphasis on the things that could bias the standard errors than there is on the things that could bias the coefficient estimates. In fact, one of the potential sources of bias for coefficient estimates (inclusion of mediating factors) is not even mentioned other than by Angrist and Pischke (2009), and there is minimal discussion for the other biases. The large number of pages I indicate is devoted to simultaneity might be misleading, as few of these pages are devoted to identifying when it could occur and the direction of the bias. In fact, "reverse causality" is a very small part of number of pages devoted to simultaneity (and not mentioned in most of these).

**Table 1.** What the main graduate textbooks teach (number of pages on a given topic).

| | **Goldberger (1991)** | **Hayashi (2000)** | **Russell and MacKinnon (2004)** | **Angrist and Pischke (2009)** | **Greene (2012)** | **Wooldridge (2012)** |
|---|---|---|---|---|---|---|
| **Causes of bias in the standard errors** | | | | | | |
| Heteroskedasticity | 0 | 10 | 11 | 2 | 2 | 8 |
| Multicollinearity | 7 | 0 | 1 | 0 | 2 | 0 |
| **Causes of bias in the coefficient estimates** | | | | | | |
| Simultaneity | 6 | 3 | 13 | 0 | 22 | 32 |
| Omitted-variables bias | 1 | 1 | 0 | 4 | 1 | 1 |
| Measurement error | 0 | 3 | 2 | 1 | 2 | 6 |
| Mediating factors | 0 | 0 | 0 | 4 | 0 | 0 |
| **Other important topics** | | | | | | |
| Holding other factors constant | 0 | 1 | 0 | 0 | 0 | 0 |
| Fixed effects | 0 | 23 | 5 | 12 | 12 | 9 |
| Bayesian critique of *p*-values | 0 | 0 | 0 | 0 | 0 | 0 |
| Correctly indicates an insignificant coef. estimate does not mean accept the null | No | No | Yes | N/A | No | N/A |

Note: "N/A" in the last row indicates that I do not believe the topic of how to interpret insignificant estimates is discussed.

Furthermore, in most of these books, there is for the most part no discussion on the intuition behind "holding other factors constant" and what exactly happens when you do so, and there is no discussion in any of the books on the Bayesian critique of *p*-values. In addition, three of the four books that discuss hypothesis tests incorrectly state that an insignificant coefficient estimate indicates that one should accept the null hypothesis.

Assuming that econometrics courses mirror these books, there are many changes needed in the teaching of graduate econometrics, as the typical emphasis in econometrics appears to be on things that diverge from the reality of the problems that practitioners face.

## 3. Research Topics with Decades of Research Errors

Responding to Leamer (1983) critique on the unreliability of econometric research, Angrist and Pischke (2010) argued that better data and better research designs have improved the credibility of econometric research. I imagine that overall, there have been improvements in research. However, plenty of unreliable research continues to be published.

I will discuss in this section the following three research topics in which the investigators failed to recognize likely biases and did not realize it for decades:

(A) The hot hand in basketball, continuing the discussion from the Introduction;
(B) The public-finance/macroeconomic topic of how state tax rates affect Gross State Product;
(C) How occupation-specific bonuses affect the probability of reenlistment in the military.

### *3.1. The Hot Hand in Basketball*

The world is not necessarily better off with knowledge of whether the hot hand in basketball is real or not. However, if it turns out that there is no hot hand, which would stand in contrast with what the population believes, then this would be indicative of a mass cognitive illusion. However, in my view, the real value in the research comes from the arc of the story on the research and the mistakes made.

As discussed in the Introduction, no researcher in the first 25 years of study on the hot hand in basketball found any evidence for the hot hand. These studies were based on runs tests, conditional-probability tests, and stationarity tests for individual players, finding no statistically significant evidence for a hot-hand effect—see Bar-Eli et al. (2006) for a review of the early studies. The researchers (and Nobel Prize winners writing about this research) claimed that the "hot hand is a myth" or a "figment of our imaginations". However, in Arkes (2010), I pooled all players into a player-fixed-effects model (to generate more power) that regressed "whether a player made a second free throw in a set of two or three free throws" on "whether the player made the first free throw". I found a small but significant hot-hand-effect of 3 to 5 percentage points. Still, this study turned out to be flawed.

The first major error the researchers made is that they interpreted an insignificant estimate as proof of non-existence. However, as the saying goes, absence of evidence is not evidence of absence. The correct interpretation should have been that there is no evidence for the hot hand. This is a common logical error made throughout academia (not just Economics and Statistics), and it was highlighted in Amrhein et al. (2019), which I discuss below.

The second major error is that the researchers (including myself this time) failed to recognize what should have been an obvious bias: measurement error. Stone (2012) noted that the hot hand means that a player is in a state in which he/she has a higher probability than normal of making a shot (which contrasts with the conventional thought that a player "can't miss"). This means that a player can be in the hot-hand state and miss a shot, and the player can be in the normal state and make a few shots in a row, making it seem as if he/she is in that hot-hand state.

This means that the crude indicator I used for being in the hot-hand state in Arkes (2010)—making the first of two free throws—and the indicators that others have used (e.g., making the last three shots) could very well occur in the normal state. In addition, having missed the prior shot(s) could still occur in the "hot hand" state. The misclassification (measurement error) likely caused a downward bias, and this certainly could have contributed to the failure of most studies to detect the hot hand. In addition, the hot-hand effect I found for free throws in Arkes (2010) was probably a gross understatement due to the measurement error.

Miller and Sanjurjo (2018) found another major error in this research, related to the Gambler's Fallacy, that was not so obvious. They demonstrated that if you take all "heads" in a finite sequence of coin flips, the probability that the following flip is "heads" is actually less than 50%—yes, this is true! They then applied this to the hot-hand application and, with the correction, actually found a significant hot-hand effect with the data used in the seminal hot-hand study (Gilovich et al. 1985).

This source of bias again highlights the flaws in the original interpretations that the lack of evidence for the hot hand proved it was a myth. Not only had most of the literature misinterpreted the significance tests, but they had not given the model a thorough scrutiny of the potential biases that could speak to whether the lack of any estimated effect was correct.

### *3.2. How State Income Tax Rates Affect Gross State Product*

Another research topic that has had questionable modeling strategies has been on how state tax rates affect state economic growth. The convention has been to use a Cobb-Douglas model as the theoretical framework underpinning the econometric model. The Cobb-Douglas model has state economic growth as a function of tax rates, as well as other economic factors such as labor and capital. Therefore, models typically include control variables reflecting labor and capital. For example, several studies include a measure of the unemployment rate as a control variable (Mofidi and Stone 1990; Bania et al. 2007). Others use the amount of capital (Reed 2008; Yeoh and Stansel 2013). And, some studies even include state personal income per capita (Wasylenko and McGuire 1985; Poulson and Kaplan 2008) or the wage rate (e.g., Wasylenko and McGuire 1985; Funderburg et al. 2013) as control variables.

Including these variables may not have been the best approach. Bartik (1991) raised an important consideration for these models that has been largely ignored in the literature: that several factors of economic growth are endogenous. Variables such as the average wage rate, the labor supply (proxied by the unemployment rate), the level of capital, and capital growth are all factors of economic growth and, at the same time, are measures of economic growth that could depend on the tax rate. They are what Angrist and Pischke (2009) described as "bad controls" in that they come after the treatment (taxes) and control for part of the effect of the tax rate. Or, Arkes (2019) considers such variables as potential mediating factors for how the tax rate affects economic growth, i.e., tax rates could affect how much investment and employment growth there is, which in turn affect economic growth. We can also think of investment, the unemployment rate, employment growth, and personal income per capita being themselves outcomes of tax rates.

Including these variables means that what is being estimated is something akin to (but not exactly) the effect of tax rates on Gross State Product beyond the effects on employment growth, investment, and/or personal income per capita. This is no longer informative on how tax rates affect Gross State Product. The counterargument is that excluding these factors from the model could cause omitted-variables bias. At the very least, the issue of mediating factors versus omitted-variables bias should be acknowledged by the researchers.

### *3.3. How Occupation-Specific Bonuses Affect the Probability of Reenlistment*

This is an important research issue for the military services, as they try to set the optimal bonuses to efficiently achieve a required reenlistment rate in an occupation. In over 40 years of research on this topic, all studies have been subject to numerous biases, some of which were only recently recognized. The typical model would be:

$$R_{io} = \beta_1 \times (\mathrm{BONUS})_{io} + X_{io}\beta_2 + \mu_o + \varepsilon_{io},$$

where $R_{io}$ is the reenlistment/retention decision for serviceperson i in occupation o, BONUS is either a dollar amount or a multiple-of-basic-pay determining the amount a serviceperson would receive, X would be a set of other factors, such as year, home-state unemployment rate, and more, and $\mu_o$ represents occupation fixed effects.

Arkes (2018) describes four major sources of bias in these studies:

1. There is the obvious bias of reverse causality in that lower reenlistment rates lead to higher bonuses.
2. Enlisted personnel often have latitude on when they reenlist, so if they were planning to reenlist, they may time it to when the bonus appears to be higher than normal; so, the bonus is endogenous in that it is chosen to some extent by those reenlisting. This is an indirect reverse causality, in that the choice of reenlisting or not (R) would affect the timing of the reenlistment; and those choosing to reenlist would tend to do so when the bonus is relatively high within their reenlistment window.
3. There is likely bias from measurement error, as servicepersons often have a few different bonuses during their reenlistment window, and the one most often recorded is the one at the official reenlistment date, not the one when they sign the new contract (which is not among the available data and can be up to two months earlier).
4. There is unobserved heterogeneity because excess supply for reenlistments can mean that we only observe whether a person reenlists rather than whether he/she is willing to reenlist (or actual reenlistment supply). Excess supply of reenlistments could result from reduced demand from the military (e.g., occupations being eliminated) or worsening civilian prospects for the skill. Excess supply when an occupation is eliminated (and the reenlistment rate and bonus equals zero) could lead to a large exaggeration of the bonus effect.

In the numerous studies on this topic—see Arkes (2018) for a list of some of the more recent studies—none recognized the third and fourth sources of bias, and only one (Goldberg 2001) recognized the second source of bias. Furthermore, most studies attempted to address the reverse causality with separate occupation and year fixed effects. However, any variation across occupations in changes in the propensity to reenlist (due to changing civilian-economy opportunities or military environment) would still result in this reverse causality. I used occupation-fiscal-year-interacted fixed effects to reduce the bias from reverse causality, but I acknowledged that it likely led to greater bias from measurement error (Arkes 2018), as occurs often with fixed effects (see below).

The ultimate result of all this is that with the historical and current reenlistment rules and inadequate data, this is a research question that just cannot be accurately answered with any adequate degree of confidence that the potential biases are being addressed. Indeed, Hansen and Wenger (2005) note that different assumptions in such models produce widely different results. Even random assignment would probably not work well, as servicepersons would likely know whether they received a high or low bonus, and any perceived inequity could have its own effects on retention.

### 3.4. Summary

These three topics highlight how the entire literature on a topic can go decades without recognizing likely sources of bias. I cannot speak to how extensive this is among the many big topics in empirical economics, but there must be other topics that have had similar problems. For example, there is a literature on how the state unemployment rate (or other measures of local economies) affects various health or social outcomes—I have had several articles in this literature. I cannot recall one (including mine) that recognized that using state fixed effects (or controls) exacerbates any bias from measurement error in the economic measure, which would likely cause attenuation bias—a lesson I learned far too late into my career. This is not terribly harmful in this case, as at least it is a bias against finding significant estimates, but it has been an unrecognized bias, nonetheless.

## 4. Recommended Changes and New Topics for Graduate Econometrics

In this section, I propose seven new recommendations for redesigning graduate econometrics courses, most of which follow from the premises from Section 2. However, first, I contend that the first two of Angrist and Angrist and Pischke (2017) recommendations for changing undergraduate econometrics would work well for changes to graduate econometrics. These recommended changes are:

- Replace the math with intuition and examples
- Focus on choosing the best set of control variables.

The first is consistent with basic tenets of pedagogical theory, as the practice of some skill to learn about certain concepts can be much more instructive than learning the abstract equations underlying the concepts. The second one is important for avoiding potential biases, and it goes hand in hand with my first new added recommendation below.

What follows are my new recommended changes. These are the new components, changes in pedagogy, and shifts in emphasis that should help to develop effective and responsible academics and practitioners. The recommended changes and shifts I will discuss are:

A. Increase emphasis on some regression basics ("holding other factors constant" and regression objectives)
B. Reduce emphasis on getting the standard errors correct
C. Adopt new approaches for teaching how to recognize biases
D. Shift focus to the more practical quasi-experimental methods
E. Add emphasis on interpretations on statistical significance and $p$-values
F. Advocate less complexity
G. Add a simple ethical component

A. Increase emphasis on some regression basics ("holding other factors constant" and regression objectives)

These two concepts of "holding other factors constant" and the various regression objectives are important building blocks needed to understand when there could be potential bias to a coefficient estimate and for determining the optimal set of control variables to use—Angrist and Pischke's second point. In addition, together they should help foster understanding of why modeling strategies should be different depending on the objectives of a regression analysis.

I believe it is commonly assumed that students will understand "holding other factors constant" from the few pages, if that, devoted to the concept in textbooks. However, in my view, this is usually not the case. Lessons on this topic should include a discussion of the purpose of holding other factors constant, a demonstration of what happens when you do so, and in what circumstances would you not want to hold certain factors constant. In Arkes (2019), I describe a simple issue of whether adding cinnamon to your chocolate-chip cookie improves the taste. In this example, I ask which is the better approach: (1) make two batches from scratch, adding cinnamon to one; or (2) make one batch, split it in two, and add cinnamon to one of them. Most would agree that the second would be a better test because you do not want any other factor that could affect the outcome of taste (butter, sugar, and chocolate chips) to vary as you switch from the no-cinnamon to cinnamon batch, i.e., you want to hold those other factors constant. This is the point of multivariate models: design the model so that the only relevant factor that changes is the treatment or key explanatory variable. That said, with interval (quantitative) variables, it is impossible to perfectly control for the variable, and so perhaps the best that can be said is that one is attempting to adjust for the variable.

In Arkes (2019), I describe what I believe are the four main objectives of regression analyses: (1) estimating causal effects; (2) forecasting/predicting an outcome; (3) determining predictors for an outcome; and (4) analyzing relative performance by removing the influence of contextual factors, which is similar to the concept of "anomaly detection."I proceed to describe how the choice of control variables (what should be held constant) should depend on the objective. For example, a causal-effects model might attempt to estimate the effect of a college degree on the probability of getting in a car accident in a given year. An insurance company, on the other hand, might be more interested in predicting the probability of a person getting in a car accident—the second objective above. One potential control variable in both analyses would be whether the person has a white-collar job. That could be a mediating factor (a "bad control") for how a college degree affects the probability of an accident, so it would be best to exclude that variable in the causal-effects analysis. However, the insurance company might find that variable to be a valuable contributor to obtain a more accurate prediction of the probability of an accident. The insurance company does not care about obtaining the correct estimate of how a college degree affects the likelihood of an accident. Likewise, forecasting GDP (or Gross State Product, GSP) growth would involve a different strategy from that for estimating the effects of tax rates on GDP/GSP growth. In these cases of predicting an outcome or forecasting, including explanatory variables is not meant to hold other factors constant but rather to improve the prediction/forecast.

Some textbooks, e.g., Greene (2012), indicate that the *adjusted* $R^2$ could be used as part of the "model selection criteria". However, any measure of goodness-of-fit would primarily be useful for determining whether a variable should be included for forecasting/prediction. For estimating causal effects, whether a potential control variable contributes to explaining the dependent variable should not be a factor in determining whether it should be included in the model. These are just a few examples of why understanding the objective of the regression is important.

B. Reduce emphasis on getting the standard errors correct

This was a passing point by Angrist and Pischke (2017). However, in my view, it deserves status as one of the main recommendations. The justification for this recommendation is partly based on one of the premises from Section 2: that biases to standard errors are typically minimal compared to the potential biases to coefficient estimates. To this point, Harford (2014) argues that sampling bias can be much more harmful than sampling error, as demonstrated by the 1936 Literary Digest poll that found a 55-41 advantage for Landon over Roosevelt in the Presidential election. The 2.4-million sample size (and tiny standard errors) did not matter when there was sampling bias. This idea goes back to Leamer (1988), who argued that corrections for heteroscedasticity are mere "white-washing" if there is no consideration of the validity of the coefficient estimates.

Further justification for reducing the emphasis on corrections for standard errors comes from the vagueness of the $p$-value and statistical significance. Getting the standard errors correct is typically meant to make proper confidence intervals or correct conclusions on hypothesis tests, which are usually based on t-stats or $p$-values meeting certain thresholds. However, as I learned not too long ago (and far too late into my career), the $p$-value by itself actually has little meaning, given the Bayesian critique of $p$-values. This is discussed by Ioannidis (2005), who points out that the probability that an empirical relationship is real depends on: (1) the t-statistic; (2) the *a priori* probability that there could be an empirical relationship; and (3) the statistical power of the study (and this depends on the probability of a false negative and requires an alternative hypothesized value). The $p$-value is based just on the first one, the t-statistic. The less likely there is such a relationship, *a priori*, the less likely any given t-statistic indicates a significant relationship, as Nuzzo (2014) demonstrates. For example, even for an *a priori* toss-up (50% chance there is a relationship), $p$-values of 0.05 and 0.01 translate to only 71% and 89% probabilities that the relationship is real.

Unfortunately, it is nearly impossible to know beforehand what the probability is that there is an effect of one variable on another. This uncertainty means that higher levels of significance than is the current convention would be needed to make any strong conclusions about statistical relationships being real. Given the vagueness of the $p$-value and that high levels of significance should be used to make any strong conclusions, errors in the standard errors would tend to be much less impactful to those conclusions than would potentially much larger biases in coefficient estimates.

Correcting standard errors for heteroskedasticity and clustering is still important, yet it is easy to recognize when it is needed and typically takes only a few characters of code to correct for. Recognizing and addressing biases to coefficient estimates is more difficult and takes much more practice to become proficient, and so greater emphasis should go towards those concepts.

C. Adopt new approaches to teach how to recognize biases

As I described above, I do not believe that "conditional mean dependence of the error term" is an effective concept to teach how to recognize biases. I believe that calling a source of bias what it is (e.g., reverse causality) rather than what it does (conditional dependence of the error term) is a good starting point. I believe it would be more effective if we were to list the most common sources of bias, provide some visual depictions of the biases (when possible), and give examples of the various types of situations in which they might arise. In Arkes (2019), I list what I believe are the six most common biases for coefficient estimates when estimating causal effects: reverse causality, omitted-variables bias, self-selection bias, measurement error, and including mediating factors or outcomes as control variables. In addition, I give guidance on how to recognize such biases. These are the main *alternative stories* that need to be considered before making conclusions from results. (I since added a 7th bias, from *improper reference groups*[2].) Useful visual depictions could be the "directed acyclical graphical" (DAG) approach (Pearl and Mackenzie 2018; Cunningham 2018), basic flowcharts (Arkes 2019), and animations produced by Nick Huntington-Klein on his website: http://nickchk.com/causalgraphs.html. These tools demonstrate when there could be bias and what needs to be controlled for.

As an example of using visualizations (with flowcharts), let us take again the research issue from Section 3 on how occupation-specific bonuses affect retention decisions in the military:

$$R_{io} = \beta_1 \times (\text{BONUS})_{io} + X_{io}\beta_2 + \mu_o + \varepsilon_{io},$$

Figure 1 demonstrates the concept of reverse causality and omitted-variables bias. An arrow in such a pictorial representation of a model would represent the causal effect of a one-unit change in the pointing variable on the pointed-to variable. The objective would be to estimate **A**, the average causal effect of the occupational-specific bonus on the probability that a serviceperson reenlists. We hope that $\hat{\beta_1}$ is an unbiased estimate of **A** in Figure 1. However, $\hat{\beta_1}$ captures all the reasons why the bonus and the retention decision might move together (or not), after adjusting for the factors in X.

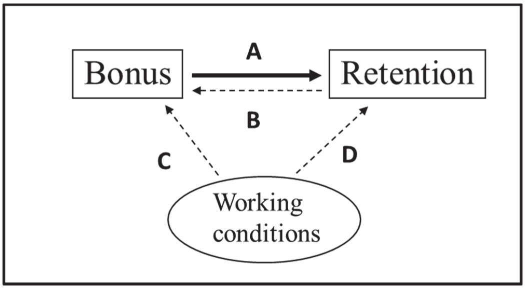


**Figure 1.** A visual representation of reverse causality.

To determine whether there is any potential bias, it does not require a formal theoretical model with assumptions on what factors affect what other factors. Rather, one would start by considering the

[2] The new source of bias on improper reference groups is available under the "eResources" tab at https://www.routledge.com/Regression-Analysis-A-Practical-Introduction-1st-Edition/Arkes/p/book/9781138541405 or at https://tinyurl.com/yytmnq65.

reasons why the bonus and retention variables move (or do not move) together, other than the bonus affecting the retention decision. This would also determine what needs to be controlled for.

The question to ask for reverse causality is whether the probability of reenlistment could affect the bonus, represented by the arrow labeled **B** in Figure 1. It very likely could, as a decrease in the probability of reenlistment for people in a certain occupation (due perhaps to increases in civilian labor market demand for the skill or increases in the deployment rates for the occupation) would cause the military service to have to increase the bonus; and an increase in the probability of reenlistment would allow the service to reduce the bonus.

Because **B** is likely negative, there would be a negative bias from the reverse causality on the estimated effect of the bonus on the probability of reenlistment. This bias would cause $\hat{\beta_1}$ to be lower than the value of **A** in Figure 1. (It requires much deeper and more-convoluted thought to determine the sign of the bias from an argument based on conditional mean dependence of the error term.) Thus, we would have an alternative story for why the estimated effect of the bonus is what it is—i.e., alternative to the causal-effects story. Attempts to address this with fixed effects would need to make sure that within the fixed-effects group, there still would not be any potential reverse causality (or omitted-variables bias).

For omitted-variables bias, in my experience, students have a hard time thinking of whether any variable might affect both the treatment and the outcome. Therefore, I found it to be more effective to use three steps: (a) What factors are the main drivers of why some have high vs. low (or 1 vs. 0) values of the treatment? (b) Which of those (if any) can you not adequately control for? (c) Could any of those factors affect the outcome beyond any effects through the treatment?

In Figure 1, we would need to think of what causes variation in the bonus in the sample. I imagine the list would include the occupation, the year (and factors specific to a given year, such as national economic conditions), the particular demand for the skills of servicepersons in a given occupation, and the working conditions for those in that occupation—such working conditions could change over time, and they would probably be rougher (say, more negative in theory) during periods of wars or increased deployments. All of these factors could affect the outcome beyond any effects through the bonus, and so if not controlled for, they would cause omitted-variables bias. I demonstrate this in Figure 1, with the omitted factor being working conditions for those in the occupation, using an oval to represent that we do not have a measure for it. Therefore, if we cannot adequately control for this, then better working conditions for an occupation in a given year would negatively affect the bonus (directly or indirectly through higher retention, leading to reverse causality) and positively affect retention, so **C** $< 0$ and **D** $> 0$. Thus, not adequately controlling for working conditions (and other things that could impact both the bonus and retention) for the occupation would lead to a negative omitted-variables bias for $\hat{\beta_1}$ (the product of **C** and **D** in Figure 1 would be negative). These are perhaps the most common sources of bias, and they follow directly from such a figure. However, there are other sources of bias, such as measurement error, that need to be considered.

Figure 2 demonstrates another type of omitted-variables bias based on the research issue of how the state unemployment rate (representing the strength of the labor market) affects marijuana use for teenagers. Therefore, **A** is the true average effect of a one-percentage-point increase in the state unemployment rate on the probability of a teenager using marijuana.

The problem is that whereas there is probably not any general factor that systematically affects both the state unemployment rate and teenage marijuana use, it still could be that states that have a higher general propensity to use marijuana (outside the influence of the economy) tend to have higher or lower unemployment rates, but not due to any systematic relationship. Therefore, whereas the occupational bonus and retention propensity for an occupation, in Figure 1, might have "spurious correlation" (due to a systematic relationship) contributing to why the variables move together, the state unemployment rate and propensity for teenage marijuana use might have "incidental correlation" that contributes to why they move together. If so, this would cause omitted-variables bias. (Line **C**, without arrows, is indicative of an incidental correlation that does not have an underlying systematic relationship).

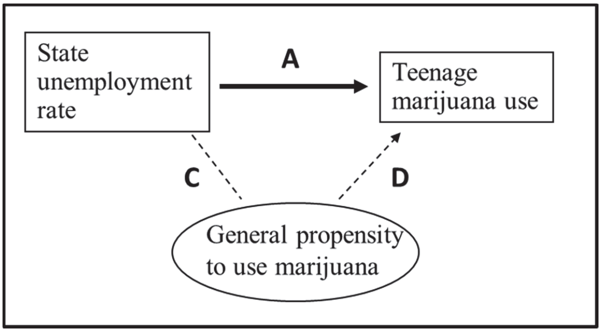


**Figure 2.** A visual representation of omitted-variables bias from incidental correlation.

Alternatively, you could put a specific state (say, California) as the potential omitted factor at the bottom of the figure and then have an arrow pointing from the California variable to both the unemployment rate (a positive effect, as California tends to have higher unemployment rates than the U.S.) and teen marijuana use (I am guessing positive). In this case, there would be positive omitted-variables bias from not controlling for California. For other states, it could be different. And, as a whole, it is quite possible that either the negative or positive biases could dominate the other, leading to a non-trivial bias in the estimate.

In this case, controlling for the states (with dummy variables or state fixed effects) would help towards addressing this problem. However, based on the concept mentioned at the end of Section 3, using fixed effects (or, controlling for a categorization) when the treatment has some error could cause greater bias from measurement error. In this case, a higher proportion of the usable variation in the state unemployment rate within states would be due to measurement error. (That is, in the ratio of variation due to measurement error divided by overall usable variation, state fixed effects reduces the denominator significantly but does not reduce the numerator.)

Let me give two tangents. First, omitted-variables bias is not a problem if the regression objective is forecasting or determining predictors of an outcome—this provides an example of how an understanding of regression objectives (recommendation A above) is important. Second, Arkes (2019) notes that the conventional definition of omitted-variables bias needs some modification. Note that in Figures 1 and 2, the correlation between the treatment and the omitted variable is based on the omitted variable affecting the treatment or incidental correlation. If, on the other hand, the omitted variable were a mediating factor and were affected by the treatment, then there would not be omitted-variables bias by excluding the variable; rather, there would be bias by *including* the variable. Therefore, the conventional definition that an omitted variable is correlated with the treatment and affects the outcome needs to add as a condition that the correlation is not solely due to the treatment affecting the omitted variable. Seeing this in a flow-chart demonstration can help with this concept.

In another visual lesson to recognize the direction of the biases, the bias from non-differential measurement error can be demonstrated with a simple bar graph of an outcome (say, income) for two groups (no-college-degree and college-degree). One can then easily see what would happen to the difference in income if people get randomly misclassified as to whether they have a college degree. The two averages would converge, and the estimated effect of a college degree would be biased downwards, at least from measurement error.

Finally, any teaching of how to recognize biases would be served well by having numerous examples to apply the concept to. This is consistent with the lessons from the book, *How Learning Works* (Ambrose et al. 2010), in which the authors argue that mastery of a subject requires much practice applying the topic and knowing when to and which topic to apply to a new situation. This is also consistent with higher levels of understanding, based on Bloom's Taxonomy. Furthermore, seeing research mistakes in action could provide meaningful lessons. And, dissecting media reports on research and gauging how trustworthy that research is (based on simply reading the media report) might be worthwhile for developing intuition on scrutinizing research.

D. Shift focus to the more practical quasi-experimental methods

This recommendation actually is in the same spirit as but diverges from the third of Angrist and Pischke (2017) recommendations. They espouse a shift in the focus of econometrics classes to randomized control trials (RCT) and quasi-experimental methods. One method they mention is the Regression-Discontinuity (RD) approach, which has appeared to become the new favorite approach for graduate students. This strategy parallels an earlier article (Angrist and Pischke 2010) and their book (Angrist and Pischke 2009).

However, I would argue there could be a better approach. The methods Angrist and Pischke espouse are more for academics who can search for randomness or a discontinuity and build a topic from that. It is not as effective for non-academics and other academics who are trying to address a specific policy question that will probably not afford the opportunity to apply an RCT or RD to the problems they are given. Furthermore, it limits the usefulness of economists. As Sims (2010) stated:

> "If applied economists narrow the focus of their research and critical reading to various forms of pseudo-experimental, the profession loses a good part of its ability to provide advice about the effects and uncertainties surrounding policy issues".

Sims (2010) also suggested that many of the quasi-experimental studies have limited scope with regards to the extrapolation of the results. This could occur, for example, due to non-linearities or just the nature of Local Average Treatment Effects that some quasi-experimental methods estimate. This further limits the usefulness of economics research.

Meanwhile, the less-complicated quasi-experimental methods might be more fruitful for most people conducting economic research and may be less limiting in extrapolation of the estimates. In particular, from my experiences in the non-academic world, a fixed-effects model is often the only plausible approach to addressing some potential sources of bias.

Given that the fixed-effects method would likely be a more useful tool than RD and other quasi-experimental methods, more emphasis should be placed on the nuances of fixed effects. These include many particulars of fixed effects that I wish I had learned in graduate school. For example:

(1) As described above in some things I had missed in my own research, bias from measurement error can be exacerbated by fixed effects;
(2) The estimated treatment effect with fixed effects is a weighted average of the estimated treatment effects within each fixed-effects group;
(3) The natural regression weights of the fixed-effects groups with a higher variance of the treatment are disproportionately higher—this concept and the correction is described in Gibbons et al. (2018) and Arkes (2019, Section 8.3); and so shifting the natural weight of a group could partly explain why fixed-effects estimates are different from the corresponding estimates without fixed effects; also besides fixed effects, this concept applies to cases in which one simply controls for categories (e.g., race). Reweighting observations can help address this problem.

Although the RCT is the most valid type of study, it is easy to analyze and few will have the resources to conduct one. The RD approach is a rare occurrence, and it is more for "finding a topic to use the method" than "finding a method for a topic". The fixed-effects method is much more widely used, and so shifting focus to the nuances of fixed effects would be more practical and useful to most students.

E. Add emphasis on interpretations on statistical significance and $p$-values

Perhaps the most important topic for which interpretations need to be taught better is on statistical significance and insignificance. In a recent article in *Nature*, Amrhein et al. (2019) call for an end to statistical significance and *p*-values and instead to use confidence intervals. There have been similar calls for teaching statistical analysis beyond the "*p*-value" approach over the last few decades—e.g., Gigerenzer (2004), Wasserstein and Wasserstein and Lazar (2016), and Wasserstein et al. (2019). Wasserstein et al. (2019) said: "Statistics education will require major changes at all levels to move to a post '$p < 0.05$' world". However, most textbooks continue to teach hypothesis tests based on the conventional approach that uses *p*-values.

I am aware of only two textbooks that discuss the problems of *p*-values and potential solutions: Paolella (2018) and Arkes (2019). Paolella (2018) points out all the problems with hypothesis tests, the *p*-value, and even the use of confidence intervals. He makes the point that hypothesis tests should not be used, but he notes how the *p*-values still might be useful. A single study with a low *p*-value provides little evidence for a theory or an empirical relationship. However, repeated studies with low *p*-values would provide stronger evidence. This confirms the value of the importance of replications. Furthermore, as Paolella argues, finding a *p*-value of 0.06 on a new drug that could cure cancer does not mean that society should discard any further research on the drug. Rather, the result should be interpreted as "something might be there" and it should be further investigated (Paolella 2018).

One area that could also use better instruction is on the various possible explanations for insignificance. Amrhein et al. (2019) find that over half of 791 articles across five journals made the mistake of interpreting insignificance as meaning that there is no effect—and these do not include the hot-hand studies. Aczel et al. (2018) find an even worse statistic for three leading psychology journals: 72% of 137 studies from 2015 with negative results had incorrect interpretations of those results. What highlights the problem with these interpretations is that an insignificant estimate may still provide more evidence for the alternative hypothesis than for the null (Aczel et al. 2018). Abadie (2020) makes the argument that an insignificant estimate might have more information than a significant estimate. In addition, there is always the possibility that a biased coefficient estimate has caused the insignificance; and a bias could cause significance when there is no causal effect. Furthermore, as described in Arkes (2019), if a treatment were to positively affect some and negatively affect others, then it could be that an insignificant effect is the average of these positive and negative effects that are, to some extent, cancelling each other out. Thus, it would be improper to conclude that the treatment has no effect based on an insignificant estimate. That said, a precisely estimated coefficient very close to zero ("precise nulls", as some call them), if free from potential biases, could mean that there is evidence for no meaningful *average* effect.

In light of the problems with the traditional *p*-value approach and the misinterpretations of insignificant estimates, lessons from Kass and Raftery (1995) or Startz (2014) on how to calculate posterior odds and on determining the most likely hypothesis would be useful components to the teaching on any statistical testing. Unfortunately, these often introduce an inconvenient vagueness in properly interpreting a hypothesis test. However, it is the proper approach to interpreting statistical tests. In addition, introducing the Bayesian critique should give the important lesson that strong conclusions on an empirical relationship should require quite high levels of significance.

Another important consideration for hypothesis tests would be the costs (loss) from a wrong conclusion. Therefore, such costs should be considered when determining the optimal significance level for the hypothesis test (Kim and Ji 2015; Kim 2020). Adjustments to the optimal significance level should be made for quite large samples. In addition, there should be some discussion on statistical significance for a meaningful effect size (rather than using zero as a baseline effect).

In the end, perhaps the post-($p < 0.05$) world should be one without hypothesis tests. Even the correct conclusions of "fail to reject" and "reject" (and not including "accept") come across as more conclusive than they actually are. And, they do not account for the potential biases and the practicality of the estimated relationship.

F. Advocate less complexity

The current go-to model for the Department of Defense (DoD) for evaluating the effects of various manpower policies (including bonuses) on retention is the Dynamic Retention Model (DRM). This is a complex model that only relatively recently has become able to be estimated, given the huge computing power it requires. Even though I had a (very) minor role in an application of it, I do not have a strong understanding of the model. And, my educated speculation is that no one at DoD funding such studies understands the model neither.

However, I do understand the model enough to know that the DRM is deficient in many ways, as described in Arkes et al. (2019). In retention models, the DRM estimates complex concepts, such as the discount rate and a taste-for-military parameter. However, it fails to control for basic factors that could partly address the reverse causality for bonuses I describe above, such as military occupation, fiscal year, and their interactions (Arkes 2018). Furthermore, the DRM will never be able to address the other problems noted above of measurement error and excess supply, as we only observe whether a person reenlists, not their willingness to reenlist. Thus, the DRM will probably not give a more reliable answer than the simpler and more-direct models. And, in my view, guesses from subject-matter-experts would be more reliable than what any model would tell us.

These empirical challenges are probably not well known to DoD officials. Therefore, they appear to be enamored by the complexity of the model. Some may put more faith in complex models. However, the simpler models are often more credible, as they rely on fewer assumptions.

One lesson may come from the history of instrumental-variables models. Early studies tended to not pay much attention to the validity of the instruments. For example, Sims (2010) noted that Ehrlich (1975) research on capital punishment lacked any discussion on the validity of the numerous instruments that were used, such as lagged endogenous variables. Later studies (e.g., Bound et al. 1995) noted the major problems with instrumental variables if assumptions were violated. This is an example of how the problems with complex models come out as people start understanding them better.

G. Add a simple ethical component

We conduct research to help inform society on the best public policies, health behaviors, business practices, and more. What we hope to see in others' research is the product of the optimal model they can develop, not the product of their efforts to find statistical significance. This means that our goal in conducting research should not be to find statistical significance, but rather to develop the best model to answer a research question and to give a responsible assessment of that model.

I recommend a few basic lessons in ethics (or good research practices). The first one would stress honesty in research and would give examples of when or how people might not be honest, such as with p-hacking. This could include some efforts to detect p-hacking, as described in Christensen and Miguel (2018). The second lesson would be the simple concept that "significance is not the goal of research". This is obvious to my students when they hear it (after they have taken other statistics and econometrics classes), but it is new to them and proves to be a valuable lesson. One student said, in an end-of-term reflection paper, that she had an insignificant estimate on her treatment variable in her thesis. She had the temptation to change the model to find significance, but she resisted that temptation based on this simple lesson that significance is not the goal. Other students, before hearing this lesson, tell me that something must be wrong with their model because their main coefficient estimate was insignificant. A simple statement on the order of "insignificant estimates are okay" might help change the culture. The third lesson in ethics would be on the importance of making responsible conclusions. This should involve being completely forthright about all potential pitfalls and biases to the coefficient estimates that could not be addressed and being careful with the conclusions on significance based on the Bayesian critique of $p$-values. This is important for society to properly synthesize the meaning and conclusions that can be drawn from a study. Overall, having textbooks incorporate lessons on the ethics of research might be a good step towards contributing to more honest research.

These lessons may also benefit from what Baicker et al. (2013) did for the study on how an expansion of Medicaid in Oregon affected health outcomes. They developed their model and published the research plan before implementing it. New resources, such as from the Center for Open Science, are promoting the online posting of research plans[3].

## 5. Implications for Undergraduate Econometrics

It follows logically that if my argument is correct that graduate econometrics training needs to be changed as I suggest, so too does undergraduate econometrics. Here is an equation from the textbook assigned in the undergraduate econometrics class I took many years ago, which remains in the current edition:

$$\hat{\beta_2} = \frac{\sum\ (y_i x_{2i})\left(\lambda^2 \sum\ x_{2i}^2 + \sum\ v_i^2\right) - (\lambda \sum\ y_i x_{2i} + \sum\ y_i v_i)\left(\lambda \sum\ x_{2i}^2\right)}{\sum\ x_{2i}^2\left(\lambda^2 \sum\ x_{2i}^2 + \sum\ v_i^2\right) - \left(\lambda \sum\ x_{2i}^2\right)^2}$$

It makes me wonder what would be a more efficient use of students' time: deciphering equations such as this or learning how to recognize biases.

One colleague said to me as I was writing my textbook, "Undergraduate econometrics is taught as if everyone will go on to a Ph.D. Economics program."I would take that statement further and argue that undergraduate econometrics is generally taught as if everyone will become an econometric theorist. However, few will.

To highlight how misguided it might be to use a high-level math approach rather than a more practical approach, consider these numbers. There are about 26,500 undergraduate economics majors per year (Stock 2017). And, according to the American Economic Association, there are about 1000 new Economics Ph.D.'s each year[4]. I will guess that no more than 10% of those Ph.D.'s become econometric theorists. There are also some Economics Ph.D. students who may not have had undergraduate econometrics. Therefore, less than 4% of undergraduate econometrics students end up receiving an Economics Ph.D., and easily less than 1% of them end up becoming econometric theorists.

Just as with graduate econometrics, I would agree with the first two of Angrist and Pischke (2017) recommended changes to undergraduate econometrics:

- replace the math with examples, which is a basic tenet of fostering student motivation (Ambrose et al. 2010)
- increase the emphasis on choosing the correct set of control variables.

However, I would argue that there is even less justification (than for graduate econometrics) for their third point (increase emphasis on RCT and quasi-experimental methods), at least for most quasi-experimental methods that have limited opportunities to be applied. More so than graduate students, few undergraduate econometric students will become academics, and so few will have the opportunity to search for randomness, valid instrumental variables, or discontinuities. Rather, they will mostly have to make the best of non-random data. Therefore, lessons should focus on developing skills for dealing with such data, understanding what the potential sources of bias are, figuring out how (if possible) to address the potential biases, and making responsible conclusions. These are the skills that will be needed for most people using regression analysis to try to solve problems. Learning how to conduct regression analysis without learning how to properly scrutinize a model and interpret results (in terms of causality and significance) has the potential to do more harm than good.

Table 2 is similar to Table 1, but it is for the top undergraduate textbooks, as used by Angrist and Pischke (2017), and with a few more I added. It shows, again, my estimate for the number of pages centered around a given topic. There is the same problem as with graduate textbooks that important concepts are not covered much or at all, while much space (and likely time in undergraduate classes) is

[3] See https://www.cos.io/our-services/prereg?_ga=2.152997817.1848170691.1585117163-115791253.1585117163.

[4] This comes from https://www.aeaweb.org/resources/students/careers/the-economics-profession.

devoted to concepts that are relatively minor for what would be useful to learn, in my view. Although the things that could cause bias in standard errors appear to have a large emphasis, there remains minimal coverage of things that could bias coefficient estimates. Furthermore, while those books that do discuss how to interpret an insignificant coefficient estimate do so correctly (albeit, briefly for each of them), it appears that only four of the eight books discuss it in the main discussion of hypothesis tests.

**Table 2.** What the main undergraduate textbooks teach (number of pages on a given topic).

| | **Kennedy (2008)** | **Gujarati and Porter (2009)** | **Studenmund (2010)** | **Baltagi (2011)** | **Wooldridge (2015)** | **Angrist and Pischke (2015)** | **Dougherty (2016)** | **Stock and Watson (2018)** |
|---|---|---|---|---|---|---|---|---|
| **Causes of bias in the standard errors** | | | | | | | | |
| Heteroskedasticity | 5 | 47 | 1 | 12 | 28 | 1 | 17 | 5 |
| Multicollinearity | 10 | 31 | 28 | 3 | 5 | 0 | 9 | 3 |
| **Causes of bias in the coefficient estimates** | | | | | | | | |
| Simultaneity | 18 | 32 | 27 | 24 | 20 | 0 | 4 | 4 |
| Omitted-variables bias | 2 | 6 | 8 | 0 | 5 | 13 | 9 | 9 |
| Measurement error | 7 | 5 | 4 | 1 | 7 | 9 | 7 | 3 |
| Mediating factors | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 |
| **Other important topics** | | | | | | | | |
| Holding other factors constant | 2 | 5 | 5 | 2 | 9 | 11 | 2 | 1 |
| Fixed effects | 6 | 15 | 0 | 3 | 8 | 0 | 6 | 12 |
| Bayesian critique of $p$-values | 19 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Correctly indicates an insignificant coef. estimate does not mean accept the null | N/A | Yes | Yes (in a footnote) | N/A | Yes | N/A | Yes | N/A |

Note: "N/A" in the last row indicates that I do not believe the topic of how to interpret insignificant estimates is discussed.

With a more practical approach, there can be useful lessons that would actually be applicable for most of the undergraduate econometrics students. I grant that not all undergraduate econometrics students will have a job using econometrics. However, perhaps the greater skill they should come away with is the ability to recognize sources of bias. This could help them understand why correlation could but does not always mean causation. It could help them understand other important statistical concepts such as omitted-variables bias and Type I errors, both of which have applications to many workplace situations. Learning about biases could help engender a healthy skepticism in the statistics and research they hear about every day. And, these are skills that could form the foundation for more efficient learning in graduate econometrics, for those who take that route.

## 6. Conclusions and Topics for Further Discussion

The goal we should have as econometrics instructors is to teach the skills that would encourage solid, honest, and responsible research that can help improve the world. Being able to have a voice for improving the world requires trust that what we produce is valid. Therefore, efforts in instruction should foster honesty, responsibility, and the skills and research practices that produce valid research.

This means that we need to assess what concepts and what methods of instruction are the most important for producing solid researchers. Based on this idea, I have made the case, building on Angrist and Pischke (2017), that we need to shift emphases.

The teaching of graduate (and undergraduate) econometrics needs to be revamped. As instructors, we need to think about what most students will be doing with their skills, what are the most practical lessons from econometrics, what potential problems are most likely to affect the validity of a study, and how do we produce ethical and responsible researchers.

Not everyone is going to be an academic with the freedom to search the world for random assignment and choose their own topics from the randomness they find. Rather, most students will become non-academic practitioners who will need to address important problems with data that do not have random assignment. Their task would be to recognize the potential sources of bias, design the optimal method to address the issue, choose the optimal set of control variables, recognize the remaining sources of bias that could not be addressed, and make responsible conclusions. Or, as consumers of research, they should have the tools to recognize biases, which can also apply to the everyday statistics they hear in the news or even properly assessing events by considering alternative stories that could explain why two variable move (or do not move) together. These are the things that econometrics courses should be aimed towards, both at the graduate and undergraduate levels.

Certainly, such shifts would impact certain fields in which there would be methods particular to that field, such as Macroeconomics. And, anyone studying econometric theory would need a new course on the high-level math underlying econometrics. However, such shifts would make sense to spend class and student-studying time more efficiently, avoiding spending time on field-specific methods or the high-level math for students who would never need such material. Furthermore, a class that spent more time giving examples that demonstrate the nuances of certain methods should help students better understand the mathematical theory behind the models.

Let me end by calling for a larger assessment of what skills Ph.D. economists need in their research. Would most Ph.D. students benefit from a shift in focus from the high-level math to something more practical? Should basic graduate econometrics be any different from undergraduate econometrics? For what I believe is a large share of Ph.D. economists, two good low-math undergraduate courses (that incorporate the changes I describe above), along with applied graduate courses and plenty of practice, should be sufficient to prepare them to become successful researchers. Based on my experiences at research organizations and in academia, I believe that these lessons would have been sufficient for most of my colleagues. The redesign and shifts that I have discussed, as I have argued in this article, would have helped me avoid most of my research mistakes.

**Funding:** This research received no external funding.

**Acknowledgments:** I would like to thank Thomas Ahn, Jiadi Chen, Judith Hermis, Ercio Munoz, Daniel Stone, and anonymous referees for their helpful comments.

**Conflicts of Interest:** The strategy for the teaching of econometrics that I espouse in this article is consistent with much of (although not all of) the teachings in my own textbook. There are ideas in this paper that go beyond what is in my textbook, and at least one idea that is inconsistent with my textbook. If this article were to lead to more book sales, there would be a very-minor financial benefit. That said, my views expressed here were based on: (1) my perspectives that shaped my textbook; and (2) other of my perspectives that have evolved since the publication of my textbook.

## References

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

Aczel, Balazs, Bence Palfi, Aba Szollosi, Marton Kovacs, Barnabas Szaszi, Peter Szecsi, Mark Zrubka, Quentin F. Gronau, Don van den Bergh, and Eric-Jan Wagenmakers. 2018. Quantifying support for the null hypothesis in psychology: An empirical investigation. *Advances in Methods and Practices in Psychological Science* 1: 357–66. [CrossRef]

Ambrose, Susan A., Michael W. Bridges, Michele DiPietro, Marsha C. Lovett, and Marie K. Norman. 2010. *How Learning Works: Seven Research-Based Principles for Smart Teaching*. Hoboken: John Wiley & Sons.

Amrhein, Valentin, Sander Greenland, and Blake McShane. 2019. Scientists rise up against statistical significance. *Nature* 567: 305–7. [CrossRef] [PubMed]

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2010. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives* 24: 3–30. [CrossRef]

Angrist, Joshua D., and Jörn-Steffen Pischke. 2015. *Mastering Metrics: The Path from Cause to Effect*. Princeton: Princeton University Press.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2017. Undergraduate econometrics instruction: Through our classes, darkly. *Journal of Economic Perspectives* 31: 125–44. [CrossRef]

Arkes, Jeremy. 2010. Revisiting the hot hand theory with free throw data in a multivariate framework. *Journal of Quantitative Analysis in Sports* 6. [CrossRef]

Arkes, Jeremy. 2018. Empirical biases and some remedies in estimating the effects of selective reenlistment bonuses on reenlistment rates. *Defence and Peace Economics* 29: 475–502. [CrossRef]

Arkes, Jeremy. 2019. *Regression Analysis: A Practical Introduction*. Oxford: Routledge.

Arkes, Jeremy, Thomas Ahn, Amilcar Menichini, and William Gates. 2019. *Retention Analysis Model (RAM) for Navy Manpower Analysis*. Report NPS-GSBPP-19-003. Monterey: Naval Postgraduate School.

Baicker, Katherine, Sarah L. Taubman, Heidi L. Allen, Mira Bernstein, Jonathan H. Gruber, Joseph P. Newhouse, Eric C. Schneider, Bill J. Wright, Alan M. Zaslavsky, and Amy N. Finkelstein. 2013. The Oregon experiment—effects of Medicaid on clinical outcomes. *New England Journal of Medicine* 368: 1713–22. [CrossRef]

Baltagi, Badi H. 2011. *Econometrics*, 3rd ed. Heildelberg: Springer.

Bania, Neil, Jo Anna Gray, and Joe A. Stone. 2007. Growth, Taxes, and Government Expenditures: Growth Hills for U.S. States. *National Tax Journal* 60: 193–204. [CrossRef]

Bar-Eli, Michael, Simcha Avugos, and Markus Raab. 2006. Twenty years of "hot hand" research: Review and critique. *Psychology of Sport and Exercise* 7: 525–53. [CrossRef]

Bartik, Timothy J. 1991. *Who Benefits from State and Local Economic Development Policies?* Kalamazoo: W.E. Upjohn Institute for Employment Research.

Bocskocsky, Andrew, John Ezekowitz, and Carolyn Stein. 2014. The hot hand: A new approach to an old 'fallacy'. Paper presented at 8th Annual MIT Sloan Sports Analytics Conference, Boston, MA, USA, 28 February 2014–1 March 2014; pp. 1–10.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 90: 443–50. [CrossRef]

Brooks, David. 2013. The Philosophy of Data. *New York Times*. February 4. Available online: http://www.nytimes.com/2013/02/05/opinion/brooks-the-philosophy-of-data.html (accessed on 6 November 2015).

Christensen, Garret, and Edward Miguel. 2018. Transparency, Reproducibility, and the Credibility of Economics Research. *Journal of Economic Literature* 56: 920–80. [CrossRef]

Cready, William M., Bo Liu, and Di Wang. 2019. A Content Based Assessment of the Relative Quality of Leading Accounting Journals. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3438405 (accessed on 4 September 2020).

Cunningham, Scott. 2018. Causal Inference: The MIX Tape (v. 1.7). Available online: tufte-latex.googlecode.com (accessed on 4 September 2020).

Dougherty, Christopher. 2016. *Introduction to Econometrics*. New York: Oxford University Press.

Ehrlich, Isaac. 1975. The deterrent effect of capital punishment: A question of life and death. *American Economic Review* 65: 397–417.

Ettinger, Bruce, Gary D. Friedman, Trudy Bush, and Charles P. Quesenberry Jr. 1996. Reduced mortality associated with long-term postmenopausal estrogen therapy. *Obstetrics & Gynecology* 87: 6–12.

Funderburg, Richard, Timothy J. Bartik, Alan H. Peters, and Peter S. Fisher. 2013. The impact of marginal business taxes on state manufacturing. *Journal of Regional Science* 53: 557–82. [CrossRef]

Gertler, Paul. 2004. Do conditional cash transfers improve child health? Evidence from PROGRESA's control randomized experiment. *American Economic Review* 94: 336–41. [CrossRef]

Gibbons, Charles E., Juan Carlos Suárez Serrato, and Michael B. Urbancic. 2018. Broken or fixed effects? *Journal of Econometric Methods* 8. [CrossRef]

Gigerenzer, Gerd. 2004. Mindless statistics. *The Journal of Socio-Economics* 33: 587–606. [CrossRef]

Gill, D. 2018. Why It's so Hard to Study the Impact of Minimum Wage Increases. Available online: https://qz.com/work/1415401/why-minimum-wage-research-is-full-of-conflicting-studies/ (accessed on 11 June 2020).

Gilovich, Thomas, Robert Vallone, and Amos Tversky. 1985. The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology* 17: 295–314. [CrossRef]

Goldberg, Matthew S. 2001. *A Survey of Enlisted Retention: Models and Findings*. Report # CRM D0004085.A2. Alexandria: Center for Naval Analyses.

Goldberger, Arthur Stanley. 1991. *A Course in Econometrics*. Harvard: Harvard University Press.

Gould, Stephen Jay. 1989. The streak of streaks. *Chance* 2: 10–16. [CrossRef]

Greene, W. 2012. *Econometric Analysis*, 7th ed. Upper Saddle River: Prentice Hall.

Gujarati, Damodar N., and D. Porter. 2009. *Basic Econometrics*. New York: McGraw-Hill.

Hansen, Michael L., and Jennie W. Wenger. 2005. Is the pay responsiveness of enlisted personnel decreasing? *Defence and Peace Economics* 16: 29–43. [CrossRef]

Harford, Tim. 2014. Big data: A big mistake? *Significance* 11: 14–19. [CrossRef]

Hayashi, Fumio. 2000. *Econometrics*. Princeton: Princeton University Press.

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

Ioannidis, John P. A. 2005. Why Most Published Research Findings Are False. *PLoS Medicine* 2: e124. [CrossRef]

Kahneman, Daniel. 2011. *Thinking Fast and Slow*. New York: Farrar, Straus and Giroux.

Kass, Robert E., and Adrian E. Raftery. 1995. Bayes factors. *Journal of the American Statistical Association* 90: 773–95. [CrossRef]

Kennedy, Peter. 2008. *A Guide to Econometrics*, 6th ed. Hoboken: Wiley-Blackwell.

Kim, Jae H. 2020. Decision-theoretic hypothesis testing: A primer with R package OptSig. *The American Statistician*. in press. [CrossRef]

Kim, Jae H., and Philip Inyeob Ji. 2015. Significance testing in empirical finance: A critical review and assessment. *Journal of Empirical Finance* 3: 1–14. [CrossRef]

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. 2007. Experimental analysis of neighborhood effects. *Econometrica* 75: 83–119. [CrossRef]

Leamer, Edward E. 1983. Let's take the con out of econometrics. *Modelling Economic Series* 73: 31–43.

Leamer, Edward E. 1988. 3 Things that bother me. *Economic Record* 64: 331–35. [CrossRef]

Manson, JoAnn E., Aaron K. Aragaki, Jacques E. Rossouw, Garnet L. Anderson, Ross L. Prentice, Andrea Z. LaCroix, Rowan T. Chlebowski, Barbara V. Howard, Cynthia A. Thomson, Karen L. Margolis, and et al. 2017. Menopausal Hormone Therapy and Long-term All-Cause and Cause-Specific Mortality: The Women's Health Initiative Randomized Trials. *Journal of the American Medical Association* 318: 927–38. [CrossRef]

Miller, Joshua B., and Adam Sanjurjo. 2018. Surprised by the hot hand fallacy? A truth in the law of small numbers. *Econometrica* 86: 2019–47. [CrossRef]

Mofidi, Alaeddin, and Joe A. Stone. 1990. Do state and local taxes affect economic growth? *The Review of Economics and Statistics* 72: 686–91. [CrossRef]

Nuzzo, Regina. 2014. Statistical errors. *Nature* 506: 150–52. [CrossRef] [PubMed]

Paolella, Marc S. 2018. *Fundamental Statistical Inference: A Computational Approach*. Hoboken: John Wiley & Sons, vol. 216.

Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.

Petersen, Mitchell A. 2009. Estimating standard errors in finance panel data sets: Comparing approaches. *The Review of Financial Studies* 22: 435–80. [CrossRef]

Poulson, Barry W., and Jules Gordon Kaplan. 2008. State Income Taxes and Economic Growth. *Cato Journal* 28: 53–71.

Reed, W. Robert. 2008. The robust relationship between taxes and US state income growth. *National Tax Journal* 61: 57–80. [CrossRef]

Rossouw, Jacques E., Garnet L. Anderson, Ross L. Prentice, Andrea Z. LaCroix, Charles Kooperberg, Marcia L. Stefanick, Rebecca D. Jackson, Shirley A. A. Beresford, Barbara V. Howard, Karen C. Johnson, and et al. 2002. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: Principal results from the Women's Health Initiative randomized controlled trial. *JAMA* 288: 321–33.

Russell, Davidson, and James G. MacKinnon. 2004. *Econometric Theory and Methods*. New York: Oxford University Press.

Sims, Christopher A. 2010. But economics is not an experimental science. *Journal of Economic Perspectives* 24: 59–68. [CrossRef]

Startz, Richard. 2014. Choosing the More Likely Hypothesis. *Foundations and Trends in Econometrics* 7: 119–89. [CrossRef]

Stock, Wendy A. 2017. Trends in economics and other undergraduate majors. *American Economic Review* 107: 644–49. [CrossRef]

Stock, James H., and Mark W. Watson. 2018. *Introduction to Econometrics*, 4th ed. London: Pearson.

Stone, Daniel F. 2012. Measurement error and the hot hand. *The American Statistician* 66: 61–66. [CrossRef]

Studenmund, A. H. 2010. *Using Econometrics: A Practical Guide*. London: Pearson Education.

Thaler, Richard H., and Cass R. Sunstein. 2009. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New York: Penguin.

Wasserstein, Ronald L., and Nicole A. Lazar. 2016. The ASA statement on p-values: Context, process, and purpose. *The American Statistician* 70: 129–33. [CrossRef]

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

Wasylenko, Michael, and Therese McGuire. 1985. Jobs and taxes: The effect of business climate on states' employment growth rates. *National Tax Journal* 38: 497–511.

Wooldridge, Jeffrey M. 2012. *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press.

Wooldridge, Jeffrey M. 2015. *Introductory Econometrics: A Modern Approach*. Toronto: Nelson Education.

Yeoh, Melissa, and Dean Stansel. 2013. Is Public Expenditure Productive: Evidence from the Manufacturing Sector in US Cities, 1880–1920. *Cato Journal* 33: 1.

*econometrics*

*Article*

# The Replication Crisis as Market Failure

**John Quiggin**

School of Economics, The University of Queensland, St Lucia QLD 4072, Australia; j.quiggin@uq.edu.au

Received: 30 September 2019; Accepted: 18 November 2019; Published: 24 November 2019

**Abstract:** This paper begins with the observation that the constrained maximisation central to model estimation and hypothesis testing may be interpreted as a kind of profit maximisation. The output of estimation is a model that maximises some measure of model fit, subject to costs that may be interpreted as the shadow price of constraints imposed on the model. The replication crisis may be regarded as a market failure in which the price of "significant" results is lower than would be socially optimal.

**Keywords:** replication crisis; profit maximization; market failure

## 1. Introduction

Concerns about the inadequacy of standard practices in statistics and econometrics have been long-standing. Since the 1980s, criticisms of econometric practice, including those of Leamer (1983); Lovell (1983); McCloskey (1985) have given rise to a large literature. Kim and Ji (2015) provide a survey. Even more significant, within the economics profession was the Lucas (1976) critique of Keynesian economic modelling Lucas and Sargent (1981).

Most of the concerns raised in these critiques apply with equal force to other social science disciplines and to fields such as public health and medical science. Ioannidis (2005) offered a particularly trenchant critique, concluding that "most published research findings are false".

The emergence of the "replication crisis", first in psychology and then in other disciplines, has attracted the broader public to some of these concerns. The applicability of the term "crisis" is largely due to the fact that, unlike previous debates of this kind, concern over replication failures has spilled over disciplinary boundaries and into the awareness of the educated public at large.

The simplest form of the replication crisis arises from the publication of a study suggesting the existence of a causal relationship between an outcome of interest $y$ and a previously unconsidered explanatory variable $x$ followed by studies with a similar design that fail to find such a relationship.

Most commonly, this process takes place in the context of classical inference. In this framework, the crucial step is the rejection of a null hypothesis of no effect, with some specified level of significance, typically 95 per cent or 90 per cent. In the most commonly used definition, a replication failure arises when a subsequent study testing the same hypothesis with similar methods but with a different population fails to reject the null.[1]

For example, Kosfeld et al. (2005) found that exposure to oxytocin increases trust in humans. This finding created substantial interest in possible uses of oxytocin to change mood, potentially for malign as well as benign purposes. Similar results were published by Mikolajczak et al. (2010). However, a subsequent attempt at replication by the same research team, Lane et al. (2015), was unsuccessful.

[1] As an anonymous referee points out, this characterisation of replication failure is too stringent since some failures to reject the null are to be expected. More stringent definitions of replication failure are that the null is not rejected using the pooled data or that the parameter estimates from the two studies are (statistically) significantly different.

The crisis was brought to the wider public's attention by the publication by Open Science Collaboration (2015) of a systematic attempt to replicate 100 experimental results published in in three psychology journals. Replication effects were half the magnitude of original effects, representing a substantial decline. Whereas ninety-seven percent of original studies had statistically significant results, only thirty-six percent of replications did.

A variety of responses have been offered in response to the replication crisis. These include tightening the default P-value threshold to 0.005 (Benjamin et al. 2018), procedural improvements such as the maintenance of data sets and the preregistration of hypotheses ((Nosek and Lakens 2014), attempts to improve statistical practice within the classical framework, for example, through bootstrapping (Lubke and Campbell 2016)), and the suggestion that Bayesian approaches might be less vulnerable to these problems (Gelman 2015).

This paper begins with the observation that the constrained maximisation central to model estimation and hypothesis testing may be interpreted as a kind of profit maximisation. The output of estimation is a model that maximises some measure of model fit, subject to costs that may be interpreted as the shadow price of constraints imposed on the model. This approach recalls the observation of Johnstone (1988) "That research workers in applied fields continue to use significance tests routinely may be explained by forces of supply and demand in the market for statistical evidence, where the commodity traded is not so much evidence, but "statistical significance."[2]

In mainstream economics, an unsatisfactory market outcome is taken as prima facie evidence of a "market failure", in which prices are not equal to social opportunity costs.[3]

In this paper, we will consider the extent to which the replication crisis, along with broader problems in statistical and econometric practice, may be seen as a failure in the market that generates published research.

## 2. Model Selection as an Optimisation Problem

In the general case, we consider data $(X, y)$ where $y$ is the variable (or vector of variables) of interest, and $X$ is a set of potential explanatory variables. We consider a finite set of models $\mathcal{M}$, with typical element $m$. The set of models may be partitioned into classes $\mathcal{M}_\kappa$, where $\kappa = 1 \ldots K$. Typically, although not invariably, lower values of $\kappa$ correspond to more parsimonious and, therefore, more preferred models.

For a given model $m$, the object of estimation is to choose parameters $\beta^* (m; X, y)$ to maximise a value function $V(\beta; X, y)$, such as log likelihood or explained sum of squares. Define

$$V^* (m; X, y) = \max_\beta V(\beta; X, y) = V(\beta^* (m; X, y); X, y). \tag{1}$$

The model selection problem is to choose $m$ to maximise the global objective function

$$\Pi (m; X, y) = V^* (m; X, y) - C(m), \tag{2}$$

where $C(m)$ is a cost function. Given the interpretation of $V$ as a value function and $C$ as a cost function, $\Pi$ may be regarded as a profit function.

[2] I am indebted to an anonymous referee for pointing out this article, foreshadowing my central point.

[3] Quiggin (2019) extends this analysis to encompass issues such as unemployment and inequality.

### 2.1. Linear Case

We will confine our attention to linear models with a single variable of interest $y$ and $N$ observations on a set of $K$ potential explanatory variables $X = (x_1 ... x_K)$. The generic model of this form is

$$\begin{aligned} Y &= X\beta + \varepsilon \\ R\beta &= v, \end{aligned} \tag{3}$$

where

$Y$ is an $N \times 1$ vector of observations on $y$;
$\mathbf{X}$ is an $N \times K$ matrix of observations on $X$;
$\beta$ is a $K \times 1$ vector of parameters;
$\varepsilon$ is an $N \times 1$ error term;
$R$ is a $J \times K$ vector of constraints, where $J < K$ and $R$ has full rank;
the special case of ordinary least squares (OLS) is that of $k$ unconstrained explanatory variables.

In this case, $J = K - k$ and $R = \begin{pmatrix} \mathbf{0}_k \\ \mathbf{I}_{K-k} \end{pmatrix}$. The model may be written without explicit constraints as

$$Y = \mathbf{X}\beta + \varepsilon, \tag{4}$$

where

$Y$ is an $N \times 1$ vector; of observations on $y$
$\mathbf{X}$ is an $N \times k$ matrix of observations on $(x_1 ... x_k)$;
$\beta$ is a $k \times 1$ vector of parameters;
$\varepsilon$ is an $N \times 1$ error term, distributed as iid.$N(0, \sigma)$

### 2.2. Value Functions, Cost Functions, and Profit Functions

#### 2.2.1. Value

The most commonly used value measures are
(i) measures related to likelihood,

$$\mathcal{L} = \prod_n p\left(y_n | \beta\right); \tag{5}$$

(ii) those based on the proportion of variation in $y$ explained by the model. The canonical measure is

$$R^2 = \frac{RSS}{TSS}. \tag{6}$$

These value measures may be interpreted in terms of explanation and prediction. If the objective of the model is taken to be the explanation of observed outcomes in terms of a given model, the best explanation may be seen as the choice of $\beta$ that maximises the likelihood of the observed $y$. If the objective of the model is taken to be prediction of $y$ given $X$, $R^2$ is the value function implied by the minimisation of a quadratic loss function.

#### 2.2.2. Cost

The simplest cost functions $C$ are monotonic functions of $k$, the number of non-zero parameters in the model. More generally, we may attach a cost $c$ to the relaxation of a constraint $R\beta = 0$ suggested by theory.

Any cost function gives rise to a partition of the set of models into classes $\mathcal{M}_\kappa$ where $\kappa = 1 \ldots K$, where the equivalence relation is defined by the property that if $\kappa(m) = \kappa(m')$, $C(m) = C(m')$.

Standard choices include

$$\begin{aligned} C(m) &= k, \\ C(m) &= \frac{k-1}{N-k-1}, \\ C(m) &= K - J. \end{aligned} \tag{7}$$

2.2.3. Profit

In a general setting, profit may be defined as

$$\Pi = py - wx, \tag{8}$$

where $y$ is an output measure, $p$ is the output price, $x$ is an input measure, and $w$ is the wage or factor price. Treating output as numeraire, we may simplify to

$$\Pi = y - wx. \tag{9}$$

Given the interpretation above, profit maximisation is represented by the choice of the best-fitting model, subject to a cost associated with rejecting restrictions suggested by theory or by concern about Type 1 error.

As in a market setting, profit maximisation will be consistent with a broader notion of optimality if $y$ and $x$ are measured correctly and if the price $w$ represents social cost.

## 3. Model Selection and Hypothesis Testing

To examine problems of hypothesis testing and model selection, we introduce an adjacency relationship between models. Informally, two models are adjacent if they differ by a single step such as the inclusion of an additional variable or the imposition of a linear restriction. As these examples indicate, it will typically be the case that adjacent models $m, m'$ are ranked by parsimony so that, if $m$ is derived by imposing a restriction in $m'$, then $m'$ $\kappa(m) < \kappa(m')$, and there exists no $m''$ such that $\kappa(m) < \kappa(m'') < \kappa(m')$. The adjacency relation is directed and will be denoted by $m \to m'$ (for the case $\kappa(m) < \kappa(m')$).

### *3.1. Global Model Selection Criteria*

Many widely used model selection criteria may be interpreted as profit functions, with a reversal of sign to make the problem one of maximisation rather than minimisation

Examples include the Akaike information criterion (AIC),

$$-AIC = 2\log(\mathcal{L}) - 2k, \tag{10}$$

the corrected AIC,

$$-AICc = AIC - \frac{2k^2 + 2k}{N-k-1}, \tag{11}$$

and the Bayesian information criterion (BIC)

$$-BIC = 2\log(\mathcal{L}) - \log(N)\,k. \tag{12}$$

It is also possible to include a cost function in measures based on $R^2$.

For example, we may write

$$V = R^2 - \frac{k-1}{N-k-1}. \tag{13}$$

Closely related is the $\bar{R}^2$ criterion, which satisfies

$$1-\bar{R}^2=\left(1-R^2\right)\frac{N-1}{N-k-1} \tag{14}$$

or

$$\bar{R}^2=\frac{ESS/\left(k-1\right)}{RSS/\left(N-k-1\right)}. \tag{15}$$

### 3.2. Local Model Selection Criteria

**Definition 1.** *Let $m\rightarrow m'$ be adjacent models. Then $m'$ is a stepwise profit improvement (reduction) over $m$ if $\Pi\left(m';X,y\right)>\left(<\right)\Pi\left(m';X,y\right).$*

This definition immediately suggests consideration of the stepwise regression algorithm proposed by Efroymson (1960). Let the profit function $\Pi$ be $\bar{R}^2$. From an initial $m$, choose the adjacent $m'$ that maximises $\Pi\left(m';X,y\right)$ terminating when, for all adjacent $m'', \Pi\left(m';X,y\right)\geq\Pi\left(m'';X,y\right).$ Forward selection adds the requirement $m\rightarrow m'$ so that selection proceeds by adding variables. Conversely, backward selection requires $m'\rightarrow m$.

### 3.3. Hypothesis Testing

The problems of model selection and hypothesis testing are commonly treated separately. From the viewpoint offered here, they may be regarded as global and local versions of the same problem, that of profit maximisation. This point may be illustrated with respect to the large class of classical hypothesis tests encompassed by the Wald, likelihood ratio, and Lagrange multiplier statistics Engle (1984), and also with respect to the $t$ and $F$ tests used in hypothesis testing in the OLS framework.

Model selection may be considered locally in terms of a pairwise choice between models, so that $m$ is preferred to $m'$ if and only if $\Pi\left(m;X,y\right)>\Pi\left(m';X,y\right)$, which may be written as

$$V^*\left(m;X,y\right)-C\left(m\right)>V^*\left(m';X,y\right)-C\left(m'\right) \tag{16}$$

or

$$V^*\left(m;X,y\right)-V^*\left(m';X,y\right)>C\left(m\right)-C\left(m'\right). \tag{17}$$

If $V\left(m;X,y\right)=\log\left(\mathcal{L}\right)$, the left-hand side of (16) and (17) is a log likelihood ratio. If $m'$ is obtained from $m$ by the imposition of a vector of $J$ constraints $R\beta=v$, then under the standard assumptions of the general linear model, the Likelihood Ratio (LR) statistic is distributed as $\chi^2$ with $J$ degrees of freedom in the limit. Hence, if we set $C\left(m'\right)=0$ and $C\left(m'\right)$ equal to a suitably chosen critical value for $\chi^2\left(J\right)$, we obtain the usual LR test.

Compare this approach to the various information criteria discussed in the previous section. The domains of the two approaches coincide in the case of nested models where the restrictions imposed in $m'$ consist of setting some coefficients equal to zero. The canonical criteria set out above have in common that more parsimonious models are always preferred, at least weakly. In formal terms, the classical hypothesis-testing framework shares this characteristic. The only difference is in the choice of cost function.

The same point may be made with respect to value functions based on explained variation. The natural starting point is the $F$-test,

$$F=\frac{\left(RSS\left(m\right)-RSS\left(m'\right)\right)/J}{RSS\left(m\right)/\left(N-k\right)}. \tag{18}$$

Alternatively, we may consider the framework of constrained optimisation. The objective for a linear model $m$ may be written as

$$V(m; X, y) - \lambda (R\beta - v), \tag{19}$$

where $\lambda$ is a vector of Lagrange multipliers, such that $\lambda (R\beta - v) = 0$. In the context of constrained optimisation, as observed by Breusch and Pagan (1980), the Lagrange multiplier represents the shadow price of relaxing a constraint.

The most usual restriction is that of setting the coefficient on some variable equal to zero. In this case, the $F$ statistic is typically replaced by its square root, and the test is a $t$-statistic.

Such tests of significance give rise to an estimation strategy in which variables are included sequentially in the model. The most usual criterion is the $t$-statistic associated with the added variable, typically with a stopping criterion associated with 5 per cent significance.

As Dhrymes (1970) observes, this strategy is closely related to the criterion of maximising $\bar{R}^2$. However, $\bar{R}^2$ is (locally) optimised by including variables whenever their $t$-statistic is greater than one and, therefore, implies a lower price for including variables. A similar point may be made with respect to the relationship between information criteria such as the AIC and the likelihood ratio test.

## 4. Market Failure

When the model selection problem is interpreted as one of profit maximisation, it is natural to interpret problems within econometric practice as market failures. In particular, the profit function being maximised by researchers differs from that which would be suggested by an objective of social welfare maximisation.

In relation to the replication crisis, the core of the problem, as it is generally perceived, is that the price of including a variable of interest as statistically or economically significant in the reported model is too low.

On the one hand, the bias against publication of negative results means that the benefit to researchers of reporting models with no significant variables of interest is limited and, in many fields, close to zero.

On the other hand, the availability of the set of techniques pejoratively referred to as "p-hacking" means that the test statistics reported in published studies are more likely to arise through chance than would be suggested by the classical hypothesis-testing framework.

Taken together, these observations suggest that the price to researchers of including a variable of interest in a published model is lower than would be socially optimal. As a result, too many positive results are reported.

But what is the socially optimal price? The classical hypothesis-testing framework is of limited value here. Informal reasoning about Type 1 and Type 2 errors implies that there must exist a trade-off that can be expressed in terms of relative prices. But the relationship between that trade-off and the notion of statistical significance is opaque, to say the least. Holding the size (likelihood of Type 1 error) constant means that the power (likelihood of Type 2 error) depends on a combination of sample size and the variance of the error term in a way that bears no obvious relationship to the relative desirability of avoiding the two types of error.

Bayesian approaches based on loss functions are much closer to the spirit of the approach being suggested here. The main difference between the profit maximisation approach proposed here and the loss function approach (apart from a trivial change of sign) is the inclusion of an explicit cost associated with the estimation of a more general model.

The central point of the market failure analogy is that excessive publication of fragile results implies that the price of a positive result is too low. As with other market failures, a solution may be sought either in regulation or in pricing.

Regulatory approaches include measures such as the preregistration of estimation strategies. To develop an appropriate model of socially optimal prices, we must consider how, as a society, we respond to research publications.

This is, in some sense, an equilibrium outcome. If we accept at face value either the classical interpretation of a finding of statistical significance or its natural subjective misinterpretation (the probability that the variable in question is causally related to the variable of interest is near one), then the fact that many findings cannot be replicated is a major social problem. The initial result would lead us to act on a belief that cannot be substantiated.

Such an outcome cannot be sustained, however. Most people now understand that a reported result, whatever the supposed statistical significance, is not conclusive in the absence of replication, and certainly not at the stated level of significance. The question, therefore, is whether the publication of the result is, on balance, socially beneficial.

In this view, the most important requirement is that the price should be high enough to offset the inevitable effects of publication bias. Publication of a result showing a statistically significant result should be treated as an indication that further research is warranted, rather than conclusive or even highly probable evidence that the reported relationship is real.

## 5. Concluding Comments

Statistical research is a social and economic enterprise aimed at discovering, testing, and ultimately acting on relationships between variables of interest. The economic concepts of profit maximisation and market failure provide a way of thinking about statistical research that reflects its social role.

**Funding:** This research received no external funding.

**Conflicts of Interest:** The author declares no conflicts of interest.

## References

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E. -J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, and Colin Camerer. 2018. Redefine statistical significance. *Nature Human Behaviour* 2: 6–10. [CrossRef] [PubMed]

Breusch, Trevor S., and Adrian R. Pagan. 1980. The Lagrange Multiplier Test and Its Applications to Model Specification in Econometrics. *Review of Economic Studies* 47: 239–53. [CrossRef]

Dhrymes, Phoebus. J. 1970. On the Game of Maximizing R Bar Square. *Australian Economic Papers* 9: 177–85. [CrossRef]

Efroymson, M. 1960. Multiple regression analysis. In *Mathematical Methods for Digital Computers*. Edited by A. Ralston and H. S. Wilf. New York: Wiley.

Engle, Robert F. 1984. Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. *Handbook of Econometrics* 2: 775–826.

Gelman, Andrew. 2015. The Connection Between Varying Treatment Effects and the Crisis of Unreplicable Research: A Bayesian Perspective. *Journal of Management* 41: 632–43. [CrossRef]

Ioannidis, John PA. 2005. Why most published research findings are false (Essay). *PLoS Medicine* 2: e124. [CrossRef] [PubMed]

Johnstone, David. 1988. Comments on Oakes on the foundation of statistical inference in the social and behavioral sciences: The market for statistical significance. *Psychological Reports* 63: 319–31. [CrossRef]

Kim, Jae H., and Philip Inyeob Ji. 2015. Significance testing in empirical finance: A critical review and assessment. *Journal of Empirical Finance* 34: 1–14. [CrossRef]

Kosfeld, Michael, Markus Heinrichs, Paul J. Zak, Urs Fischbacher, and Ernst Fehr. 2005. Oxytocin increases trust in humans. *Nature* 435: 673. [CrossRef]

Lane, Anthony, Moïra Mikolajczak, Evelyne Treinen, Dana Samson, Olivier Corneille, Philippe de Timary, and Olivier Luminet. 2015. Failed Replication of Oxytocin Effects on Trust: The Envelope Task Case. *PLoS ONE* 10: e0137000. [CrossRef]

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

Lovell, Michael 1983. Data mining. *Review of Economics and Statistics* 45: 1–12.

Lubke, Gitta H., and Ian Campbell. 2016. Inference Based on the Best-Fitting Model Can Contribute to the Replication Crisis: Assessing Model Selection Uncertainty Using a Bootstrap Approach. *Structural Equation Modeling: A Multidisciplinary Journal* 23: 479–90. [CrossRef] [PubMed]

Lucas, Robert E., Jr. 1976. Econometric Policy Evaluation: A Critique. In *Carnegie-Rochester Conference Series on Public Policy*. Edited by Karl Brunner and Alan Meltzer. Amsterdam: North-Holland.

Lucas, Robert E., and Thomas J. Sargent. 1981. *Rational Expectations and Econometric Practice*. London: George Allen & Unwin.

McCloskey, Donald N. 1985. The loss function has been mislaid: The rhetoric of significance tests. *American Economic Review* 75: 201–5.

Mikolajczak, Moïra, James J. Gross, Anthony Lane, Olivier Corneille, Philippe de Timary, and Olivier Luminet. 2010. Oxytocin Makes People Trusting, Not Gullible. *Psychological Science* 21: 1072–74. [CrossRef] [PubMed]

Nosek, Brian A., and Daniël Lakens. 2014. Registered reports: A method to increase the credibility of published results. *Social Psychology* 45: 137–41. [CrossRef]

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349: aac4716. [CrossRef] [PubMed]

Quiggin, John. 2019. *Economics in Two Lessons: Why Markets Work So Well, and Why They Can Fail So Badly*. Princeton: Princeton University Press.

econometrics

MDPI

*Article*

# A Frequentist Alternative to Significance Testing, *p*-Values, and Confidence Intervals

**David Trafimow**

Department of Psychology, MSC 3452, New Mexico State University, P.O. Box 30001, Las Cruces, NM 88003-8001, USA; dtrafimo@nmsu.edu

Received: 27 March 2019; Accepted: 24 May 2019; Published: 4 June 2019

**Abstract:** There has been much debate about null hypothesis significance testing, *p*-values without null hypothesis significance testing, and confidence intervals. The first major section of the present article addresses some of the main reasons these procedures are problematic. The conclusion is that none of them are satisfactory. However, there is a new procedure, termed the a priori procedure (APP), that validly aids researchers in obtaining sample statistics that have acceptable probabilities of being close to their corresponding population parameters. The second major section provides a description and review of APP advances. Not only does the APP avoid the problems that plague other inferential statistical procedures, but it is easy to perform too. Although the APP can be performed in conjunction with other procedures, the present recommendation is that it be used alone.

**Keywords:** a priori procedure; null hypothesis significance testing; confidence intervals; *p*-values; estimation; hypothesis testing

## 1. A Frequentist Alternative to Significance Testing, *p*-Values, and Confidence Intervals

Consistent with the purposes of the *Econometrics* special issue, my goal is to explain some of the problems with significance testing, point out that these problems are not solved satisfactorily using *p*-values without significance testing, and show that confidence intervals are problematic too. The second major section presents a frequentist alternative. The alternative can be used on its own or in conjunction with significance testing, *p*-values, or confidence intervals. However, my preference is for the alternative to be used on its own.

## 2. Discontent with Significance Testing, *p*-Values, and Confidence Intervals

### *2.1. Significance Testing*

Researchers use widely the null hypothesis significance testing (NHST) procedure, whereby the researcher computes a *p*-value, and if that value is under a threshold (usually 0.05), the result is declared statistically significant. Once the declaration has been made, the typical response is to conclude that the null hypothesis is unlikely to be true, reject the null hypothesis based on that conclusion, and accept the alternative hypothesis instead (Nickerson 2000). It is well-known that this sort of reasoning invokes a logical fallacy. That is, one cannot validly make an inverse inference from the probability of the obtained effect size or a more extreme one, given the null hypothesis; to the probability of the null hypothesis, given the obtained effect size (e.g., Cohen 1994; Fisher 1973; Nickerson 2000; Trafimow 2003).[1] The error is so common that it has a name: the modus tollens fallacy.

[1] This is an oversimplification. In fact, the *p*-value is computed from a whole model, which includes the null hypothesis as well as countless inferential assumptions. That the whole model is involved in computing a *p*-value will be addressed carefully later. For now, we need not consider the whole model to bring out the logical issue at play.

To see the reason for the name, consider that if the probability of the obtained effect size or a more extreme one given the null hypothesis were zero; then obtaining the effect size would guarantee that the null hypothesis is not true, by the logic of modus tollens (also termed denying the consequent). However, modus tollens does not work with probabilities other than zero and the unpleasant fact of the matter is that there is no frequentist way to compute the probability of the null hypothesis conditional upon the data. It is possible to use the famous theorem by Bayes; but most frequentists are unwilling to do that. Trafimow (2003, 2005) performed the Bayesian calculations and, not surprisingly, obtained very different findings from those obtained via the modus tollens error.[2] Thus, there is not only a logical invalidity; but a numerical one too.

An additional difficulty with making the modus tollens error is that to a frequentist, hypotheses are either true or false, and do not have probabilities between zero and unity. From this perspective, the researcher is simply wrong to assume that $p < 0.05$ implies that the probability of the null hypothesis given the data also is less than 0.05 (Nickerson 2000).

To be sure, researchers need not commit the modus tollens error. They can simply define a threshold level, such as 0.05, often termed an alpha level; and reject the null hypothesis whenever the obtained $p$-value is below that level and fail to reject the null hypothesis whenever the obtained $p$-value is not below that level. There is no necessity to assume anything about the probability of the null hypothesis.

But there are problems with threshold levels (see (Trafimow and Earp 2017) for a review). An important problem is that, under the null hypothesis, $p$-values have a uniform distribution between zero and one (0, 1). Therefore, whether the researcher obtains a $p$-value under the alpha level is partly a matter of luck. Just by getting lucky, the researcher may obtain a large sample effect size and a small $p$-value, thereby enabling publication. But in that event, the finding may be unlikely to replicate. Although this point is obvious from a statistical regression standpoint, the Open Science Collaboration effort (2015) showed empirically that the average effect size in the replication cohort of studies was less than half that in the original cohort of studies (from 0.403 to 0.197). Locascio (2017a) has argued that the most important disadvantage of NHST is that it results in scientific literatures replete with inflated effect sizes.[3]

Some have favored reducing the alpha level to lower values, such as 0.01 or 0.005 (e.g., Melton 1962; Benjamin et al. 2018); but these suggestions are problematic too. The most obvious problem is that having a more stringent alpha level merely increases statistical regression effects so that published sample effect sizes become even more inflated (Trafimow et al. 2018). Furthermore, more stringent thresholds need not increase replication probabilities. As Trafimow et al. (2018) pointed out, applying a more stringent alpha level to both the original and replication studies would render replication even more difficult. To increase the replication probability, the researcher would need to apply the more stringent alpha level to the original study and a less stringent alpha level to the replication study. And then there is the issue of what justifies the application of different alpha levels to the two categories of studies.

There are many problems with NHST, of which the present is only a small sampling. But this small sampling should be enough to render the reader highly suspicious. Trafimow and Earp (2017) presented a much larger list of problems and the subsequent section includes yet additional problems. Critique of the use of Fisherian $p$-values and Neyman-Pearson hypothesis testing for model selection has even made its way into (graduate) statistics textbooks; see, e.g., Paolella (2018, sct. 2.8).

[2] Also see Kim and Ji (2015) for Bayesian calculations pertaining to significance tests in empirical finance.

[3] The interested reader should consult the larger discussion of the issue in the pages of *Basic and Applied Social Psychology* (Grice 2017; Hyman 2017; Kline 2017; Locascio 2017a, 2017b; Marks 2017).

### 2.2. *p*-Values without Significance Testing

The American Statistical Association (Wasserstein and Lazar 2016) admits that *p*-values provide a poor justification for drawing conclusions about hypotheses. But might *p*-values be useful for something else? An often-touted possibility is to use *p*-values to obtain evidence against the statistical model of which the null hypothesis—or test hypothesis more generally—is a subset. It is a widely known fact—though not widely attended—that *p*-values depend on the whole model and not just on the test hypothesis. There are many assumptions that go into the full inferential statistical model, in addition to the test hypothesis, such as that the researcher sampled randomly and independently from a defined population (see (Trafimow), for a taxonomy of model assumptions). As Berk and Freedman (2003) pointed out, this assumption is practically never true in the soft sciences. Worse yet, there are many additional assumptions too numerous to list here (Armhein et al. 2019; Berk and Freedman 2003; Trafimow). As Armhein et al. (2019) concluded (p. 263): "Thus, statistical models imply countless assumptions about the underlying reality." It cannot be overemphasized that *p*-values pertain to models—and their countless assumptions—as opposed to just hypotheses.

Let us pretend, for the moment, that it is a valuable exercise for the researcher to obtain evidence against the model.[4] How well do *p*-values fulfill that objective? One problem with *p*-values is that they are well known to be unreliable (e.g., Halsey et al. 2015), and thus cannot provide strong evidence with respect to models or to the hypotheses that are subsets of models.[5] In addition to conceptual demonstrations of this point (Halsey et al. 2015), an empirical demonstration also is possible, thanks to the data file the Open Science Collaboration (2015) displayed online (https://osf.io/fgjvw/). The data file contains a wealth of data pertaining to the original cohort of published studies and the replication cohort. After downloading the data file, I obtained exact *p*-values for each study in the cohort of original studies; and for each study in the cohort of replication studies. After correlating the two columns of *p*-values, I obtained a correlation coefficient of 0.0035.[6] This empirical demonstration of the lack of reliability of *p*-values buttresses previous demonstrations involving mathematical or computer simulations.

And yet, there is a potential way out. Greenland (2019) suggested performing a logarithmic transformation of *p*-values: $(-1) \cdot \log_2(p)$. The logarithmic transformation causes *p*-values to be expressed as the number of bits of information against the model. For example, suppose that $p = 0.05$. Applying the logarithmic transformation implies that there are approximately four (the exact number is 4.32) bits of information against the model. An advantage of the transformation, as opposed to the untransformed *p*-value, is that the untransformed *p*-value has endpoints at 0 and 1, with the restriction of range potentially reducing the correlation that can be obtained between original and cohort sets of *p*-values. As an empirical demonstration, when I used the logarithmic transformation on the two columns of *p*-values obtained from the Open Science Collaboration data file, the original *p*-value correlation of 0.0035 jumped to 0.62! Thus, not only do transformed *p*-values have a more straightforward interpretation than do untransformed *p*-values; but they replicate better too. Clearly, if one insists on using a *p*-value to index evidence against the model, it would be better to use a transformed one than an untransformed one. However, there nonetheless remains the issue of whether it is worthwhile to gather evidence against the model in the first place.

Let us return to the issue that a *p*-value—even a transformed *p*-value—is conditioned not only on the test hypothesis; but rather on the whole model. For basic research, where one is testing a

[4] As will become clear later, this is not a valuable exercise; but the pretense is nevertheless useful to make an important point about logarithmic transformations of *p*-values.

[5] Sometimes *p*-value apologists admit *p*-value unreliability but point out that such unreliability has been known from the start. Although this contention is correct, it fails to justify *p*-values. That a procedure has been known from the start to be unreliable does not justify its use!

[6] This correlation, rounded to 0.004, was mentioned by Trafimow and de Boer (2018); but these researchers did not assess transformed *p*-values.

theory, there are assumptions in the theory (theoretical assumptions) as well as auxiliary assumptions used to connect non-observational terms in the theory to observational terms in empirical hypotheses (Trafimow). If the researcher is to employ descriptive statistics, such as means, standard deviations, and so on; there are statistical assumptions pertaining to issues such as whether means should be used or whether some other location statistic should be used, whether standard deviations should be used or whether some other dispersion statistic should be used, and so on (Trafimow 2019a). Finally, there are inferential assumptions such as those pointed out by Berk and Freedman (2003), especially concerning random and independent sampling. As suggested earlier, it is a practical certainty that not all the assumptions are precisely true, which means that the model is wrong (Armhein et al. 2019; Berk and Freedman 2003; Trafimow 2019b, forthcoming). The question then arises: What is the point in gathering evidence against a model that is already known to be wrong? The lack of a good answer to this question is a strong point against using statistics of any sort, even transformed $p$-values, to gather evidence against the model.

A counter can be generated out of the cliché that although all models are wrong, they might be close enough to correct to be useful (e.g., Box and Draper 1987). As an example of the cliché, suppose that based on the researcher's statistical model, she considers sample means as appropriate to estimate population means. In addition, suppose that the sample means really are close to their corresponding population means, though not precisely correct. It would be reasonable to argue that the sample means are useful, despite not being precisely correct, because they give the researcher a good approximation of the population means. Can this argument be extended to $p$-values or transformed $p$-values?

The answer is in the negative. To see why, consider that in the case of a $p$-value or a transformed $p$-value, there is no corresponding population parameter; the researcher is not estimating anything! And if the researcher is not estimating anything, closeness is irrelevant. Being close counts for much if the goal concerns estimation; but being close is irrelevant in the context of making accept/not accept decisions about hypotheses. In the case of such binary decisions, one cannot be close; one can only be correct or incorrect. It is worthwhile to reiterate. The Box and Draper (1987) quotation makes sense in estimation contexts; but it fails to save $p$-values or transformed $p$-values because nothing is being estimated. Obtaining evidence against a known wrong model fails to provide useful information.

### 2.3. Confidence Intervals

Do the foregoing criticisms of $p$-values, either with or without NHST, provide a strong case for using confidence intervals (CIs) instead? Not necessarily.

To commence, most researchers use CIs like they use $p$-values. That is, if the critical value falls outside the CI, it is statistically significant; and if the critical value does not fall outside the CI, it is not statistically significant. Used in this way, CIs are plagued with all the problems that plague $p$-values when used for NHST.

Alternatively, researchers may use CIs for parameter estimation. For example, many researchers believe that constructing a 95% CI around the sample mean indicates that the population mean has a 95% chance of being within the constructed interval. However, this is simply false. There is no way to know the probability that the population mean is within the constructed interval. To understand what a 95% CI really entails, it is necessary to imagine the experiment performed an indefinite number of times, with the researcher constructing a 95% CI each time. In this hypothetical scenario, 95% of the constructed 95% CIs would enclose the population mean but there is no way to know the probability that any single constructed interval contains the population mean. Interpreting a CI as giving the probability that the population parameter is within the constructed interval constitutes another way of making an inverse inference error, not dissimilar to that discussed earlier with respect to $p$-values. Furthermore, if one is a frequentist, it does not even make sense to talk about such a probability as the population parameter is either in the interval or not, and lack of knowledge on the part of the researcher does not justify the assignment of a probability.

Sophisticated CI aficionados understand the foregoing CI misinterpretations and argue instead that the proper use of CIs is to provide information about the precision of the data. A narrow CI implies high precision and a wide CI implies low precision. But there is an important problem with this argument. Trafimow (2018b) showed that there are three types of precision that influence the size of CIs. There is sampling precision: larger sample sizes imply greater sampling precision under the usual assumptions. There is measurement precision: the more the random measurement error, the lower the measurement precision. Finally, there is precision of homogeneity: the more similar the people in the sample are to each other, the easier it is to discern the effect of the manipulation on the dependent variable. CIs confound the three types of precision.[7]

The obvious counter is that even a confounded precision index might be better than no precision index whatsoever. But Trafimow (2018b, 2019a) showed that provided the researcher has assessed the reliability of the dependent variable, it is possible to estimate the three kinds of precision separately, which provides superior precision information than a triply confounded CI. The fact that the three types of precision can be estimated separately places the CI aficionado in a dilemma. On the one hand, if the aficionado is honestly interested in precision, she should take the trouble to assess the reliability of the dependent variable and estimate the three types of precision separately. On the other hand, if the aficionado is not interested in precision, there is no reason for her to compute a CI anyhow. Thus, either way, there is no reason to compute a CI.

A further problem with CIs, as is well-known, is that they fluctuate from sample to sample. Put another way, CIs are unreliable just as *p*-values are unreliable. Cumming and Calin-Jageman (2017) have attempted to justify that this is okay because most CIs overlap with each other. But unfortunately, the extent to which CIs overlap with each other is not the issue. Rather, the issue—in addition to the foregoing precision issue—is whether sample CIs are good estimates of the CI that applies to the population. Of course, in normal research, the CI that applies to the population is unknown. But it is possible to perform computer simulations on user-defined population values. Trafimow and Uhalt (under submission) performed this operation on CI widths, as well as upper and lower CI limits. The good news is that as sample sizes increase, sample CI accuracy also increases. The bad news is that unless the sample size is much greater than those typically used, the accuracy of CI ranges and limits is very poor.

In summary, CIs are triply confounded precision indexes, they tend not to be accurate, and they are not useful for estimating population values. To what valid use CIs can be put is far from clear.

### 2.4. Bayesian Thinking

There are many ways to “go Bayesian.” In fact, Good (1983) suggested that there are at least 46,656 ways![8] Consequently, there is no feasible way to do justice to Bayesian statistical philosophy in a short paragraph. The interested reader can consult Gillies (2000), who examined some objectivist and subjectivist Bayesian views, and described important dilemmas associated with each of them. There is no attempt here to provide a critical review, except to say that not only do Bayesians disagree with frequentists; but Bayesians often disagree with each other too. For researchers who are not Bayesians, it would be useful to have a frequentist alternative that is not susceptible of the problems discussed

[7] To understand why, consider that CIs are based largely on the standard error. In turn, the standard error is based on the standard deviation and the sample size. Finally, the standard deviation is influenced by random measurement error but also by systematic differences between people. Thus, the standard deviation in the numerator of the standard error calculation is influenced by both measurement precision and precision of homogeneity; and the denominator of the standard error includes the sample size, thereby implicating the importance of sampling precision. Thus, all three types of precision influence the standard error. This triple confound is problematic for interpreting CIs.

[8] Good (1983) stated that Bayesians can make a variety of choices with respect to a variety of facets. In calculating that the number of Bayesian categories equals 46,656, Good also pointed out that this is larger than the number of professional statisticians so there are empty categories (p. 21).

with respect to NHST, $p$-values, and CIs. Even Bayesians might value such an alternative. We go there next.

## 3. The A Priori Procedure (APP)

The APP takes seriously that the researcher wishes to estimate population parameters based on sample statistics. To see quickly the importance of having sample statistics that are at least somewhat indicative of corresponding population parameters, imagine Laplace's Demon, who knows everything, and who warns researchers that their sample means have absolutely nothing to do with corresponding population means. Panic would ensue in science because there would be no point in obtaining sample means.

In contrast to the Demon's disastrous pronouncement, let us imagine a different pronouncement. Suppose the Demon offered researchers the opportunity to specify how close they wish their sample statistics to be to corresponding population parameters; and the probability of being that close. And the Demon would provide the necessary sample sizes to achieve specifications. This would be of obvious use, though not definitive. That is, researchers could tell the Demon their specifications, the Demon could answer with necessary sample sizes, and researchers could then go out and obtain those sample sizes. Upon obtaining the samples, researchers could compute their descriptive statistics of interest under the comforting assurance that they have acceptable probabilities of being within acceptable distances of the population values. After all, it is the researcher who told the Demon the specifications for closeness and probability. Because the researcher is assured that the sample statistics have acceptable probabilities of being within acceptable distances of corresponding population parameters, the need for NHST, $p$-values, or CIs is obviated.

Of course, there is no Demon; but the APP can take the Demon's place. As a demonstration, consider the simplest possible case where the researcher is interested in a single sample with a single mean, and where participants are randomly and independently sampled from a normally distributed population. That these assumptions are rarely true will be addressed later. For now, though, let us continue with this ideal and simple case to illustrate how APP thinking works.

Trafimow (2017) provided an accessible proof of Equation (1) below:

$$n = \left(\frac{Z_C}{f}\right)^2, \tag{1}$$

where

- $f$ is the fraction of a standard deviation the researcher defines as sufficiently "close,"
- $Z_C$ is the $z$-score that corresponds to the desired probability of being close, and
- $n$ is the minimum sample size necessary to meet specifications for closeness and confidence.

For example, suppose the researcher wishes to have a 95% probability of obtaining a sample mean that is within a quarter of a standard deviation of the population mean. Because the $z$-score that corresponds to 95% confidence is approximately 1.96, Equation (1) can be solved as follows: $n = \left(\frac{1.96}{0.25}\right)^2 = 61.47 \approx 62$. In other words, the researcher will need to recruit 62 participants to meet specifications for closeness and confidence.[9]

The researcher now can see the reason for the name, a priori procedure. All the inferential work is performed ahead of data collection and no knowledge of sample statistics is necessary. But the fact that the APP is an a priori procedure does not preclude its use in a posteriori fashion. To see that this is indeed possible, suppose that a researcher had already performed the study and a second researcher wishes to estimate the closeness of the reported sample mean to the population mean, under

[9] Because participants do not come in fractions, it is customary to round upwards to the nearest whole number.

the typical value of 95% for confidence. Equation (2) provides an algebraic rearrangement of Equation (1), to obtain a value for $f$ given that $n$ has already been published:

$$f = \frac{Z_C}{\sqrt{n}}. \tag{2}$$

For example, suppose that the researcher had 200 participants. What is the closeness? Using Equation (2) implies the following: $f = \frac{1.96}{\sqrt{200}} = 0.14$. In words, the researcher's sample mean has a 95% probability of being within 0.14 standard deviations of the population mean.

### 3.1. *j* Groups

An obvious complaint to be made about Equations (1) and (2) is that they only work for a single mean. But it is possible to extend to as many groups as the researcher wishes. Trafimow and MacDonald (2017) derived Equation (3), that allows researchers to calculate the sample size per condition needed to ensure that all ($j$) sample means are within specifications for closeness and probability:

$$n = \left( \frac{\Phi^{-1}\left( \frac{\sqrt[j]{p(j\ means)}+1}{2} \right)}{f} \right)^2 \tag{3}$$

where

- $j$ is the number of groups,
- $p(j\ means)$ is the probability of meeting the closeness specification with respect to the $j$ groups,
- and $\Phi^{-1}$ is the inverse of the cumulative distribution function of the normal distribution.

Algebraic rearrangement of Equation (3) renders Equation (4) that can be used to estimate the precision of previously published research:

$$f = \left( \frac{\Phi^{-1}\left( \frac{\sqrt[j]{p(j\ means)}+1}{2} \right)}{\sqrt{n}} \right) \tag{4}$$

(Trafimow and Myüz (forthcoming) used Equation (4) to analyze a large sample of published papers in lower-tier and upper-tier journals in five areas of psychology. They found that although precision was unimpressive in all five areas of psychology, it was worst in cognitive psychology and least bad in developmental and social psychology.

### 3.2. Differences in Means

In some research contexts, researchers may be more interested in having the difference in two sample means be close to the population difference, than in having each individual mean be close to its corresponding population mean. Researchers who are interested in differences in means may use matched or independent samples. If the samples are matched, Trafimow et al. (forthcoming) derived the requisite equation:

$$t_{\frac{\alpha}{2},\, n-1} \leq \sqrt{n} f, \tag{5}$$

where $t_{\frac{\alpha}{2},\, n-1}$ is the critical $t$-score, analogous to the use of the $z$-score from Equation (1). Unfortunately, Equation (5) cannot be used in the simple manner that Equations (1)–(4) can be used. For instance, suppose a researcher wishes to specify $f = 0.2$ at 95% confidence for matched samples. The researcher might try $n = 99$, so the right side of Equation (2) is 1.99, which satisfies Equation (5). That is, $t_{\frac{\alpha}{2},\, n-1} = 1.81 \leq \sqrt{99} \cdot 0.2 = 1.99$. Alternatively, the researcher might try $n = 98$, which does not satisfy

Equation (5): $t_{\frac{\alpha}{2},\,n-1} = 1.9847 \nleq \sqrt{99}{\cdot}0.2 = 1.9799$. Because $n = 99$ satisfies Equation (5) whereas $n = 98$ does not, the minimum sample size necessary to meet specifications for precision and confidence is $n = 99$. Equation (5) is best handled with a computer that is programmed to try different values until convergence on the smallest sample size that fulfils requirements.

Equation (5) can be algebraically rearranged to give closeness, as Equation (6) shows:

$$f \geq \frac{t_{\frac{\alpha}{2},\,n-1}}{\sqrt{n}} \tag{6}$$

If the researcher uses independent samples, as opposed to matched samples, Equation (5) will not work and it is necessary instead to use Equation (7). When there are independent samples, there is no guarantee that the sample sizes will be equal, and it is convenient to designate that there are $n$ participants in the smaller group and $m$ participants in the larger group, where $k = \frac{n}{m}$. Using $k$, Trafimow et al. (forthcoming) derived Equation (7):

$$t_{\frac{\alpha}{2},\,q} \leq \sqrt{\frac{n}{k+1}}f, \tag{7}$$

where $t_{\frac{\alpha}{2},q}$ is the critical $t$-score that corresponds to the level of confidence level $1-\alpha$ and degrees of freedom $q = n + \left[\frac{n}{k}\right] - 2$ in which $\left[\frac{n}{k}\right]$ is rounded to the nearest upper integer.

If the researcher has equal sample sizes, Equation (3) reduces to Equation (4):

$$t_{\frac{\alpha}{2},\,2(n-1)} \leq \sqrt{\frac{n}{2}}f. \tag{8}$$

Like Equation (5), Equation (7) or Equation (8) is best handled using a computer to try out different sample sizes. Again, the lowest sample sizes for which the equations remain true are those required to meet specifications.

Alternatively, if the researcher is interested in the closeness of already published data, Equation (7) can be algebraically rearranged to render Equation (9); and Equation (8) can be algebraically rearranged to render Equation (10).

$$f \geq \frac{t_{\frac{\alpha}{2},\,q}}{\sqrt{\frac{n}{k+1}}} \tag{9}$$

$$f \geq \frac{t_{\frac{\alpha}{2},\,2(n-1)}}{\sqrt{\frac{n}{2}}} \tag{10}$$

### 3.3. Skew-Normal Distributions

Most researchers assume normality and consequently would use Equations (1)–(10). But many distributions are skewed (Blanca et al. 2013; Ho and Yu 2015; Micceri 1989), and so Equations (1)–(10) may overestimate the sample sizes needed to meet specifications for closeness and confidence. The family of skew-normal distributions is more generally applicable than the family of normal distributions. This is because the family of skew-normal distributions employs three, rather than two, parameters. Let us first consider the two parameters of normal distributions: mean $\mu$ and standard deviation $\sigma$. For skew-normal distributions, these are replaced by the location parameter $\xi$ and scale parameter $\omega$, respectively. Finally, skew-normal distributions also include a shape parameter $\lambda$. When $\lambda = 0$, the distribution is normal, and $\xi = \mu$ and $\omega = \sigma$. But when $\lambda \neq 0$, then the distribution is skew-normal, and $\xi \neq \mu$ and $\omega \neq \sigma$. Although the mathematics are too complex to render here, skew-normal equations have been derived analogous to Equations (1)–(10) (e.g., Trafimow et al. 2019). In addition, Wang et al. (2019a) have shown how to find the number of participants necessary to meet specifications for closeness and confidence with respect to estimating the shape parameter. Finally, Wang et al. (2019b) have shown how to find the number of participants necessary to meet specifications for closeness and confidence with respect to estimating the scale parameter (or standard deviation if normality is assumed).

### *3.4. Limitations*

Although much progress has been made in the approximately two years since the APP was invented, there nevertheless remain limitations. The most important limitations are conceptual. Unlike many other inferential statistical procedures, the APP does not dictate what hypotheses to accept or reject. For those researchers who believe that other inferential statistical procedures, such as NHST, really do validly dictate what hypotheses to accept or reject, this is an important limitation. Hopefully, the remarks in the first major section of the present article have disabused the reader that any procedure is valid for making decisions about hypotheses. If the reader is convinced, then although the limitation remains serious in the absolute sense that it would be nice to have an inferential procedure that validly dictates what hypotheses to accept or reject; the limitation is not serious in the relative sense that other inferential procedures that make the promise fail to deliver, so nothing is lost by using the APP.[10]

A second conceptual limitation is suggested by one of the arguments against *p*-values. That is, the model is known wrong and so there is no point using *p*-values to gather evidence against it. It is tempting to apply model wrongness to the APP, where the models again will not be precisely correct. However, as we pointed out earlier, closeness counts heavily with respect to estimation; but does not count at all for binary decisions, that are either correct or incorrect. Because the APP is for estimation, if the model is reasonably close to being correct, though it cannot be precisely correct, the estimate should be reasonably good. And this point can be explained in more specific terms. Suppose that the model is close but not perfect, thereby resulting in a sample size that is slightly larger or slightly smaller than the precisely correct sample size necessary to meet specifications. In the case of the larger sample size, the result will be that the researcher will have slightly better closeness, and little harm is done, except that the researcher will have put greater than optimal effort into data collection. In the case of the smaller sample size, the researcher's sample statistics of interest will not be quite as close to their corresponding population parameters as desired; but may nevertheless be close enough to be useful. Therefore, although model wrongness always is important to consider, it need not be fatal for the APP whereas it generally is fatal for *p*-values.

There are also practical limitations. Although work is in progress concerning complex contrasts involving multiple means (assuming a normal distribution) or multiple locations (assuming a skew-normal distribution), the requisite APP equations do not yet exist. Similarly, for researchers interested in correlations, regression weights, and so on; although work is in progress; the requisite APP equations do not yet exist. Another practical limitation is the lack of an APP computer program that will allow researchers to perform the calculations without having to do their own programming. And yet, work proceeds and we hope to be able to address these practical limitations in the very near future.

### *3.5. APP versus Power Analysis*

To some, the APP may seem like merely an advanced way to perform power analysis. However, this is not so as can be shown in both a general sense and in two specific senses. Speaking generally, the APP and power analysis have very different goals. The goal of power analysis is to find the number of participants needed to have a good chance of obtaining a *p*-value that comes in under threshold

[10] I thank an anonymous reviewer for pointing out that, due to the lack of cutoff points, this limitation can be considered a strength. According to the reviewer, "There is no cutoff point, so potentially all estimates could be viable."

(e.g., $p < 0.05$) when the null hypothesis is meaningfully violated.[11] In contrast, the APP goal is to find the number of participants necessary to reach specifications for closeness and confidence.

This general difference results in specific mathematical differences too.[12] First, power analysis depends importantly on the anticipated effect size. If the anticipated effect size is small, a power analysis will indicate that many participants are necessary for adequate power; but if the anticipated effect size is large, a power analysis will indicate that only a small sample size is necessary for adequate power. In contrast, the anticipated effect size plays no part whatsoever in APP calculations. For example, suppose that anticipated effect size for a single sample experiment is 0.80. A power analysis would show that only 13 participants are needed for power = 0.80; but an APP calculation would nevertheless demonstrate that the closeness value is a woeful 0.54.[13]

A second specific difference is that APP calculations are influenced, importantly, by the desired level of closeness. In contrast, power calculations are completely uninfluenced by the desired level of closeness. Moreover, absent APP thinking, few researchers would even consider the issue of the desired level of closeness. In summary, the APP is very different from power analysis, both with respect to general goals and with respect to specific factors that influence how the calculations are performed.

### *3.6. The Relationship between the APP and Idealized Replication*

Much recent attention concerns replication probabilities across sciences. For example, the Open Science Foundation (2015) publication indicates that most published papers in top psychology journals failed to replicate. One of the many disadvantages of both *p*-values and CIs is that they fail to say much about the extent to which experiments would be likely or unlikely to replicate. In contrast, as Trafimow (2018a) explained in detail, the results of the APP strongly relate to reproducibility.

To understand the relationship, it is necessary to make two preliminary points. The first point is philosophical and concerns what we would expect a successful replication to entail. Because of scientists' addiction to NHST, most consider a successful replication to entail statistically significant findings in the same direction, in both the original and replication studies. However, once NHST is admitted as problematic, defining a successful replication in terms of NHST is similarly problematic. But the present argument extends beyond NHST to effect sizes more generally.

Consider the famous Michelson and Morley (1887) experiment that disconfirmed the presence of the luminiferous ether that researchers had supposed to permeate the universe.[14] The surprise was that the effect size was near zero, thereby suggesting that there is no luminiferous ether, after all.[15] Suppose a researcher today wished to replicate. Because larger effect sizes correspond with lower *p*-values, it should be clear that going by replication conceptions involving *p*-values, it is much more difficult to replicate smaller effect sizes than larger ones.[16] Thus, according to traditional NHST thinking, it should be extremely difficult to replicate Michelson and Morley (1887), though physicists do not find it so. This is one reason it is a mistake to let effect sizes dictate replication probabilities. In contrast, using APP thinking, a straightforward conceptualization of a successful replication is if the descriptive statistics of concern are close to their corresponding population parameters in both the original and replication studies.[17] An advantage of this conceptualization is that it treats large and

[11] For those who prefer CIs, an alternative goal would be to find the number of participants required to obtain sample CIs of desired widths.

[12] For elaborated mathematical discussions of the differences, see Trafimow and Myüz (Trafimow and Myüz) and Trafimow (2019b).

[13] See (Trafimow and Myüz (forthcoming) for details.

[14] Michelson received his Nobel Prize in 1907.

[15] It is interesting that Carver (1993) reanalyzed the data using NHST and obtained a statistically significant effect due to the large number of data points. As Carver pointed out, had Michelson and Morley used NHST, the existence of the luminiferous ether would have been supported, with incalculable consequences for physics (also see Trafimow and Rice 2009).

[16] A counter might be to use equivalence testing; but this is extremely problematic because it involves the computation of at least two *p*-values, whereas we already have seen that even one *p*-value is problematic.

[17] If specifications are not met in one of the two studies, that constitutes a failure to replicate.

small effect sizes equally. What matters is not the size of the effect; but rather how close the sample effect is to the population effect, in both the original and replication studies.

The second point is to imagine an idealized universe, where all systematic factors are the same in both the original and replication study. Thus, the only differences between the original and replication study are due to randomness.

Remembering that our new conceptualization of a successful replication pertains to the sample statistics of interest being close to their corresponding population parameters, in both the original and replication studies, invoking an idealized universe suggests a simple way to calculate the probability of replication. Specifically, the probability of replication in the idealized universe is simply the square of the probability of being close in a single experiment (Trafimow 2018a). We have already seen that all APP equations can be algebraically rearranged to yield closeness given the sample size used in the original study. Well, then, it is equally possible to fix closeness at some level and algebraically rearrange APP equations to yield the probability of meeting the closeness specification given the sample size used. Once this has been accomplished, the researcher merely squares that probability to obtain the probability of replication in the idealized universe. Trafimow (2018a) described the mathematics in detail and showed how specifications for closeness and sample size influence the probability of replication.

A way to attack the usefulness of the APP conceptualization of a successful replication is to focus on the necessity to invoke an idealized universe to carry through the calculations. But the attack can be countered in both general and specific ways. The general counter is that scientists often have suggested idealized universes, such as the ideal gas law in chemistry, Newton's idealized universe devoid of friction, and so on. The history of science shows that idealized conceptions often have been useful, though not strictly correct (Cartwright 1983). More specifically, however, consider that the difference between the APP idealized universe and real universe is that only random factors can hinder replication in the idealized universe whereas both random and systematic factors can hinder replication in the real universe. Because there is more that can go wrong in the real universe than in the idealized universe, it should be clear that the probability of replication in the idealized universe sets an upper limit on the probability of replication in the real universe. Because Trafimow (2018a) showed that most research has a low probability of replication even in the idealized universe, the probability of replication in the real universe must be even lower. In summary, whereas $p$-values and CIs have little to say about the probability of replication, the APP has much to say about it.

### *3.7. Criteria*[18]

An unaddressed issue concerns the setting of criteria with respect to closeness and the probability of being close. The temptation is to set cutoffs; for example, that closeness must be at the 0.02 level or better, at 95% probability or better, to justify publication. However, this temptation must be resisted, lest the APP degenerate into dichotomous thinking that is currently problematic in the sciences. Instead of cutoffs, it would be better to have graduated verbal descriptions for what constitutes different levels of closeness and probabilities of being close. However, even graduated verbal descriptions may be problematic because researchers in different fields, or areas within fields, might justifiably differ with respect to what constitutes suitable verbal descriptions. For example, closeness at the 0.40 level might be "poor" in some fields or areas, and "acceptable" in others. It would be a mistake to impose such criteria from outside.[19]

One way to address the issue would be to have conferences, workshops, or symposia where people in similar fields and areas meet; become familiar with the APP, with a solid understanding about closeness and the probability of being close; and engage in serious discussions about criteria for

[18] I thank an anonymous reviewer for suggesting this issue.

[19] Trafimow (2018a) used some graduated verbal descriptions but also emphasized that these should not be taken very seriously.

graduated verbal descriptions. An alternative possibility would be to use the social media. Yet another alternative would be for editors of substantive journals to promulgate special issues devoted to setting criteria for graduated verbal descriptions such that substantive experts can provide perspectives.

Given academia's publish or perish culture, it cannot be overemphasized what a mistake it would be to turn APP thinking dichotomous, with publication thresholds for closeness and probability of being close. The ability of researchers to meet criteria for various verbal descriptions should only be one consideration for publication. Many factors should influence publication decisions, including the worth of the theory, the execution of the study, the feasibility of obtaining participants, the writing clarity, and others. Hopefully, having graduated verbal descriptions, instead of dichotomous cutoffs, that differ across fields and areas; will facilitate journal editors and reviewers to engage in more nuanced thinking that weighs many relevant factors.

## 4. Conclusions

The first major section considered NHST, *p*-values without NHST, and CIs. All were found wanting. Consequently, the second major section focused on a new way to think: the APP. The APP differs from the others because all the inferential work is performed before data collection. This is not to say that the others involve no work, whatsoever, before data collection. The setting of threshold levels, for instance, is work that is done before data collection when one uses NHST. But there nevertheless is a strong difference. With traditional procedures, once the data have been collected, it is still necessary to calculate *p*-values or CIs. In contrast, using the APP, the only inferential work that needs to be done after data collection is to acknowledge the results of the descriptive work. The researcher can be assured that the descriptive statistics have acceptable probabilities of being acceptably close to corresponding population parameters. After all, it is the researcher who decided what constitutes an acceptable probability and an acceptable degree of closeness and collected the requisite sample size to meet specifications. And if reviewers or editors believe that the investigator was too liberal in setting specifications, they have the option to reject the manuscript or insist that the researcher augment the sample. For example, if the researcher uses $f = 0.4$ and the editor favors $f = 0.1$, it is transparent how to calculate how much the researcher needs to augment the sample to reach the editor's specification.

A reasonable person might agree that the APP is a good thing; but also argue that NHST, *p*-values without NHST, or CIs are good too. As was stated earlier, there is nothing about APP calculations performed before data collection that renders impossible the calculation of *p*-values or CIs after data collection. Thus, it is possible to use all the procedures for the same study. This possibility need not be inconvenient for setting the APP apart from other procedures. If researchers were to routinely use the APP, they also would become accustomed to APP thinking. In turn, this would result in their eventually perceiving just how barren *p*-values and CIs are if one wishes to advance science. This is not to say that scientists should not test hypotheses. They should. But they should not depend on automatized decision-makers such as *p*-values and CIs to do it. Instead, researchers should perform much of their thinking up front and make a priori specifications for closeness and confidence. Then they should take their descriptive results seriously; with such seriousness being warranted by a priori specifications of acceptable closeness at acceptable probabilities. Of course, there are other factors that also influence the trust researchers place in descriptive statistics, such as the worth of the theory, the validity of the auxiliary assumptions, and so on. Whether or how much to believe substantive hypotheses, or the larger theories from which they are derived, is a process that cannot be automated. There will always remain an important role for expert judgment. The APP recognizes this explicitly.

**Funding:** This research received no external funding.

**Conflicts of Interest:** The author declare no conflict of interest.

## References

Armhein, Valentin, David Trafimow, and S. Sander Greenland. 2019. Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician* 73: 262–70.

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, and Colin Camerer. 2018. Redefine statistical significance. *Nature Human Behavior* 33: 6–10. [CrossRef]

Berk, Richard. A., and David A. Freedman. 2003. Statistical assumptions as empirical commitments. In *Law, Punishment, and Social Control: Essays in Honor of Sheldon Messinger*, 2nd ed. Edited by Thomas G. Blomberg and Sheldon Cohen. New York: Aldine de Gruyter, pp. 235–54.

Blanca, Maria, Jaume J. Arnau, Dolores López-Montiel, Roser Bono, and Rebecca Bendayan. 2013. Skewness and kurtosis in real data samples. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 9: 78–84. [CrossRef]

Box, George E. P., and Norman R. Draper. 1987. *Empirical Model-Building and Response Surfaces*. New York: John Wiley and Sons.

Cartwright, Nancy. 1983. *How the Laws of Physics Lie*. Oxford: Oxford University Press.

Carver, Ronald P. 1993. The case against statistical significance testing, revisited. *Journal of Experimental Education* 61: 287–92. [CrossRef]

Cohen, Jacob. 1994. The earth is round ($p < 0.05$). *American Psychologist* 49: 997–1003.

Cumming, Geoff, and Robert Calin-Jageman. 2017. *Introduction to the New Statistics: Estimation, Open Science, and Beyond*. New York: Taylor and Francis Group.

Fisher, Ronald A. 1973. *Statistical Methods and Scientific Inference*, 3rd ed. London: Collier Macmillan.

Gillies, Donald. 2000. *Philosophical Theories of Probability*. London: Taylor and Francis.

Good, Irving J. 1983. *Good Thinking: The Foundations of Probability and Its Applications*. Minneapolis: University of Minnesota Press.

Greenland, Sander. 2019. The Unconditional Information in *p*-values, and Its Refutational Interpretation via S-values. *The American Statistician* 73: 106–14. [CrossRef]

Grice, James W. 2017. Comment on Locascio's results blind manuscript evaluation proposal. *Basic and Applied Social Psychology* 39: 254–55. [CrossRef]

Halsey, Lewis G., Douglas Curran-Everett, Sarah L. Vowler, and Gordon B. Drummond. 2015. The fickle *P* value generates irreproducible results. *Nature Methods* 12: 179–85. [CrossRef]

Ho, Andrew D., and Carol C. Yu. 2015. Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects. *Educational and Psychological Measurement* 75: 365–88. [CrossRef]

Hyman, Michael. 2017. Can 'results blind manuscript evaluation' assuage 'publication bias'? *Basic and Applied Social Psychology* 39: 247–51. [CrossRef]

Kim, Jae H., and Philip I. Ji. 2015. Significance testing in empirical finance: A critical review and empirical assessment. *Journal of Empirical Finance* 34: 1–14. [CrossRef]

Kline, Rex. 2017. Comment on Locascio, results blind science publishing. *Basic and Applied Social Psychology* 39: 256–57. [CrossRef]

Locascio, Joseph. 2017a. Results blind publishing. *Basic and Applied Social Psychology* 39: 239–46. [CrossRef] [PubMed]

Locascio, Joseph. 2017b. Rejoinder to responses to "results blind publishing". *Basic and Applied Social Psychology* 39: 258–61. [CrossRef] [PubMed]

Marks, Michael J. 2017. Commentary on Locascio. *Basic and Applied Social Psychology* 39: 252–53. [CrossRef]

Melton, Arthur. 1962. Editorial. *Journal of Experimental Psychology* 64: 553–57. [CrossRef]

Micceri, Theodore. 1989. The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin* 105: 156–66. [CrossRef]

Michelson, Albert A., and Edward W. Morley. 1887. On the relative motion of earth and Luminiferous ether. *American Journal of Science, Third Series* 34: 333–45. [CrossRef]

Nickerson, Raymond S. 2000. Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods* 5: 241–301. [CrossRef]

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349: aac4716. [CrossRef]

Paolella, Marc S. 2018. *Fundamental Statistical Inference: A Computational Approach*. Chichester: John Wiley and Sons.

Trafimow, David. 2003. Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes's theorem. *Psychological Review* 110: 526–35. [CrossRef] [PubMed]

Trafimow, David. 2005. The ubiquitous Laplacian assumption: Reply to Lee and Wagenmakers. *Psychological Review* 112: 669–74. [CrossRef]

Trafimow, David. 2017. Using the coefficient of confidence to make the philosophical switch from *a posteriori* to *a priori* inferential statistics. *Educational and Psychological Measurement* 77: 831–54. [CrossRef] [PubMed]

Trafimow, David. 2018a. An *a priori* solution to the replication crisis. *Philosophical Psychology* 31: 1188–214. [CrossRef]

Trafimow, David. 2018b. Confidence intervals, precision and confounding. *New Ideas in Psychology* 50: 48–53. [CrossRef]

Trafimow, David. 2019a. My ban on null hypothesis significance testing and confidence intervals. In *Structural Changes and Their Economic Modeling*. Edited by Vladik Kreinovich and Songsak Sriboonchitta. Cham: Springer, pp. 35–48.

Trafimow, David. 2019b. What to do instead of null hypothesis significance testing or confidence intervals. In *Beyond Traditional Probabilistic Methods in Econometrics*. Edited by Vladik Kreinovich, Nguyen Ngoc Thack, Nguyen Duc Trung and Dang Van Thanh. Cham: Springer, pp. 113–28.

Trafimow, David, and Michiel de Boer. 2018. Measuring the strength of the evidence. *Biomedical Journal of Scientific and Technical Research* 6: 1–7. [CrossRef]

Trafimow, David, and Brian D. Earp. 2017. Null hypothesis significance testing and Type I error: The domain problem. *New Ideas in Psychology* 45: 19–27. [CrossRef]

Trafimow, David, and Justin A. MacDonald. 2017. Performing inferential statistics prior to data collection. *Educational and Psychological Measurement* 77: 204–19. [CrossRef]

Trafimow, David, and Hunter A. Myüz. Forthcoming. The sampling precision of research in five major areas of psychology. *Behavior Research Methods*.

Trafimow, David, and Stephen Rice. 2009. What if social scientists had reviewed great scientific works of the past? *Perspectives in Psychological Science* 4: 65–78. [CrossRef]

Trafimow, David, Valentin Amrhein, Corson N. Areshenkoff, Carlos J. Barrera-Causil, Eric J. Beh, Yusef K. Bilgic, Roser Bono, Michael T. Bradley, William M. Briggs, Héctor A Cepeda-Freyre, and et al. 2018. Manipulating the alpha level cannot cure significance testing. *Frontiers in Psychology* 9: 699. [CrossRef] [PubMed]

Trafimow, David, Tonhui Wang, and Cong Wang. 2019. From a sampling precision perspective, skewness is a friend and not an enemy! *Educational and Psychological Measurement* 79: 129–50. [CrossRef] [PubMed]

Trafimow, David, Cong Wang, and Tonghui Wang. Forthcoming. Making the a priori procedure (APP) work for differences between means. *Educational and Psychological Measurement*.

Trafimow, David. Forthcoming. A taxonomy of model assumptions on which P is based and implications for added benefit in the sciences. *International Journal of Social Research Methodology*. [CrossRef]

Wang, Cong, Tonghui Wang, David Trafimow, and Hunter A. Myüz. 2019a. Desired sample size for estimating the skewness under skew normal settings. In *Structural Changes and Their Economic Modeling*. Edited by Vladik Kreinovich and Songsak Sriboonchitta. Cham: Springer, pp. 152–62.

Wang, Cong, Tonghui Wang, David Trafimow, and Xiaoting Zhang. 2019b. Necessary sample size for estimating the scale parameter with specified closeness and confidence. *International Journal of Intelligent Technologies and Applied Statistics* 12: 17–29.

Wasserstein, Ronald. L., and Nicole A. Lazar. 2016. The ASA's statement on p-values: context, process, and purpose. *The American Statistician* 70: 129–33. [CrossRef]

econometrics

MDPI

*Short Note*

# On Using the *t*-Ratio as a Diagnostic

**Jan R. Magnus**

Department of Econometrics and Operations Research, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands; jan@janmagnus.nl

Received: 8 February 2019; Accepted: 28 May 2019; Published: 29 May 2019

**Abstract:** The *t*-ratio has not one but two uses in econometrics, which should be carefully distinguished. It is used as a test and also as a diagnostic. I emphasize that the commonly-used estimators are in fact pretest estimators, and argue in favor of an improved (continuous) version of pretesting, called model averaging.

**Keywords:** *t*-statistic; pretest estimator; model averaging

## 1. Introduction

The concept of 'statistical significance' appears in almost all scientific papers in order to form or strengthen conclusions, and the *p*-value or the *t*-ratio are commonly used to quantify this concept. Unfortunately, there is a lot of confusion about statistical significance. Almost 25 years have passed since Hugo Keuzenkamp and I wrote on this issue (Keuzenkamp and Magnus 1995), but the confusion persists and does not seem to diminish over time.

Most importantly, despite many warnings in textbooks, there is confusion about the difference between significance and importance. Statistical significance does not imply importance. This and other misuses of the *p*-value were recently well summarized by (Wasserstein and Lazar 2016).

In this note, which draws heavily on my recent undergraduate textbook (Magnus 2017), I concentrate on another (mis)use of the *t*-ratio (or of the *p*-value)—one which is not mentioned in (Wasserstein and Lazar 2016), but also needs attention and warning. This concerns the role of the *t*-ratio as a diagnostic. My aim is to explain that the *t*-ratio has not one but two uses in econometrics, which should be carefully distinguished; to emphasize (again) the difference between significance and importance; to show that the estimators that are used in practice are pretest (or post-selection) estimators (Leeb and Pötscher 2005); and to argue in favor of an improved (continuous) version of pretesting, called model averaging.

## 2. Two Uses of the *t*-Ratio

The *t*-ratio can be viewed in two ways. We could, for example, be interested in testing the hypothesis that $\beta_j = 0$ in the linear model $y = X\beta + u$. In that case the *t*-ratio $t_j$ can be fruitfully employed, because under certain assumptions (such as normality) $t_j$ follows Student's *t*-distribution under the null hypothesis and if we fix the significance level of the test (say at 5%) then we can reject or not reject the hypothesis.

The *t*-ratio, however, is also commonly employed in a different manner. Suppose we are primarily interested in the value of another $\beta$-coefficient, say $\beta_i$. Then, $t_j$ is often used as a diagnostic rather than as a test statistic in order to decide whether we wish to keep the $j$th regressor $x_j$ in the model or not. In this situation the 5% level is also typically used, but why? The two situations are quite different because in the first case we are interested in $\beta_j$ while in the second case we are interested in $\beta_i$. In the first case we ask: Is it true that $\beta_j = 0$? In the second case: Does inclusion of the $j$th regressor improve the estimator of $\beta_i$? These are two different questions and they require different approaches.

## 3. Significance and Importance

Suppose you are an econometrician working on a problem and some famous expert comes by, looks over your shoulder, and tells you that she knows the data-generation process (DGP). Of course, you yourself do not know the DGP. You use models but you do not know the truth; this expert does. Not only does the expert know the DGP but she is also willing to tell you, that is, she tells you the specification, not the actual parameter values. So now, you actually have the true model. What next? Is this the model that you are going to estimate?

The answer, surprisingly perhaps, is no. The truth, in general, is complex and contains many parameters, nonlinearities, and so on. All of these need to be estimated and this will produce large standard errors. There will be no bias if our model happens to coincide with the truth, but there will be large standard errors. A smaller model will have biased estimates but also smaller standard errors. Now, if we have a parameter in the true model whose value is small (so that the associated regressor is unimportant), then setting this parameter to zero will cause a small bias, because the size of the bias depends on the size of the deleted parameter. Setting this unimportant parameter to zero also means that we don't have to estimate it. The variance of the parameters of interest will therefore decrease, and this decrease does not depend on the size of the deleted parameter. Thus, deleting a small unimportant parameter from the model is generally a good idea, because we will incur a small bias but may gain much precision.

This is true even if the estimated parameter happens to be highly 'significant', that is, has a large $t$-ratio. Significance indicates that we have managed to estimate the parameter rather precisely, possibly because we have many observations. It does not mean that the parameter is important.

Note the proviso 'if our model happens to coincide with the truth' in the second paragraph. When we omit relevant variables we get biased estimators (which is bad), but a smaller variance (which is good). This, however, is only true when we compare the restricted model with an unrestricted model which coincides with the DGP. If, which is much more likely, we compare two models one of which is small (the restricted model) and the other is somewhat larger (the unrestricted model), but both are smaller than the DGP, then the estimator from the unrestricted model is also biased and, in fact, this bias may be larger than the bias from the restricted model; see (De Luca et al. 2018).

We should therefore omit from the model all aspects that have little impact, so that we end up with a small model—one, which captures the essence of our problem.

## 4. Pretesting

Let us consider the situation where $t_j$ is used as a diagnostic in more detail. In fact, we have three estimators of $\beta_i$: the estimator from the unrestricted model, $\hat{\beta}_{iu}$; the estimator from the restricted model (where $\beta_j = 0$), $\hat{\beta}_{ir}$; and the estimator after a preliminary test,

$$b_i = w\hat{\beta}_{iu} + (1-w)\hat{\beta}_{ir}, \qquad w = \begin{cases} 1 & \text{if } |t_j| > c, \\ 0 & \text{if } |t_j| \le c, \end{cases}$$

for some $c > 0$, such as $c = 1.96$ or $c = 1$. The estimator $b_i$ is called the pretest estimator.

The estimators $\hat{\beta}_{ir}$ and $\hat{\beta}_{iu}$ are linear and (under standard assumptions) normally distributed, but $b_i$ is nonlinear, because its distribution depends on a random restriction. The pretest estimator is therefore much more complicated than the other two estimators. But it is the pretest estimator that is commonly used in applied econometrics, because in applied econometrics we typically use $t$- and $F$-statistics as diagnostics to select the most suitable model. That in itself is not ideal, but what is worse is that we typically ignore the model selection aspect when reporting properties of our estimators.

The pretest estimator is kinked and therefore inadmissible. Its poor features are well-studied; see for example (Magnus 1999). Surely we should be able to come up with an estimator which performs better than the pretest estimator. This is where model averaging comes in.

## 5. Model Averaging

In (Magnus 2017) I tell the following story.

> A King has twelve advisors. He wishes to forecast next year's inflation and calls each of the advisors in for his or her opinion. He knows his advisors and obviously has more faith in some than in others. All twelve deliver their forecast, and the King is left with twelve numbers. How to choose from these twelve numbers? The King could argue: which advisor do I trust most, who do I believe is most competent? Then I take his or her advice. The King could also argue: all advisors have something useful to say, although not in the same degree. Some are more clever and better informed than others and their forecast should get a higher weight. Which way of thinking is better?

Intuitively most people, and I also, prefer the second method (model averaging), where all pieces of advice are taken into account. In standard econometrics, however, it is the first method (pretesting) which dominates.

There are theoretical and practical problems with the pretest estimator. One practical problem is the property that—if 1.96 is our cut-off point—for $t_j = 1.95$ we would choose one estimator and for $t_j = 1.97$ another, while in fact there is little difference between 1.95 and 1.97. This is not satisfactory.

These and other considerations lead us to reconsider the estimator

$$b_1 = w\hat{\beta}_{iu} + (1-w)\hat{\beta}_{ir}$$

by allowing $w$ to be a smoothly increasing function of $|t_j|$. This is model averaging in its simplest form, and we see that it is just the continuous counterpart to pretesting. In model averaging we give weight to all models of interest, but not in the same degree, while in pretesting we select one model after a preliminary test, precisely as the King in the story above.

In practice, econometricians use not one but many models. One of these is the largest and one is the smallest. Neither is probably the most suitable for the question at hand. If we use diagnostic tests to search for the best-fitting model, then we need to take into account not only the uncertainty of the estimates in the selected model, but also the fact that we have used the data to select a model. In other words, model selection and estimation should be seen as a combined effort, not as two separate efforts. This is what model averaging does. It incorporates the uncertainty arising from estimation and model selection jointly. Failure to do so may lead to misleadingly precise estimates.

**Funding:** This research received no external funding.

**Conflicts of Interest:** The author declares no conflicts of interest.

## References

De Luca, Giuseppe, Jan R. Magnus, and Franco Peracchi. 2018. Balanced variable addition in linear models. *Journal of Economic Surveys* 32: 1183–200. [CrossRef]

Keuzenkamp, Hugo A., and Jan R. Magnus. 1995. On tests and significance in econometrics. *Journal of Econometrics* 67: 5–24. [CrossRef]

Leeb, Hannes, and Benedikt M. Pötscher. 2005. Model selection and inference: Facts and fiction. *Econometric Theory* 21: 21–59. [CrossRef]

Magnus, Jan R. 1999. The traditional pretest estimator. *Theory of Probability and Its Applications* 44: 293–308. [CrossRef]

Magnus, Jan R. 2017. *Introduction to the Theory of Econometrics*. Amsterdam: VU University Press.

Wasserstein, Ronald L., and Nicole A. Lazar. 2016. The ASA's statement on $p$-values: Context, process, and purpose. *The American Statistician* 70: 129–33. [CrossRef]

*econometrics*

MDPI

*Article*

# Interval-Based Hypothesis Testing and Its Applications to Economics and Finance

**Jae H. Kim [1,*] and Andrew P. Robinson [2]**

[1] Department of Economics and Finance, La Trobe University, Bundoora, VIC 3086, Australia

[2] School of Mathematics and Statistics, University of Melbourne, Parkville, VIC 3010, Australia; apro@unimelb.edu.au

* Correspondence: j.kim@latrobe.edu.au

Received: 26 March 2019; Accepted: 7 May 2019; Published: 15 May 2019

**Abstract:** This paper presents a brief review of interval-based hypothesis testing, widely used in bio-statistics, medical science, and psychology, namely, tests for minimum-effect, equivalence, and non-inferiority. We present the methods in the contexts of a one-sample $t$-test and a test for linear restrictions in a regression. We present applications in testing for market efficiency, validity of asset-pricing models, and persistence of economic time series. We argue that, from the point of view of economics and finance, interval-based hypothesis testing provides more sensible inferential outcomes than those based on point-null hypothesis. We propose that interval-based tests be routinely employed in empirical research in business, as an alternative to point null hypothesis testing, especially in the new era of big data.

**Keywords:** equivalence; minimum-effect; non-inferiority; point-null hypothesis testing; zero probability paradox

**JEL Classification:** C12

---

> *Genuinely interesting hypotheses are neighbourhoods, not points. No parameter is exactly equal to zero; many may be so close that we can act as if they were zero.*
>
> Edward Leamer (1988)

## 1. Introduction

The paradigm of point null hypothesis testing has been almost exclusively adopted in all areas of empirical research in business, including accounting, economics, finance, management, and marketing. The procedure involves forming a sharp null hypothesis (typically the value of a parameter equal to zero, to represent no effect) and using the "$p$-value less than $\alpha$" criterion to reject or fail to reject the null hypothesis, or in the Neyman–Pearson tradition, determining whether the test statistic lies in a region defined by $\alpha$, the test size. Although the alternative hypothesis is often unspecified, the rejection of a null hypothesis of no effect is frequently taken as evidence for the existence of a non-zero effect.

As a hybrid of Fisher's approach to significance testing and Neyman–Pearson decision-theoretic approach, the procedure is often conducted in an automatic manner without considering the key factors of statistical research, such as effect size, statistical power, relative loss, and prior beliefs (see, for example, Kim and Choi 2019). This practice has been criticized by many authors, for example, Gigerenzer (2004) calls it the "null ritual"; while McCloskey and Ziliak (1996) warn against widespread practice of "asterisk econometrics" and "sign econometrics". Despite numerous calls for change for years, little improvement has been made in the practice of "mindless statistics" (Gigerenzer 2004). The consequences include serious distortion of scientific process (Wasserstein and Lazar 2016), an embarrassing number of false positives (Kim and Ji 2015; Harvey 2017; Kim et al. 2018) replication crises in many fields of science (see, for example, Open Science Collaboration 2015), and publication bias (Basu and Park 2014; Kim and Ji 2015).

With increasing availability of large or massive data sets in the business disciplines in recent years, the current paradigm has become even more problematic, and arguably deficient. This is because, in reality, any null hypothesis is violated even when it is (practically or economically) true (see, for example, De Long and Lang 1992). Rao and Lovric (2016) call this phenomenon the *zero-probability paradox*, providing a mathematical proof for a simple case. Its consequence is that the $p$-value is a decreasing function of sample size, even when the null hypothesis is violated by an economically or scientifically negligible margin (see Kim and Ji 2015). As a result, the probability of a false positive increases with sample size, as also noted by Ohlson (2018). As Spanos (2017) points out, there is nothing paradoxical about this, since it is a reflection of the consistency property of a test. As Kim and Ji (2015) and Kim et al. (2018) report from their respective meta-analytic surveys, many empirical researchers routinely adopt large or massive samples under the current paradigm, with a high chance that their scientific findings represent false positives. It is also problematic in the context of model specification testing, since any model may be judged to be mis-specified when the sample size is large enough (Spanos 2017).

In view of the above points, Rao and Lovric (2016) argue that "in the 21st century, statisticians will deal with large data sets and complex questions, it is clear that the current point-null paradigm is inadequate" and that "next generation of statisticians must construct new tools for massive data sets since the current ones are severely limited" (see also van der Laan and Rose 2010). They call for a paradigm shift in statistical hypothesis testing and suggest the Hodges and Lehmann (1954) paradigm as a possible alternative, arguing that this will substantially improve the credibility of scientific research based on statistical testing. Under the Hodges and Lehmann (1954) paradigm, the null and alternative hypotheses are formulated as *intervals*. The focus of testing is whether the parameter value belongs to an interval of no practical (or economic) significance, with its limits set by the researcher based on substantive importance. In this way, the researcher's economic reasoning or judgment can be incorporated into hypothesis testing.

In fact, the tests for interval-based hypotheses have been in existence and being used in biostatistics and psychology under the name of equivalence tests, non-inferiority tests, and minimum tests: see, for comprehensive and in-depth reviews, Wellek (2010), Murphy et al. (2014), and Lehmann and Romano (2005, sct. 13.5.2). However, the researchers in the business disciplines have little knowledge about these tests, especially those who are engaged in empirical or applied research. The purpose of this paper is to present a brief review of these tests to the researchers in business, discussing their merits and otherwise. The tests are also presented for parameter restrictions and model specification in the linear regression context, incorporating the bootstrap method. The tests are presented with three empirical applications in economics and finance. We propose that these tests be routinely employed in business research as an alternative to point null hypothesis testing. We hope that this will contribute to a paradigm shift in statistical inference, which will restore credibility and integrity in statistical research in business disciplines.

In the next section, we briefly discuss the current paradigm of point null hypothesis and its problems and consequences. In Section 3, we present a review of equivalence, non-inferiority, and minimum-effect tests for the simple $t$-test and regression $F$-test. Section 4 provides empirical applications, and Section 5 concludes the paper.

## 2. Current Paradigm and Its Deficiencies

We begin by presenting the current (frequentist) paradigm of hypothesis testing, which is widely adopted in many areas of statistical research, in the context of a simple $t$-test for a point null hypothesis. This is followed by a review of its deficiencies as a criterion of statistical evidence. We also review the problems and malpractices such as $p$-hacking and data-mining and how they are related with the current paradigm of statistical inference.

### 2.1. A Simple t-Test for a Point Null Hypothesis

Consider the case of a simple one-sample $t$-test for the population mean $\theta$, where $X_i$ $(i = 1, \dots, n)$ is independently generated from a normal distribution with mean $\theta$ and standard deviation $\sigma$. Applying the point null hypothesis paradigm, we test (assuming two-tailed alternative) for

$$H_0 : \theta = 0; H_1 : \theta \neq 0.$$

The null hypothesis most often represents the claim of "no effect". When $H_0$ is true (hereafter, *under* $H_0$), the $t$-statistic follows a $t$-distribution; while under $H_1$, it follows a non-central $t$-distribution with the non-centrality parameter $\sqrt{n}\theta/\sigma$. The decision to reject or fail to reject depends on the "$p$-value less than $\alpha$" criterion where $p$-value $\equiv Prob(|t| > t_{c,1-0.5\alpha}|H_0)$ and $t_{c,1-0.5\alpha}$ is the critical value from a central $t$-distribution at the $\alpha$ level of significance. The value of $\alpha$ conventionally adopted is 0.05, although values such as 0.01 or 0.10 are often used. When the $p$-value satisfies the criterion, the effect is said to be statistically significant at the $\alpha$ level of significance. This is what Gigerenzer (2004) calls the "null ritual", which is a hybrid of the proposal of Fisher and that of Neyman and Pearson. In practical applications, a small $p$-value is often interpreted as a strong evidence against $H_0$ and its strength is marked with the number of asterisks indicating the significance at a 0.10, 0.05, or 0.01 level of significance. More seriously, many researchers do not pay attention to the magnitude of the $\theta$ estimate, making their decisions based only on its sign and statistical significance. This practice has been branded as "asterisk econometrics" and "sign econometrics" by Ziliak and McCloskey (2008), who correctly argue that it only shows whether the effect exists or not, but nothing about economic significance or substantive importance of the effect (see also Kleijnen 1995).

### 2.2. Shortcomings of the p-Value Criterion

It is well-known that the $p$-value is not a good measure of evidence for a hypothesis. For example, Berger and Sellke (1987) shows that the $p$-value provides a measure of evidence against $H_0$ that can differ from the actual value by an order of magnitude. Johnstone and Lindley (1995) demonstrates that a $p$-value less than 0.05 may represent evidence in favor of the null, not against it, especially when the sample size is large (see, also, Kim et al. 2018). It is largely because the $p$-value does not take account of the probability under $H_1$; nor does it represent the probability that null is true given data. On this basis, the American Statistical Association expressed grave concerns against the misuse or abuse of the $p$-value criterion in empirical research, stating that this practice has led to a considerable distortion of the scientific process (Wasserstein and Lazar 2016).

Another problem of the $p$-value criterion is that the choice of its threshold $\alpha$ is arbitrary (Keuzenkamp and Magnus 1995; Lehmann and Romano 2005, p. 57). As Arrow (1960) and Leamer (1978) argue, it should be chosen in consideration of the key factors such as sample size, statistical power, and relative loss from Type I and II errors. For example, the level of significance should be set at a range of 0.3 to 0.4 when the power is low (Winer 1962); while is should be set at a small a value (such as 0.001) when the sample size is large (McCloskey and Ziliak 1996, p. 102). This is to balance the probabilities of Type I and II errors when the losses from Type I and II errors are (almost) equal. Kim and Choi (2017, 2019) provided a review of a decision-theoretic approach to the optimal level of significance with applications.

### 2.3. Zero-Probability Paradox

In practice, the null hypothesis cannot hold exactly, as shown by Rao and Lovric (2016). As Leamer (1988), De Long and Lang (1992), and Startz (2014) point out, an economic hypothesis should not be formulated as a point, but as a neighborhood or an interval since an economic effect (or parameter) cannot take a numerically exact value such as 0. The consequence is that, with observational data, the distribution under a point $H_0$ is never observed nor realized; but the $t$-statistic is always generated from the distribution under $H_1$, which is a non-central $t$-distribution. This is another reason that makes the $p$-value criterion deficient because the critical value $t_{c,\alpha}$ is obtained from a central $t$-distribution which is never observed in practice.

The problem is exacerbated as the sample size increases, because the non-centrality of the $t$-distribution ($\sqrt{n}\theta/\sigma$) also sharply increases, meaning that the $p$-value approaches 0. This occurs even when the true value of $\theta$ is practically or economically no different from 0. When the sample size is large, this distribution is so far away from the central $t$-distribution. Hence, when $H_0$ is numerically violated but it holds practically, rejection of $H_0$ occurs with certainty in large samples, as long as the level of significance $\alpha$ is maintained at a conventional value such as 0.05. In practice, many empirical researchers often take an economically negligible violation of $H_0$ as evidence for particular alternative hypothesis, committing what is called the "fallacy of rejection" (Spanos 2017). A natural solution in this context is to obtain the critical value from a non-central distribution under $H_1$, which increases with sample size. In fact, this is a proposal of interval-based hypothesis testing, as we shall see in the next section.

### *2.4. Problems and Consequences*

The deficiency and weakness of the $p$-value criterion discussed above have created a number of problems and malpractices, namely $p$-hacking (Harvey 2017), data mining (Black 1993) or data snooping (Lo and MacKinlay 1990). They generally refer to the practice of cherry-picking the results in order to achieve statistically significant outcome. Black (1993, p. 75) provides a good description of data mining:

> When a researcher tries many ways to do a study, including various combinations of explanatory factors, various periods, and various models, we often say, he is "data mining." If he reports only the more successful runs, we have a hard time interpreting any statistical analysis he does. We worry that he selected, from the many models tried, only the ones that seem to support his conclusions. With enough data mining, all the results that seem significant could be just accidental.

A consequence is an embarrassing number of false positives, as Harvey (2017) puts it. As Kim and Ji (2015), Kim et al. (2018) and Kim (2019) report, the use of alternative criteria for statistical significance (such as Bayes factors, adaptive or optimal levels of significance, or posterior probabilities for null hypotheses) gives different inferential outcomes from the $p$-value criterion in a large number of published results. This may have led to accumulation of many false stylized facts in empirical studies. For example, Black (1993) argues that most of investment anomalies identified in finance are likely to be the result of data-mining; while Kandel and Stambaugh (1996) argue that the $p$-value as measure of evidence often conflicts with economic significance in asset-allocation decisions. Kim and Choi (2017) report that many economically puzzling research outcomes (such as empirical invalidity of the purchasing power parity) based on unit root testing may be the result of incorrectly maintaining the conventional level of significance, despite extremely low power of the test. In behavioral finance, it is a stylized fact that the weather affects stock market (see, for example, Saunders 1993; Hirshleifer and Shumway 2003). However, as Kim (2017) argues, this statistical significance is the result of having power of practically one due to massive sample size. In a similar context, Kamstra et al. (2003) report the statistically significant effect of winter blues on stock market (as discussed in Section 4.1 as an application), while they find statistically insignificant effect of weather variables in the same equation. This conflicting result may be the outcome of data-mining, where statistical significance is purely accidental.

Abuse and misuse of the $p$-value criterion for statistical significance also have contributed to other serious problems which undermines the research integrity and credibility in science: namely, publication bias and replication crisis. The practice of $p$-hacking and data-mining is closely related with publication bias where statistically significant results are favored in the publication process. The meta-analytic evaluation of Kim and Ji (2015) and Kim et al. (2018) reveal unreasonably high proportions of studies published in accounting and finance journals are statistically significant. Harvey (2017) also recognizes the practice of $p$-hacking can contribute to publication bias. This is partly because many journal editors and referees favor statistically significant results, and often judge statistically insignificant studies with skepticism and suspicion. As a consequence, many studies with statistically insignificant results (at a conventional significance level) may not have been published, even though they are economically important and statistically sound. This practice can push many researchers to the malpractice of $p$-hacking or data-mining to gain higher chance of publication. "Replication crisis" refers to the problem that a high proportion of published results are not reproducible by replication exercises (Peng 2015). For example, in psychology, only 36% of the replications are found to be statistically significant, compared to 97% of the original studies that reported significance (Open Science Collaboration 2015).

As discussed in this section, the current paradigm of statistical inference has a number of problems, and has contributed to a range of serious issues that undermine research integrity and credibility. On this basis, Rao and Lovric (2016) call for a new paradigm for statistical inference, especially needed in the big data era where the $p$-value fails as a measure of statistical evidence and the conventional level of significance is inappropriate. They suggest an interval-based test as a possible alternative, which will be discussed in the next section.

## 3. Tests for Minimum-Effect, Equivalence, and Non-Inferiority

We now present the three types of interval tests, namely the equivalence, minimum-effect, and non-inferiority tests, based on the well known one-sample $t$-test or test for linear restrictions in the linear regression. Loosely speaking, the difference between the equivalence and minimum-effect tests comes down to the condition for which proof is being sought. If the status quo conjecture is characterized by equality, that is, the conjecture against which we wish to assess evidence is that one thing equals another, then we falsify the conjecture by the minimum-effect test. On the other hand, if the status quo conjecture is characterized by inequality, so the conjecture against which we wish to assess evidence is that things are unequal, then we falsify using an equivalence test. A non-inferiority test may be used when the hypothesis formulated as an open interval.

### *3.1. Test for Minimum Effect*

The minimum-effect test, originally put forward by Hodges and Lehmann (1954), has the null and alternative hypotheses of the following form:

$$H_0 : \theta_l \leq \theta \leq \theta_u; H_1 : (\theta < \theta_l) \cup (\theta_u < \theta), \tag{1}$$

where $\theta_l$ and $\theta_u$ denote the limits of practical or economic importance. Hodges and Lehmann (1954, p. 254) propose conducting separate one-tailed $t$-tests of the two one-sided hypotheses. That is,

- $H_{01} : \theta \leq \theta_u$ against alternative $H_{11} : \theta > \theta_u$ and
- $H_{02} : \theta \geq \theta_l$ against alternative $H_{12} : \theta < \theta_l$.

According to Hodges and Lehmann (1954, p. 254), we then reject $H_0$ given in (1) if either of these separate tests rejects. The size of this composite $t$-test is the sum of their separate sizes. The power of the test should depend on the power of the individual one-tailed tests associated. The decision can also be made by using the confidence interval: the null hypothesis of minimum-effect given in (1) cannot be rejected at the $\alpha$ level of significance if a two-sided $(1 - 2\alpha)$ confidence interval for $\theta$ lies entirely within the interval $[\theta_l, \theta_u]$.

Even though the interval test extends the simple $t$-test, the intention is the same: to detect a statistically significant and important difference. Rejection of the test is interpreted as a failure to detect such a difference—a failure to split. We now review tests that have the opposite effect: to detect a statistically significant and important similarity. These tests are equivalence tests. Rejection of these tests is interpreted as a failure to detect such a similarity—a failure to lump.

### 3.2. Test for Equivalence

If we switch the null and alternative hypotheses, we have what is called an equivalence test (e.g., Wellek 2010). That is,

$$H_0 : (\theta \leq \theta_l) \cup (\theta_u \leq \theta); H_1 : \theta_l < \theta < \theta_u. \tag{2}$$

The decision rule for the equivalence test can be developed by conducting two one-sided test procedures similarly to the above, which is referred to as TOST:

- $H_{01} : \theta \leq \theta_l$ against alternative $H_{11} : \theta > \theta_l$ and
- $H_{02} : \theta \geq \theta_u$ against alternative $H_{12} : \theta < \theta_u$.

Let $p_1$ be the one-sided $p$-value for the test of $H_{01}$ against $H_{11}$; and $p_2$ be the same for for the test of $H_{02}$ against $H_{12}$. For the equivalence test, the null hypothesis of no equivalence given in (2) is rejected at the $\alpha$ level of significance if $max(p_1, p_2) < \alpha$. Equivalently, it is rejected at the $\alpha$ level of significance if a two-sided $(1 - 2\alpha)$ confidence interval for $\theta$ lies entirely within the interval $[\theta_l, \theta_u]$. The power of the test should depend on the power of the individual one-tailed tests associated.

Note that the researcher should choose between the minimum-effect test and equivalence test by considering whether the evidence being sought is against similarity (minimum effect test) or difference (equivalence test). It is worth mentioning that, as the sample size increases, the confidence interval shrinks but the limits of economic significance do not change. For interval-based tests, this can be interpreted as the critical values increasing with sample size, relative to the test statistic, which is a feature not shared by point-null hypothesis testing. It is also worth mentioning that the minimum effect and equivalence tests give mutually exclusive results in that one always rejects and the other always does not, as long as the two tests share the same limits of economic importance.

### 3.3. Test for Non-Inferiority

It is often the case that testing for a one-sided (open) interval may be appropriate. The test is called the non-inferiority test or superiority test, whose null and alternative hypotheses can be written as

$$H_0 : \theta \geq \theta_l; H_1 : \theta < \theta_l, \tag{3}$$

where $\theta_l$ denotes the smallest effect size of economics importance. The non-inferiority test tests whether the null hypothesis that an effect is at least as large as $\theta_l$ can be rejected. The actual direction of the hypothesis depends on whether a higher value of the response is desirable or not. The above test can be conducted as a usual one-tailed test.

### 3.4. Interval Tests in the Linear Regression Model

Following Hodges and Lehmann (1954), Murphy and Myors (1999) approach the minimum-effect using the $F$-test, which can be presented in a regression context. In this subsection, we review their proposal and extend it to a more general setting.

Consider a regression model of the form

$$Y = \gamma_0 + \gamma_1 X_1 + ... + \gamma_K X_K + u, \tag{4}$$

where $Y$ is a dependent variable and $X$'s are independent variables. Suppose the researcher tests for a linear restriction such as $H_0 : \gamma_1 = ... = \gamma_J = 0$, where $J \leq K$. The $F$-statistic can be written as

$$F = \frac{(R_1^2 - R_0^2)/J}{(1 - R_1^2)/(T - K - 1)}, \tag{5}$$

where $R_j^2$ represents the coefficient of determination under $H_j$ ($j$ = 0, 1). Under $H_0$, the $F$-statistic follows the $F$-distribution with $J$ and $T - K - 1$ degrees of freedom, denoted as $F\,(J, T - K - 1)$. Under $H_1$, the $F$-statistic follows $F(J, T - K - 1; \lambda)$, which denotes the non-central $F$-distribution with the degrees of freedom $(J, T - K - 1)$ and the non-centrality parameter $\lambda$. Note that

$$\lambda = T\frac{R_{p1}^2 - R_{p0}^2}{1 - R_{p1}^2} \equiv T\eta, \tag{6}$$

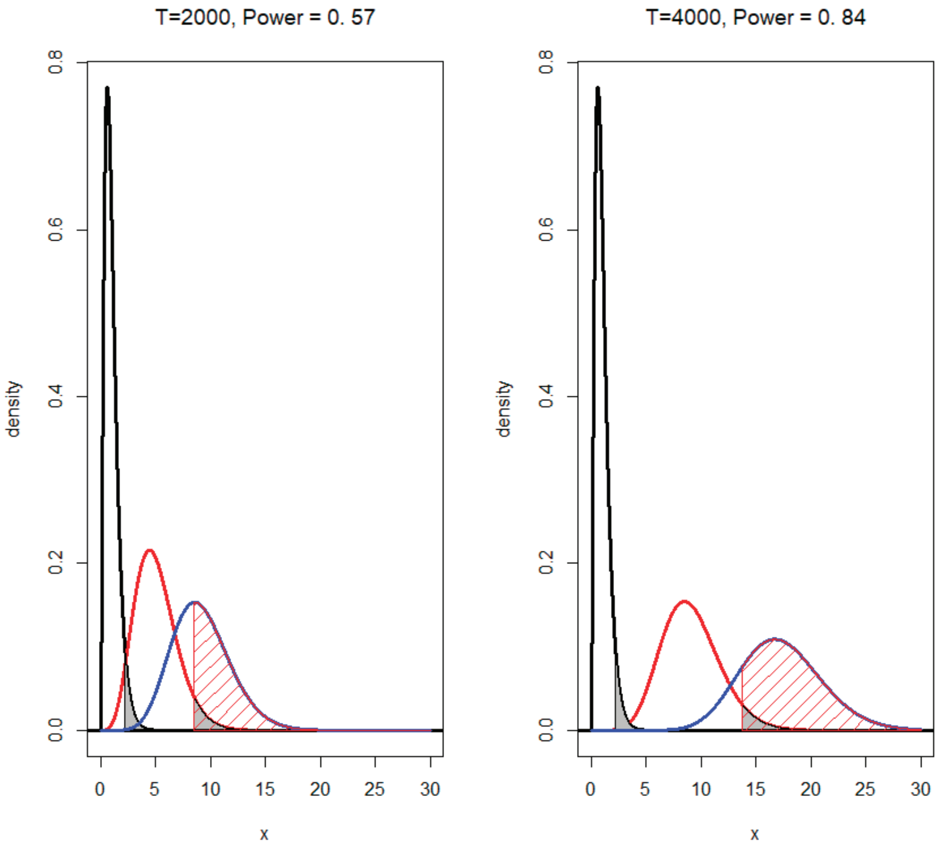
where $R_{pj}^2$ denotes the population or desired coefficient of determination under $H_j$, following from Peracchi (2001, Theorem 9.2). Note that $\eta \equiv (R_{p1}^2 - R_{p0}^2)/(1 - R_{p1}^2)$ may be called the population signal-to-noise ratio, measuring the incremental contribution of $(X_1, \ldots, X_J)$ relative to the noise to the model. The degree of non-centrality is determined as a product of sample size and signal-to-noise ratio, with the former playing a dominant role.

Hodges and Lehmann (1954, p. 253) and Murphy and Myors (1999) propose that the above non-central distribution be used to test for the minimum-effect test. As an example, consider a simple regression model $Y = \gamma_0 + \gamma_1 X_1 + u$ with $H_0 : \gamma_1 = 0$. Here, $R_{p1}^2$ measures the incremental contribution of $X_1$ for $Y$ (note that $R_{p0}^2 = 0$). The researcher wishes to test for $H_0 : 0 \leq \gamma_1 \leq \gamma_u$, where $\gamma_u$ represents the limit for the minimum-effect. The researcher can also specify the value of $R_{p1}^2$ corresponding to the value of $\gamma_u$, which is the minimum desired value of $R^2$ for $X_1$ to be economically significant (see Section 3.8 for the details as to how this value may be chosen with applications in Section 4.1). Alternatively to $H_0 : 0 \leq \gamma_1 \leq \gamma_u$, one can formulate the null hypothesis in terms of $R_{p1}^2$, namely $H_0 : 0 \leq R_{p1}^2 \leq R_{max}^2$, where $R_{max}^2$ is the maximum of $R_{p1}^2$ value for $0 \leq \gamma_1 \leq \gamma_u$, given $(Y, X_1)$; and also $0 < \lambda \leq \lambda_{max}$ corresponding to $0 \leq R_{p1}^2 \leq R_{max}^2$.

If the $F$-statistic is greater than $F_{\alpha,\lambda_{max}}$, the $\alpha$-level critical value from $F(J, T - K - 1; \lambda_{max})$, then the null hypothesis of the minimum-effect is rejected at the $\alpha$-level of significance. An interesting feature of the decision rule for the minimum-effect test is that its critical value and sampling distribution change with sample size. This is in stark contrast with those of the point-null hypothesis, which do not change with sample size. The latter property is the root cause of the "large-n problem" associated with the point-null hypothesis, as Rao and Lovric (2016) point out.

As an illustration, consider a regression where $K = 1$. For simplicity, we assume that $Var(X_1) = Var(Y)$, when the sample size $T$ takes values 2000 and 4000. Consider first the case of point-null hypothesis where $H_0 : \gamma_1 = 0$. The black curves in Figure 1 plot the density $F(J, T - K - 1)$, which is the distribution of the $F$-statistic under $H_0 : \gamma_1 = 0$, for each sample size of 2000 and 4000. It is clear that the 5% critical value does not change with increasing sample size. Since $F$-statistic is an increasing function of sample size, rejection of $H_0 : \gamma_1 = 0$ will eventually occur (except of course for the rare case that the true value of $\gamma_1$ is really numerically identical to zero).

Suppose the researcher tests for a minimum effect: $H_0 : 0 \leq \gamma_1 \leq 0.1$ against $H_1 : \gamma_1 > 0.1$. Since $R_{p1}^2 = \gamma_1^2 Var(X_1)/Var(Y)$ and $R_{p0}^2 = 0$, the null and alternative hypotheses can be formulated as $H_0 : R_{p1}^2 \leq 0.01$ against $H_1 : R_{p1}^2 > 0.01$. The red curves in Figure 1 plot the density $F(J, T - K - 1; \lambda_{max})$ associated with $R_{p1}^2 = 0.01$ for each sample size. The gray under area under represents 5%, indicated by the critical value which is the 95th percentile of the red curve. It appears that this critical value increases with sample size. The blue curve plots the density $F(J, T - K - 1; \lambda)$ associated with $H_1 : R_{p1}^2 = 0.02$ and the red shaded area represents the power of the test for $H_0 : R_{p1}^2 \leq 0.01$. It shows that the power increases with sample size.



Note: The black curve plots the density $F(J, T - K - 1)$ which is the distribution of the $F$-statistic under $H_0 : \gamma_1 = 0$. The gray area under it represents 5% associated with the corresponding critical value. The red curve plots the density $F(J, T - K - 1; \lambda_{max})$ which is the distribution of the $F$-statistics under $H_0 : \gamma_1 \leq 0.1$ or $H_0 : R^2_{p1} \leq 0.01$.The gray area under it represents 5% associated with the corresponding critical value. The blue curve plots the densit y $F(J, T - K - 1; \lambda)$ for $H_1 : R^2_{p1} = 0.02$. The red-shaded are represents the power of thetest for $H_0 : R^2_{p1} \leq 0.01$.

**Figure 1.** Test for minimum-effect: An illustration.

It is often the case in economics and finance that a test for linear restrictions is conducted involving a number of regression parameters. For example, under the point-null paradigm, the null hypothesis can be formulated as $H_0 : \gamma_1 = \gamma_2 = 0$ for a regression of $Y$ on $X_1$ and $X_2$. In the context of minimum-effect test, the null hypothesis can be written as

$$H_0 : (\gamma_{1l} \leq \gamma_1 \leq \gamma_{1u}) \cup (\gamma_{2l} \leq \gamma_2 \leq \gamma_{2u}),$$

where $\gamma_{il}$ and $\gamma_{iu}$ for $(i = 1, 2)$ denote the boundaries of economic significance. In this case, the null hypothesis can be formulated in terms of $R^2_{pj}$. That is,

$$H_0 : \eta \leq \eta_{max},$$

where $\eta_{max}$ is the maximum population signal-to-noise ratio implied by $R^2_{pj}$. The researcher can formulate the value of $R^2_{p1} - R^2_{p0}$ as the economically significant incremental contribution of $(X_1, X_2)$ to $Y$, where the value of $R^2_{p0}$ can be estimated from the regression with restriction $\gamma_1 = \gamma_2 = 0$. Let $\lambda_{max} = T\eta_{max}$, then if the $F$-statistic is greater than $F_{\alpha,\lambda_{max}}$, the $\alpha$-level critical value from $F(J, T-K-1; \lambda_{max})$, the null hypothesis of the minimum-effect is rejected at the $\alpha$-level of significance. An example in a more general setting can be found in Section 4.1.2.

### 3.5. Bootstrap Implementation

The tests introduced so far are valid under the assumption of normality. When the assumption of normality is questionable, the one-tailed tests, confidence intervals and the distribution $F(J, T-K-1; \lambda)$ can be implemented using the bootstrap (Efron and Tibshirani 1994). Since there are extensive references available for bootstrapping the $p$-value and confidence intervals for a one-sample $t$-test, the details are not given here.

For the minimum-effect test in the linear regression model, the researcher may want to obtain the bootstrap counterpart of the red curve in Figure 1, when the underlying normality is questionable. Consider a simple case of $H_0 : 0 \leq \gamma_1 \leq 0.1$ against $H_1 : \gamma_1 > 0.1$. Since $\gamma_1 = 0.1$ is associated with the maximum value of $R^2_{p1}$ of 0.01, we consider the regression model under the restriction $\gamma_1 = 0.1$. That is,

$$Y = \hat{\gamma}_0 + 0.1X_1 + e,$$

where $\hat{\gamma}_0$ is the estimator for $\gamma_0$ under the restriction $\gamma_1 = 0.1$ and $e$ represents the associated residuals. Generate the artificial data $Y^*$ given $X_1$ as

$$Y^* = \hat{\gamma}_0 + 0.1X_1 + e^*,$$

where $e^*$ is a random resample of $e$ with replacement. Calculate the $F$-statistic from $\{Y^*, X_1\}$, denoted as $F^*$. Repeat the above process sufficiently many times, say $B$, to obtain $\{F^*(i)\}_{i=1}^{B}$, which represents the bootstrap distribution for $F(J, T-K-1; \lambda)$.

When a number of parameters are involved with the linear restrictions being tested, the bootstrap can be conducted at the parameter values which maximize the value $\lambda$. As an example, consider the minimum-effect test

$$H_0 : (\gamma_{1l} \leq \gamma_1 \leq \gamma_{1u}) \cup (\gamma_{2l} \leq \gamma_2 \leq \gamma_{2u}).$$

Let $\hat{\gamma}_1$ and $\hat{\gamma}_2$ denote the values under the above $H_0$ which jointly imply the largest economic impact on $Y$. The bootstrap is conducted with the restrictions $\gamma_1 = \hat{\gamma}_1$ and $\gamma_2 = \hat{\gamma}_2$.

### 3.6. Model Equivalence Test

Lavergne (2014) proposes a general framework based on the Kullback-Leibler information to assess the approximate validity of multivariate restrictions in parametric models, which is labeled as model equivalence testing. Consider a random sample $X_t$ $(t = 1, \ldots T)$ whose probability density function is denoted as $f(X|\theta_0)$ where $\theta_0 \in \Theta$ the parameter space. Let $g(\theta_0) = 0$ denote multivariate restrictions on $\theta_0$ with $r$ number of restrictions. As a measure of closeness to the true distribution, Lavergne (2014) adopts the Kullback-Leibler information criterion, which is defined as

$$KLIC = E_{\theta_0}\left[\log \frac{f(X|\theta_0)}{f(X|\theta_0^c)}\right],$$

where $E_{\theta_0}$ denotes the expectation when $\theta_0$ is the parameter value and $\theta_0^c$ is the value which maximizes $E_{\theta_0} \log f(X|\theta_0)$ under $g(\theta_0^c) = 0$. Noting that $KLIC \geq 0$ and it is 0 when the restriction $g(\theta_0) = 0$ holds exactly, Lavergne (2014) considers the null and alternative hypotheses of the form

$$H_0 : 2KLIC \geq \delta^2/T; H_1 : 2KLIC < \delta^2/T, \tag{7}$$

where $\delta^2 \equiv T\Delta^2$ while $\Delta^2$ being the tolerance of substantive importance. Rejection of $H_0$ implies that the restriction $g(\theta_0) = 0$ is close to be valid.

According to Lavergne (2014), the above model equivalence test can be conducted using the log-likelihood ratio (LR) test, which can be written as

$$LR = 2\left[L(\hat{\theta}) - L(\hat{\theta}^c)\right], \tag{8}$$

where $\hat{\theta}$ denotes the unrestricted (quasi) maximum likelihood estimator for $\theta$ and $\hat{\theta}^c$ the restricted (quasi) maximum likelihood estimator. The LR statistic follows a non-central chi-squared distribution with $r$ degrees of freedom with the non-centrality parameter $\delta^2$, denoted as $\chi^2_{r,\delta^2}$. The null hypothesis is rejected in favor of model equivalence if the LR statistic is less than $\chi^2_{r,\delta^2}(\alpha)$, which is the $\alpha$th percentile of $\chi^2_{r,\delta^2}$.

Note that the vanishing tolerance $\delta^2/T$ is based on a theoretical consideration, as Lavergne (2014, p. 416) points out. In practical applications, a fixed tolerance $\Delta^2$ is chosen so that $\delta^2 = T\Delta^2$. This means that the degree of non-centrality of $\chi^2_{r,\delta^2}$ increases with sample size, so does the critical value of the test. This is a feature different from the point-null hypothesis testing where the critical value is obtained from a central distribution regardless of sample size. Lavergne (2014) has shown that, in the regression context, $2KLIC$ measures the loss in explanatory power coming from imposing the constraint relative to the error's variance. Hence, if the researcher sets $\Delta^2 = 0.1$, the models under $H_0$ and $H_1$ are considered to be equivalent if the loss of explanatory power due to imposing the restriction is no more than 10%. Lavergne (2014) provides further asymptotic theories of the test, along with empirical applications.

### *3.7. Equivalence Test for Model Validation*

Model validation or specification tests are often performed based on the paradigm of point null hypothesis testing, for which the null hypothesis is that the model is valid, and the alternative hypothesis is that the model is not valid. Such tests inherit the problems associated with the conventional statistical testing. As Box (1976) points out, all models are wrong, they are approximations to the true data-generation process; consequently a test based on a sharp null hypothesis is not suitable. It is possible that, in small samples, the tests may commit Type II errors due to low power, whereas in large samples all models are found to be rejected due to extreme power (see, for example, Spanos 2017).

As a consequence, Robinson and Froese (2004) recommended the use of equivalence tests for model validation, arguing that using traditional point-null hypothesis testing, as commonly done, enabled the rejection of good models when the data were too many and the failure to reject poor models when the data were too few. Furthermore, equivalence tests permit the expression of a 'region of equivalence', within which model predictions could be close enough to reality to be useful, without necessarily being exactly identical (see, e.g., Kleijnen 1995; Robinson 2019). The principle was further extended by Robinson et al. (2005), who produced an equivalence-based variant of a regression-style test originally proposed by Cohen and Cyert (1961). We now summarize Robinson et al.'s (2005) approach.

Assume that we have computer simulation results $x_i$, $i = i, \ldots, n$ that are intended to represent process observations $y_i$. For example, $y$ could be the heights of a sample of trees selected from a forest, and $x$ the predicted heights for the same trees having been computed using the tree diameter and some mathematical function that we wish to validate; $\hat{y} = x = f(d; \beta)$. Centre the predictions: $x_i^* = x_i - \bar{x}$.

Fit the linear regression model $y_i = \beta_0 + \beta_1 x_i^* + \epsilon_i;\ \epsilon_i \overset{d}{=} N(0, \sigma^2)$. Then, perform a TOST on the null hypothesis that $\beta_0 \neq 0$ as a test of model *bias* and a TOST on the null hypothesis that $\beta_1 \neq 1$ as a test of the model *fidelity*, where fidelity is taken to mean both the spread of the predictions compared to the observations and the order of the predictions compared to the observations. The estimate of the slope will reflect how well the predictions match the spread of the observations—close to 1 is good, and the standard error of the slope will reflect how well the quantiles of the predictions match the quantiles of the observations—small is good. In this way, several interpretable characteristics of model performance can be distilled from the omnibus test. Robinson (2019) provides a more detailed explanation with examples, and Robinson (2016) provides an R (R Core Team 2017) package that runs such tests[1].

### 3.8. Choosing the Limits of Economic Significance

The choice of the limits of economic significance is the most critical step for interval-based tests. Detailed discussions in the contexts of psychology and medical research appear in Murphy and Myors (1999), Walker and Nowacki (2011) and Lakens et al. (2018), among others. These limits affect the outcomes of the test, and also provide scientific credibility to the research outcome. The limits should be determined by the researcher, in consideration of economic theories and meaningful effect size. In so doing, economic reasoning or theory can be incorporated into statistical decision-making.

As Murphy and Myors (1999, p. 237) point out, the choice of limits requires "value judgment". The choice can also be "context-dependent", since it may depend on the type of dependent variable involved; and can also depend on the likelihood or seriousness of Type I and II errors. It would be desirable to have a set convention or a consensus of expert opinions in the related field as to the extent of "negligible effects" that could be economically ignored. One may also use meta-analytic evidence from past studies.

The researcher can be guided by estimation-based measures to further justify their choice. For example, one may choose the limits so that they imply the smallest effect size guided by the value of Cohen's $d$ (Cohen 1977), which is a measure of effect size (the mean difference divided by the standard deviation of the data). In the regression context, the limits may be determined so that the implied economic impact provides a certain value of (incremental) signal-to-noise ratio $\eta$ given in (6) (which is also called Cohen's $f^2$) or desired coefficient of determination $R^2_{pj}$. For example. if $Y$ is stock return and $X$ is a proposed factor, the interval can be formulated so that $X$ can explain at least 5% of the total variation of stock return ($R^2_{p1} = 0.05$; and $R^2_{p0} = 0$). This is based on the judgment that an economically meaningful factor should explain at least 5% of the stock return variation, in the absence of other factors. Again, this choice requires value judgment that can be context-dependent. For example, the choice may be different across markets depending on the market conditions such as the trading cost, regulatory framework, and development of market structure, among others. The researcher may consider a number of different values or possible candidates of this value, and make a decision considering the inferential outcomes and their economic significance. However, most ideally, the choice of the limits should be made before the researcher observes the data.

The proposed interval can be indicative of the decision when the point estimate is available. However, the point estimate is subject to sampling variability and it is necessary to conduct the test to make a more informed decision under sampling variability. Proposing such an interval may be equivalent to providing a prior distribution for the Bayesian inference. It is well known that the outcome of the Bayesian inference in large part depends on the choice of prior. But if the choice is made based on concrete economic reasoning and evidence, the Bayesian inference can provide

[1] There are two other R packages for equivalence and non-inferiority tests. One is EQUIVNONINF (Wellek and Ziegler 2017) which accompanies the book by Wellek (2010), and the other is PowerTOST (Labes et al. 2018), which contains functions to calculate power and sample size for various study designs used for bio-equivalence studies.

an informed decision. Similarly, if the interval of economic significance is proposed with concrete economic rationale, then it can help the researcher make a correct decision.

Furthermore, it is important for the researcher to include the key components of the test in reporting, such as the interval of equivalence. Doing so serves two purposes: first, it enables the reader to apply different intervals for different applications, and second, it provides a check against unscrupulous researchers choosing intervals that suit their narrative.

## 4. Empirical Applications

In this section, we provide empirical applications of the interval-based tests discussed in Section 3 to economics and finance. We present two cases where large sample size is used; and one case of a small sample.

### *4.1. A SAD Stock Market Cycle*

In empirical finance, a large number of market anomalies have been identified, where it is claimed that a stock market is systematically influenced by the factors unrelated with the market fundamentals. The evidence is at odds with the efficient market hypothesis which is a cornerstone of modern finance theories. Central to this is the findings that investors' mood systematically and negatively affects stock return. For example, it is hypothesized that less sunlight or more cloudiness negatively affect investors' mood, which in turn exerts a negative impact on stock market return. The seminal papers in this area of literature include Saunders (1993), Hirshleifer and Shumway (2003), and Kamstra et al. (2003). However, as Kim (2017) reports, the studies in this area typically show negligible effects with high statistical significance, accompanied by large sample size and negligible $R^2$ values.

Kamstra et al. (2003) study the effect of depression linked with seasonal affective disorder (SAD) on stock return. They claim that, through the link between SAD and depression, and the link between depression and risk aversion, seasonal variation in length of day can translate into seasonal variation in equity return. They consider the regression model of the following form:

$$R_t = \gamma_0 + \sum_{i=1}^{2} \gamma_i R_{t-i} + \gamma_3 M_t + \gamma_4 T_t + \gamma_5 SAD_t + \gamma_6 F_t + \gamma_7 C_t + \gamma_8 P_t + \gamma_9 G_t + \epsilon_t, \tag{9}$$

where $R_t$ denotes the stock return in percentage on day $t$; $M$ a dummy variable for Monday; $T$ a dummy for the last trading day or the first five trading days of the tax year; $F$ a dummy for fall; $C$ cloud cover, $P$ a precipitation; and $G$ temperature. $SAD_t$ is a measure of seasonal depression, which takes the value of $H_t - 12$ where $H_t$ represents the time from sunset to sunrise if the day $t$ is in the fall or winter; 0 otherwise.

Kamstra et al. (2003; p. 326) argue that lower returns should commence with autumn because depressed investors shunning risk and re-balance their portfolio in favor of safer assets (i.e., $\gamma_6 < 0$). This is followed by abnormally higher returns when days begin to lengthen and SAD-affected investors begin resuming their risky holdings (i.e., $\gamma_5 > 0$). They use the daily index return data from the markets around the world: U.S. (S&P 500, NYSE, NASGAQ, AMEX), Sweden, U.K., Germany Canada, New Zealand, Japan, Australia, and South Africa. They report, nearly for all markets, that the parameter estimate of $\gamma_5$ is positive and statistically significant at a conventional level of significance; and that of $\gamma_6$ is negative and statistically significant. These results are the basis of their evidence for the existence of the SAD effects around the world. However, the results are based on the point null hypothesis at a conventional level of significance under large sample sizes, for which Rao and Lovric (2016) among others are concerned about. In this section, we evaluate the regression results of Kamstra et al. (2003) using the interval-based tests.

### 4.1.1. Evaluating the Results of Kamstra et al.

We first conduct the interval tests using the regression results reported in Kamstra et al. (2003). Table 1 reports the sample size ($T$) and $R^2$ values of the regression (9), reproduced from Kamstra et al. (2003; Tables 2 and 4A–C). From these values, we calculate the $F$-statistic for joint significance of all slope coefficients are jointly zero ($H_0 : \gamma_1 = \cdots = \gamma_9 = 0$), as reported in Table 1. The $CR$ column reports the 5% critical values from the central $F$ distributions, which are around 1.88 regardless of sample size. It appears that the $F$-test for joint significance is clearly rejected for all markets at a conventional significance level, which indicates that the all slope coefficients of regression (9) are statistically significant. However, this is at odds with negligible $R^2$ values reported in Table 1 which indicate little predictive power for all markets.

Suppose that, for a regression model for stock return to be economically significant, it should explain at least 5% of the return variation. That is, we test for $H_0 : 0 \leq R^2_{p1} \leq 0.05$ against $H_1 : R^2_{p1} > 0.05$. The column labeled $CR_2$ reports the 5% critical values associated with $F(J, T-K-1; \lambda_{max})$ while the value of $\lambda_{max}$ is associated with $R^2_{p1} = 0.05$ (and $R^2_{p0} = 0$). According to these critical values, the null hypothesis of economically negligible effect cannot be rejected for all market indices except for US4. The critical values listed in column $CR_1$ are those associated with $H_0 : 0 \leq R^2_{p1} \leq 0.01$, which delivers rejection in four markets only. If we test for $H_0 : 0 \leq R^2_{p1} \leq 0.1$, the critical values in column labeled $CR_3$ indicate that the predictive power of the estimated models are economically negligible for all markets.

**Table 1.** Testing for the SAD effect.

| Market | $T$ | $R^2$ | $F$ | $CR$ | $CR_1$ | $CR_2$ | $CR_3$ |
|---|---|---|---|---|---|---|---|
| US1 | 18,380 | 0.011 | 20.43 | 1.83 | 26.87 | 120.25 | 245.15 |
| US2 | 9688 | 0.027 | 29.84 | 1.88 | 15.76 | 66.27 | 133.20 |
| US3 | 7083 | 0.033 | 26.82 | 1.88 | 12.33 | 49.83 | 99.26 |
| US4 | 9688 | 0.091 | 107.66 | 1.88 | 15.77 | 66.27 | 133.20 |
| SWE | 4836 | 0.017 | 9.28 | 1.88 | 9.29 | 35.46 | 69.70 |
| UK | 4534 | 0.009 | 4.57 | 1.88 | 8.87 | 33.51 | 65.69 |
| GER | 9411 | 0.008 | 8.42 | 1.88 | 15.40 | 64.53 | 129.61 |
| CAN | 8308 | 0.030 | 28.52 | 1.88 | 13.96 | 57.58 | 115.26 |
| NZ | 2627 | 0.010 | 3.31 | 1.94 | 6.79 | 23.49 | 45.04 |
| JAP | 12,783 | 0.002 | 3.20 | 1.94 | 22.11 | 96.18 | 194.77 |
| AUS | 5521 | 0.010 | 6.19 | 1.88 | 10.22 | 39.86 | 78.75 |
| SA | 7247 | 0.010 | 8.12 | 1.88 | 12.54 | 50.87 | 101.41 |

US1: United States, S&P500, from 04 January 1928 to 29 December 2000; US2: United States, NYSE, from 1962-07-05 to 2000-12-29; US3: United States, NASDAQ, from 1972-12-18 to 2000-12-29; SWE: Sweden from 1982-09-15 to 2001-12-18; UK: Britain from 1984-01-04 to 2001-12-06; GER: Germany from 1965-01-05 to 2001-12-12; CAN: Canada from 1969-01-03 to 2001-12-18; NZ: New Zealand from 1991-07-02 to 2001-12-18; JAP: Japan from 1950-04-05 to 2001-12-06; AUS: Australia from 1980-01-03 to 2001-12-18; SA: South Africa from 1973-01-03 to 2001-12-06; $T$: sample size, calculated using R package "bizdays" (Freitas 2018) from the sample ranges reported in Kamstra et al. (2003; Table 2); $R^2$: $R^2$ values reported in Kamstra et al. (2003; Table 4A–C); $F$: $F$-statistic for the joint significance of regression slope coefficients; $CR$: 5% critical values from a central $F$ distribution for $H_0 : R^2 = 0$; $CR_1$: 5% critical values for $H_0 : 0 \leq R^2 \leq 0.01$; $CR_2$: 5% critical values for $H_0 : 0 \leq R^2 \leq 0.05$; $CR_3$: 5% critical values for $H_0 : 0 \leq R^2 \leq 0.10$.

Economic significance of the magnitude of regression coefficients reported in Kamstra et al. (2003) is also questionable. For example, for the U.S. market with S&P500 index (US1), $\hat{\gamma}_6 = -0.058$ and its 90% confidence interval is $[-0.10, -0.01]$. The point estimate means that the stock return is on average lower by 0.058% during the autumn period. Suppose, for a factor to have an economically meaningful impact on stock return, its marginal effect should be at least 0.5% (either positive or negative) to justify transaction cost. Then, one can formulate the null hypothesis of economically negligible effect as $H_0 : -0.5 \leq \gamma_6 \leq 0.5$. The 90% confidence interval is clearly within this bound, so we do not reject $H_0$ at the 5% level of significance. The same inferential outcomes apply to all the other regression coefficients of (9) reported in Kamstra et al. (2003). Note that, depending on the attitude of the researcher, one can formulate the null hypothesis as $H_0 : (\gamma_6 < -0.5) \cup (\gamma_6 > 0.5)$, but it is also clearly rejected at the 5% level in favor of a negligible effect. Although Kamstra et al. (2003) justify their effect size using the annualized return, this annualized return does not take account of the underlying volatility of stock return or trading costs involved.

#### 4.1.2. Replicating the Results of Kamstra et al.

We now replicate the model (9) using the value-weighted daily returns from the NYSE composite index (CRSP). The SAD variable and other dummy variables are generated following Kamstra et al. (2003), using programming language R (R Core Team 2017). The data for weather variables ($C$, $P$, and $G$) are collected from the National Center for Environmental Information.[2] Our data for the regression ranges from January 1965 to April 1996 (7886 observations), due to the limited availability of the weather data ($C$) for New York. We have the following estimated values for the key coefficients: $\hat{\gamma}_5 = 0.032$ with $t$-statistic of 2.29; $\hat{\gamma}_6 = -0.055$ with $t$-statistic of $-2.17$; and $R^2 = 0.05$. These values are fairly close to those reported in Table 4A of Kamstra et al. (2003).

We first pay attention to the point null hypothesis that $H_0 : \gamma_5 = \gamma_6 = 0$ for joint significance of the SAD effects. The $F$-statistic is 3.18 with the $p$-value of 0.04, rejecting $H_0$ at the 5% significance level. This is despite the observation that the incremental contribution of these two variables is negligible, measured by $R_1^2 - R_0^2 = 0.0008$ with $R_1^2 = 0.0501$ and $R_0^2 = 0.0493$. Next, we consider an interval hypothesis of minimum-effect. Suppose that the incremental contribution of these variables should be at least 0.01 to be economically significant. That is,

$$H_0 : (R_{p1}^2 - R_{p0}^2) \leq 0.01.$$

Assuming $R_{p0}^2 = 0.05$, $\lambda_{max} = 83.87$ and the corresponding 5% critical value is 58.97, obtained from $F(J, T-K-1; \lambda_{max})$. With this critical value being much larger than the $F$-statistic of 3.18, the above interval null hypothesis of minimum-effect cannot be rejected at the 5% level, providing evidence that the SAD economic cycle is economically negligible in the U.S. stock market.

### *4.2. Empirical Validity of an Asset-Pricing Model*

An asset-pricing model explains the variation of asset return as a function of a range of risk factors. The most fundamental is the capital asset pricing model (CAPM) which stipulates that an asset (excess) return is a linear function of market (excess) return. The slope coefficient (often called beta) measures the sensitivity of an asset return to the market risk. While the CAPM is theoretically motivated, the market risk alone cannot fully explain the variation of asset return. In response to this, several multi-factor models have been proposed, which augment the CAPM with a number of empirically motivated risk factors such as the size premium or value premium (see, for example, Fama and French 1993). The most recently proposed multi-factor model is the five-factor model of Fama and French (2015), which can be written as

$$R_{it} - R_{ft} = a_i + b_i(R_{Mt} - R_{ft}) + s_i SMB_t + h_i HML_t + r_i RMW_t + c_i CMA_t + e_{it}, \tag{10}$$

where $R_{it}$ is the return on an asset or portfolio $i$ at time $t$ $(i = 1, \ldots, N; t = 1, \ldots, T)$, $R_{ft}$ is the risk-free rate, $R_{Mt}$ is the return on a (value-weighted) market portfolio at time $t$, $SMB_t$ is the return on a diversified portfolio of small stocks minus the return on a diversified portfolio of big stocks, the $HML_t$ is the spread in returns between diversified portfolios of high book-to-market stocks and low book-to-market stocks, $RMW_t$ is the spread in returns between diversified portfolios of stocks with robust and weak profitability, and the $CMA_t$ is the spread in returns between diversified portfolios

[2] https://www.ncdc.noaa.gov/data-access.

of low and high investment firms. The precursors to this 5-factor model include the 3-factor model of Fama and French (1993) which include $(R_{Mt} - R_{ft})$, $SMB$, and $HML$; and the 4-factor model of Carhart (1997) which adds momentum factor ($MOM$) to the 3-factor model. If these factors fully or adequately capture the variation of asset return, then the intercept terms $a_i$ (which may be may be interpreted as the risk-adjusted return) should be zero or sufficiently close to it. On this basis, the model's empirical validity is evaluated by testing for $H_0 : a_1 = ... = a_N = 0$, which is a point-null hypothesis.

### 4.2.1. GRS Test: Minimum-Effect

The $F$-test for $H_0$ is widely called the GRS test, proposed by Gibbons et al. (1989). Let $a = (a_1, \ldots, a_N)'$ be the vector of $N$ intercept terms, and $\Sigma$ be the $N \times N$ covariance matrix of error terms. The model (10) is estimated using the ordinary least-squares: $\hat{a}$ denotes the estimator for $a$ and $\widehat{\Sigma}$ the estimator for $\Sigma$. The $F$-test statistic is written as

$$F = \frac{T(T-N-K)}{N(T-K-1)} \frac{\hat{a}'\widehat{\Sigma}^{-1}\hat{a}}{1+\hat{\mu}'\widehat{\Omega}^{-1}\hat{\mu}}, \tag{11}$$

where $T$ is the sample size, $K = 5$ is the number of risk factors, $\widehat{\Omega}$ is the $K \times K$ covariance matrix of risk factors, and $\hat{\mu}$ is the $K \times 1$ mean vector. Under the assumption that the error terms $e$'s follow a multivariate normal distribution, the statistic follows the $F(N, T-N-K; \lambda)$ distribution, with the non-centrality parameter

$$\lambda = \left(\frac{T}{1+\hat{\theta}^2}\right) a'\Sigma^{-1}a = \left(\frac{T}{1+\hat{\theta}^2}\right)(\theta^{*2} - \theta^2), \tag{12}$$

where $\hat{\theta}$ is the *ex-post* maximum Sharpe ratio of $K$-factor portfolio, $\theta$ is the *ex-ante* maximum Sharpe ratio of $K$-factor portfolio, and $\theta^*$ is the slope of the *ex ante* efficient frontier based on all assets. Gibbons et al. (1989) call $\theta/\theta^*$ the proportion of the potential efficiency. Note that, under $H_0$, this ratio is equal to one and $\lambda = 0$.

However, perfect efficiency cannot exist in practice. It is unrealistic that all of $a$ values are jointly and exactly zero. On this point, it is sensible to consider an interval-based hypothesis testing. For example, consider $H_0 : 0.75 < \theta/\theta^* \leq 1$ against $H_1 : \theta/\theta^* < 0.75$. This is on the basis of judgment that the factors with the proportion of potential efficiency of 0.75 or higher provide practically efficient asset-pricing.

The data is available from French's data library monthly from 1963 to 2015 ($T = 630$).[3] We use 25 portfolio returns ($N = 25$) sorted by size and book-to-market ratio extensively analyzed by Fama and French (1993, 2015). Table 2 reports the test results. The GRS test for $H_0 : a_1 = \ldots = a_N = 0$ are clearly rejected for all models considered, with the $p$-value (not reported) practically 0 for all cases. The critical values of this test (from the central $F$ distributions) is listed in the column labeled $CR$. This results suggest that none of the asset pricing models are able to fully capture asset return variations. This is at odds with the high values of $R^2$ and small values of $|a|$, especially multi-factor models. For the 4-factor and 5-factor model, the estimated ratio of potential efficiency is much higher than other models, close to 0.7.

[3] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html.

**Table 2.** GRS test for asset-pricing models.

| Model | GRS | $R^2$ | $\|a\|$ | $CR$ | $CR_1$ | Ratio |
|---|---|---|---|---|---|---|
| CAPM | 4.41 | 0.74 | 0.25 | 1.52 | 1.88 | 0.25 |
| 3-factor | 3.61 | 0.92 | 0.10 | 1.52 | 2.62 | 0.46 |
| 4-factor | 3.07 | 0.92 | 0.09 | 1.52 | 3.70 | 0.63 |
| 5-factor | 2.79 | 0.92 | 0.09 | 1.52 | 4.03 | 0.67 |

CAPM: the model with single factor $(R_{Mt} - R_{ft})$; 3-factor: CAPM plus $SMB$ and $HML$ (Fama and French 1993); 4-factor: 3-factor plus $MOM$ (Carhart 1997); 5-factor: 3 factor plus $RMW$ and $CMA$ (Fama and French 1993); GRS: GRS test statistic $H_0 : a_1 = \ldots = a_N = 0$; $R^2$: average $R^2$ values over $N = 25$ equations; $|a|$: average intercept estimates over $N = 25$ equations; CR: 5% critical value from $F(N, T-N-K)$; $CR_1$: 5% critical value from $F(N, T-N-K, \lambda_{max})$; ratio: sample estimate of $\theta/\theta^*$.

Table 2 also reports the critical values ($CR_2$) for $H_0 : 0.75 < \theta/\theta^* \leq 1$, which is calculated from $F(N, T-N-K, \lambda_{max})$ distribution with the value of $\lambda_{max}$ implied by $\theta/\theta^* = 0.75$. It is found that, for the 4-factor and 5-factor models, $H_0 : 0.75 < \theta/\theta^* \leq 1$ cannot be rejected at the 5% level of significance. This suggests that these multi-factor model have captured the variation of asset returns adequately, with economically negligible deviation from the perfect efficiency. For the CAPM and 3-factor models, the interval-based $H_0$ is rejected at the 5% level, but this seems consistent with the estimated values of potential efficiency which are less than 0.5 for both cases. It is worth noting that the critical values $CR$ for the point-null hypothesis (based on the central $F$-distribution) are nearly identical for all cases, regardless of the estimation results such as $R^2$ and $|a|$. However, those for the interval-based tests are different, depending on the model estimation results.

#### 4.2.2. LR Test: Model Equivalence

We now test for the validity of the asset-pricing models using the model equivalence test discussed in Section 3.6. We calculate the LR test for given in (8) for $H_0 : a_1 = \ldots = a_N = 0$, which is written as

$$LR = T(\log\{\det[\hat{\Sigma}(H_0)]\} - \log\{\det[\hat{\Sigma}(H_1)]\}),$$

where $\hat{\Sigma}(H_i)$ denotes the maximum likelihood estimator for $\Sigma$ under $H_i$. For the model equivalence test given in (7), the above LR statistic follows the $\chi^2_{N,\delta^2}$ distribution with $\delta^2 = T\Delta^2$. Using the same data set as in Section 4.2.1, the LR statistic is 105.67, 88.05, 75.82, and 69.26 for the CAPM, 3-factor model, 4-factor model, and 5-factor model respectively. If we set $\Delta^2$ to 0.1, the 5% critical value is 61.12, indicating that $H_0$ is not approximately valid for all models. If we set $\Delta^2$ to 0.15, the 5% critical value is 87.19, indicating $H_0$ is approximately valid only for 4-factor and 5-factor models. If we set $\Delta^2$ to 0.20, the 5% critical value is 114.00, indicating that $H_0$ is approximately valid for all models. It appears that the results are sensitive to the choice of $\Delta^2$ values. However, at a reasonable value of $\Delta^2 = 0.15$, the results are consistent with the minimum-effect test based on the GRS test conducted above.

### *4.3. Testing for Persistence of a Time Series*

The presence of a unit root in economic and financial time series has strong implications to many economic theories and their empirical validity (see Choi 2015). For example, a unit root in the real exchange rate is evidence that the purchasing power parity does not hold (Lothian and Taylor 1996); and a unit root in the real GNP supports the view that a shock to the economy has a permanent effect, which is not consistent with the traditional (or Keynesian) view of business cycle (Campbell and Mankiw 1987). To test for the hypothesis, the unit root test proposed by Dickey and Fuller (1979) has been widely used, while a large number of its extensions and improvement have been proposed. The augmented Dickey–Fuller (ADF) test for a time series $Y$ is based on the regression of the form

$$\Delta Y_t = \delta_0 + \delta_1 t + \theta Y_{t-1} + \sum_{j=1}^{m-1} \rho_j \Delta Y_{t-j} + u_t, \tag{13}$$

where $\Delta Y_t = Y_t - Y_{t-1}$; $m$ is the autoregressive (AR) order of $Y$; and $u_t$ is an *i.i.d*. error term with zero mean and fixed variance. Note that $\theta \equiv \tau - 1$ where $\tau$ is the sum of all AR(m) coefficients in level of $Y$, measuring the degree of persistence. The test for a unit root is based on point-null hypothesis of $H_0 : \theta = 0$ against $H_1 : \theta < 0$. Under $H_0$, the $t$-test statistic asymptotically follows the Dickey–Fuller distribution, from which the critical values of the test are obtained. Under $H_1$, the $t$-test statistic asymptotically follows the standard normal distribution.

The problems of the unit root test are well documented (see, for example, Choi 2015). The most well-known is its low power (at a conventional significance level), which means that there is a high chance of committing Type II error (failure to reject a false null hypothesis). On this point, Kim and Choi (2017) propose the unit root test at the optimal level of significance, which is obtained by minimizing the expected loss from hypothesis testing. They find that the optimal level is in the 0.3 and 0.4 range for many economic time series, arguing that the exclusive use of 0.05 level has led to accumulation of false stylized facts. The other problem of the test is the discontinuity of the sampling distributions of the test statistic under $H_0$ and $H_1$. This makes the decision highly sensitive to the value specified under $H_0$.

More importantly, as discussed in Section 2.3, it is unrealistic to assume that an economic time series such as the real GNP or real exchange rate has an autoregressive root exactly equal to one. An economist may wish to test whether a time series shows a degree of persistence practically different from that of a unit root time series. The test can be conducted in the context of non-inferiority test discussed in the previous section. To do this, we need to find the value of $\tau$ or $\theta$ under which a time series shows a practically different degree of persistence from a unit root time series. According to DeJong et al. (1992), a plausible value of $\tau$ under $H_1 : \theta < 0$ is 0.85, 0.95, 0.99 for annual, quarterly and monthly data respectively, which translate to the $\theta$ values of $-0.15$, $-0.05$, and $-0.01$. On this basis, we test for the persistence of a time series using the following interval hypotheses:

$$H_0 : \theta \leq \theta_1; H_1 : \theta > \theta_1,$$

where $\theta_1 \in \{-0.15, -0.05, -0.01\}$ depending on the data frequency. The time series is practically trend-stationary under this $H_0$. This test is a standard one-sample $t$-test whose statistic asymptotically follows the standard normal distribution. However, we note that the least-squares estimator for $\tau$ or $\theta$ is biased in small samples, which may adversely affect the small sample properties of the test. As an alternative to the non-inferiority test, we also use the bias-corrected bootstrap confidence interval for $\theta$ for improved statistical inference, similar to those of Kilian (1998a, 1998b) and Kim (2004).

For a set of time series $(Y_1, \ldots . Y_T)$, we first estimate the parameters of model (13) using the bias-corrected estimators. Let $(\hat{\delta}_0, \hat{\delta}_1, \hat{\theta}, \hat{\rho}_1, \ldots, \hat{\rho}_{m-1})$ be the bias-corrected estimators; and let $\{e_t\}$ denote the corresponding residual. Generate the artificial data set as

$$Y_t^* = \hat{\delta}_0 + \hat{\delta}_1 t + \hat{\beta}_1 Y_{t-1} + \cdots + \hat{\beta}_m Y_{t-m} + e_t^*,$$

using $(Y_1, \ldots, Y_m)$ as the starting values, where $e_t^*$ is a random draw with replacement from $\{e_t\}_{t=m+1}^{T}$ and $(\hat{\beta}_1, \ldots, \hat{\beta}_m)$ are the AR coefficients in level associated with $(\hat{\theta}, \hat{\rho}_1, \ldots, \hat{\rho}_{m-1})$. Using $\{Y_t^*\}_{t=1}^{T}$, estimate the $AR(m)$ coefficients, again with bias correction, $(\hat{\delta}_0^*, \hat{\delta}_1^*, \hat{\beta}_1^*, \ldots, \hat{\beta}_m^*)$. For bias correction, we use Shaman and Stine (1988) asymptotic formula with stationarity-correction, following Kilian (1998b) and Kim (2004). We obtain $\hat{\theta}^* = \hat{\tau}^* - 1$, where $\hat{\tau}^* = \sum_{j=1}^{m} \hat{\beta}_j^*$. Repeat this process $B$ times to obtain the bootstrap distribution $\{\hat{\theta}^*(j)\}_{j=1}^{B}$, which can be used as an approximation to the sampling distribution of $\hat{\theta}$. If the confidence interval for $\theta$ obtained from $\{\hat{\theta}^*(j)\}_{j=1}^{B}$ covers $\theta_1$, then this is evidence that the time series shows a degree of of persistence practically no different from that of a trend-stationary time series.

Table 3 reports the results from the extended Nelson and Plosser (1982) data for a set of annual U.S. macroeconomic time series, setting $\theta_1 = -0.15$. Firstly, the ADF test (a point-null hypothesis test) provides the $p$-values larger than 0.05 for most of time series, providing evidence that many macroeconomic time series have a unit root. In contrast, the $t$-test (non-inferiority test) results for $H_0 : \theta \leq -0.15$ against $H_1 : \theta > -0.15$ show that we clearly cannot reject this $H_0$ at the 5% level of significance (asymptotic critical value 1.645) for the real GNP, real per capita GNP, industrial production, employment, unemployment rate, providing evidence that these time series are practically trend-stationary. As for the bootstrap inference, it is found that the 95% bias-corrected bootstrap confidence interval for $\theta$ does cover $-0.15$, for the real GNP, real per capita GNP, industrial production, employment, unemployment rate, real wage, and interest rate, indicating that these time series show the degree of persistence practically of a trend-stationary time series. The two alternative methods are in agreement in their inferential outcomes, except for real wage and interest rate.

**Table 3.** Test for persistence: Extended Nelson–Plosser Data.

| | $T$ | $p$-**Value** | $\hat{\theta}$ | $t$-Stat | $CI_1$ | $CI_2$ |
|---|---|---|---|---|---|---|
| R.GNP | 80 | 0.05 | −0.140 | −0.66 | −0.298 | −0.037 |
| N.GNP | 80 | 0.58 | −0.010 | 2.96 | −0.147 | −0.0001 |
| P.GNP | 80 | 0.04 | −0.150 | −0.82 | −0.304 | −0.043 |
| IP | 129 | 0.26 | −0.098 | −0.21 | −0.274 | −0.002 |
| Emp | 99 | 0.18 | −0.118 | −0.21 | −0.242 | −0.031 |
| Uemp | 99 | 0.01 | −0.214 | −1.53 | −0.430 | −0.074 |
| Def | 100 | 0.70 | −0.003 | 5.56 | −0.081 | −0.0001 |
| CPI | 129 | 0.91 | −0.002 | 13.22 | −0.019 | −0.0001 |
| Wages | 89 | 0.53 | −0.026 | 2.81 | −0.144 | −0.0002 |
| Rwages | 89 | 0.75 | −0.010 | 1.90 | −0.183 | −0.0002 |
| MS | 100 | 0.18 | −0.037 | 3.83 | −0.110 | −0.0016 |
| Vel | 120 | 0.78 | −0.001 | 4.65 | −0.099 | −0.0001 |
| Rate | 89 | 0.98 | −0.025 | 2.35 | −0.191 | −0.0004 |
| S&P | 118 | 0.64 | −0.021 | 2.15 | −0.144 | −0.0002 |

R.GNP: Real GNP; N.GNP: Nominal GNP: P.GNP: Real per capita GNP; IP” Industrial Production; Emp: Employment; Uemp: Unemployment Rate; Def: GNP deflator; CPI: Consumer Price Index; Wages: Wages; Rwages: Real Wages: MS: Money Stock; Vel: Velocity; Rate: Interest rate; S&P: Common Stock Price. $T$: Sample size; $p$-value: $p$-value of the ADF test for $H_0 : \theta = 0$; $\hat{\theta}$: bias-corrected estimators for $\theta$; $t$-stat: $t$-statistic for $H_0 : \theta \leq -0.15$ against $H_1 : \theta > -0.15$ based on equation (13) with 5% critical value of 1.645; $(CI_1, CI_2)$: lower and upper bounds of 95% bootstrap bias-corrected confidence interval for $\theta$; The AR orders used are same as those of Nelson and Plosser (1982).

The results for the test of persistence based on the non-inferiority test are largely consistent with those of Kim and Choi (2017) who re-evaluate the ADF test results at the optimal level of significance and report evidence that the real GNP, real per capita GNP, employment, and money stock do not have a unit root. These results are also largely consistent with the Bayesian evidence of Schotman and van Dijk (1991).

## 5. Conclusions

This paper provides a review of interval-based hypothesis testing methods, which are known under the name of minimum-effect, non-inferiority, and equivalence tests in biostatistics and psychology. Although the first proposal of such a test goes back to Hodges and Lehmann (1954), it has attracted little attention in the business disciplines of science. In the latter, the paradigm of point-null hypothesis has been the major workforce in making statistical decisions and establishing research findings. However, as a number of authors have criticized for many years, the current paradigm has a range of limitations and deficiencies, as discussed in Section 2 of this paper. These problems have become even more apparent in the big data era, where the $p$-value criterion widely and routinely adopted by statistical researchers is no longer usable in making sensible statistical decisions. The consequences are serious, with widespread practice of data-mining (Black 1993), data-snooping (Lo and MacKinlay 1990), $p$-hacking (Harvey 2017), and multiple testing (Harvey et al. 2016), which result in an embarrassing number of false positives as Harvey (2017) puts it. The related empirical evidence is provided by meta-analytic studies conducted by Kim and Ji (2015) and Kim et al. (2018).

Even more serious is systematic distortion of published results, such as publication bias (Basu and Park 2014) and replication crisis (Peng 2015). In light of these problems, Rao and Lovric (2016) call for a new paradigm to be in place for statistical testing in the 21st century, with a proposal of interval-based hypothesis testing as a possible solution.

An important point in favor of adopting an interval-based test is the fact that an economic hypothesis cannot be formulated as a point. Rather, it is more sensible when it takes a form of an interval or a neighborhood: see, for example, De Long and Lang (1992), Leamer (1988), and Startz (2014). For example, when a researcher tests for stock market efficiency, she is not testing for a perfect efficiency (as described by a point-null hypothesis), since such a perfect relationship cannot hold economically (Grossman and Stiglitz 1980). More realistically, the researcher is interested in whether the degree of market inefficiency (Campbell et al. 1997) is economically large enough to be concerned (an interval hypothesis). Hence, it makes more sense to consider an interval hypothesis for decision-making in economic or business research.

As we have seen in this paper, an interval-hypothesis can be implemented in a simple and straightforward manner, using the existing instruments of hypothesis testing such as one-tailed test, confidence interval, and non-central distributions. Its main attraction is that the critical values of these tests increase with sample size, overcoming a major deficiency of point-null hypothesis testing. A key requirement of the test is that the researcher should specify an interval of economic significance under the null or alternative hypothesis, preferably before she observes the data. This may require a value judgment depending on contexts, accompanied by a thorough economic analysis on the effect size of the relationship under investigation. This is an integral part of interval-based hypothesis testing, since it has a strong impact on the test outcome and research integrity. It is also highly desirable that the relevant research community establishes a consensus on the range of minimum effect size that matters economically.

We have applied the interval-based tests to economics and finance applications. The first is a test for market efficiency, whether investors' mood has a systematic effect on stock market return. While the effect may appear to show statistical significance under the current point-null paradigm, the minimum-effect tests cannot reject its negligible economic effect. The second is on the empirical validity of asset-pricing models. In contrast to the findings based on point-null hypothesis testing, we find that a class of multi-factor models are empirically valid based on minimum-effect and model equivalence tests. The third is on the degree of persistence of economic time series. A unit root test based on a conventional point-null hypothesis strongly favors the presence of a unit root in many macroeconomic time series such as the real GNP. According to the non-inferiority test, many time series in Nelson–Plosser data set are found to show a degree of persistence of a trend-stationary time series, especially in the real income variables. From these applications, we find that the interval-based tests are applicable to many contentious research problems in the business disciplines of science, shedding new lights on the existing results or stylized facts. We propose that interval-based hypothesis tests be widely adopted in business research, especially in the new era of big data.

**Author Contributions:** J.H.K. conceptualized, developed theory, analyzed data, wrote, and reviewed manuscript. A.P.R. conceptualized, wrote, and reviewed manuscript.

**Funding:** This research received no external funding.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

Arrow, Kenneth. 1960. Decision theory and the choice of a level of significance for the *t*-test. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Edited by Ingram Olkin. Stanford: Stanford University Press, pp. 70–78.

Basu, Sudipta, and Han-Up Park. 2014. Publication Bias in Recent Empirical Accounting Research, Working Paper. Available online: http://ssrn.com/abstract=2379889 (accessed on 31 May 2018).

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: 112–22. [CrossRef]

Black, Fischer. 1993. Beta and return. *The Journal of Portfolio Management* 20: 8–18.

Box, George E. P. 1976. Science and Statistics. *Journal of the American Statistical Association* 71: 791–99. [CrossRef]

Campbell, John Y., and N. Gregory Mankiw. 1987. Are Output Fluctuations Transitory? *Quarterly Journal of Economics* 102: 857–80. [CrossRef]

Campbell, John Y., Andrew W. Lo, and Archie Craig MacKinlay. 1997. *The Econometrics of Financial Markets*. Princeton: Princeton University Press.

Carhart, Mark M. 1997. On persistence in mutual fund performance. *Journal of Finance* 52: 57–82. [CrossRef]

Choi, In. 2015. *Almost All about Unit Roots*. New York: Cambridge University Press.

Cohen, Jacob. 1977. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. New York: LBA.

Cohen, Kalman J., and Richard M. Cyert. 1961. Computer Models in Dynamic Economics. *The Quarterly Journal of Economics* 75: 112–27. [CrossRef]

De Long, J. Bradford, and Kevin Lang. 1992. Are All Economic Hypotheses False? *Journal of Political Economy* 100: 1257–72. [CrossRef]

DeJong, David N., John C. Nankervis, N. E. Savin, and Charles H. Whiteman. 1992. Integration versus trend stationary in time series. *Econometrica* 60: 423–33. [CrossRef]

Dickey, David A., and Wayne A. Fuller. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74: 427–31.

Efron, Bradley, and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Boca Raton: Chapman & Hall, CRC Monographs on Statistics & Applied Probability.

Fama, Eugene F., and Kenneth R. French. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33: 3–56. [CrossRef]

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

Freitas, Wilson. 2018. Bizdays: Business Days Calculations and Utilities. R Package Version 1.0.6. Available online: https://CRAN.R-project.org/package=bizdays (accessed on 31 May 2018).

Gibbons, Michael R., Stephen A. Ross, and Jay Shanken. 1989. A test of the efficiency of a given portfolio. *Econometrica* 57: 1121–52. [CrossRef]

Gigerenzer, Gerd. 2004. Mindless statistics: Comment on “Size Matters”. *Journal of Socio-Economics* 33: 587–606. [CrossRef]

Grossman, Sanford J., and Joseph E. Stiglitz. 1980. On the impossibility of informationally efficient markets. *The American Economic Review* 70: 393–408.

Harvey, Campbell R. 2017. Presidential Address: The Scientific Outlook in Financial Economics. *Journal of Finance* 72: 1399–440. [CrossRef]

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

Hirshleifer, David, and Tyler Shumway. 2003. Good day sunshine: Stock returns and the weather. *Journal of Finance* 58: 1009–32. [CrossRef]

Hodges, J. L., Jr., and E. L. Lehmann. 1954. Testing the Approximate Validity of Statistical Hypotheses. *Journal of the Royal Statistical Society, Series B (Methodological)* 16: 261–68. [CrossRef]

Johnstone, D. J., and D. V. Lindley. 1995. Bayesian Inference Given Data Significant at Level $\alpha$: Tests of Point Hypotheses. *Theory and Decision* 38: 51–60. [CrossRef]

Kamstra, Mark J., Lisa A. Kramer, and Maurice D. Levi. 2003. Winter blues: A sad stock market cycle. *American Economic Review* 93: 324–43. [CrossRef]

Kandel, Shmuel., and Robert F. Stambaugh. 1996. On the Predictability of Stock Returns: An Asset-Allocation Perspective. *The Journal of Finance* 51: 385–424. [CrossRef]

Keuzenkamp, Hugo A., and Jan Magnus. 1995. On tests and significance in econometrics. *Journal of Econometrics* 67: 103–28. [CrossRef]

Kilian, Lutz. 1998a. Small sample confidence intervals for impulse response functions. *The Review of Economics and Statistics* 80: 218–30. [CrossRef]

Kilian, Lutz. 1998b. Accounting for lag-order uncertainty in autoregressions: The endogenous lag order bootstrap algorithm. *Journal of Time Series Analysis* 19: 531–38. [CrossRef]

Kim, Jae H. 2004. Bootstrap Prediction Intervals for Autoregression using Asymptotically Mean-Unbiased Parameter Estimators. *International Journal of Forecasting* 20: 85–97. [CrossRef]

Kim, Jae H. 2017. Stock Returns and Investors' Mood: Good Day Sunshine or Spurious Correlation? *International Review of Financial Analysis* 52: 94–103. [CrossRef]

Kim, Jae H. 2019. Tackling False Positives in Business Research: A Statistical Toolbox with Applications. *Journal of Economic Surveys* doi:10.1111/joes.12303. [CrossRef]

Kim, Jae H., and In Choi. 2017. Unit Roots in Economic and Financial Time Series: A Re-evaluation at the Decision-based Significance Levels. *Econometrics* 5: 41. [CrossRef]

Kim, Jae H., and In Choi. 2019. Choosing the Level of Significance: A Decision-Theoretic Approach. *Abacus: A Journal of Accounting, Finance and Business Studies*. forthcoming.

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

Kim, Jae. H., Kamran Ahmed, and Philip Inyeob Ji. 2018. Significance Testing in Accounting Research: A Critical Evaluation based on Evidence. *Abacus: A Journal of Accounting, Finance and Business Studies* 54: 524–46. [CrossRef]

Kleijnen, Jack P. C. 1995. Verification and validation of simulation models. *European Journal of Operational Research* 82: 145–62. [CrossRef]

Labes, Detlew, Helmut Schuetz, and Benjamin Lang. 2018. Power and Sample Size Based on Two One-Sided t-Tests (TOST) for (Bio)Equivalence Studies, R Package Version: 1.4-7. Available online: https://cran.r-project.org/web/packages/PowerTOST/index.html (accessed on 31 May 2018).

Lakens, Daniel, Anne M. Scheel, and Peder M. Isager. 2018. Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science* 1: 259–69. [CrossRef]

Lavergne, Pascal. 2014. Model Equivalence Tests in a Parametric Framework. *Journal of Econometrics* 178: 414–25. [CrossRef]

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

Leamer, Edward. 1988. Things that bother me. *Economic Record* 64: 331–35. [CrossRef]

Lehmann, Erich L., and Joseph P. Romano. 2005. *Testing Statistical Hypotheses*, 3rd ed. New York: Springer.

Lo, Andrew W., and A. Craig MacKinlay. 1990. Data Snooping in Tests of Financial Asset Pricing Models. *Review of Financial Studies* 10: 431–67. [CrossRef]

Lothian, James R., and Mark P. Taylor. 1996. Real exchange rate behavior: The recent float from the perspective of the past two centuries. *Journal of Political Economy* 104: 488–510. [CrossRef]

McCloskey, Deirdre N., and Stephen T. Ziliak. 1996. The standard error of regressions. *Journal of Economic Literature* 34: 97–114.

Murphy, Kevin R., and Brett Myors. 1999. Testing the Hypothesis That Treatments Have Negligible Effects: Minimum-Effect Tests in the General Linear Model. *Journal of Applied Psychology* 84: 234–48. [CrossRef]

Murphy, Kevin R., Brett Myors, and Allen Wolach. 2014. *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests*, 4th ed. New York: Routledge.

Nelson, Charles R., and Charles R. Plosser. 1982. Trends and random walks in macroeconomic time series. *Journal of Monetary Economics* 10: 139–62. [CrossRef]

Ohlson, James A. 2018. Researchers' Data Analysis Choices: An Excess of False Positives? Available online: https://ssrn.com/abstract=3089571 (accessed on 31 May 2018).

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349: 6251. [CrossRef]

Peng, Roger. 2015. The Reproducibility Crisis in Science: A Statistical Counterattack. *Significance* 12: 30–32. [CrossRef]

Peracchi, Franco. 2001. *Econometrics*. New York: Wiley.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online: https://www.R-project.org/ (accessed on 31 May 2018).

Rao, Calyampudi Radhakrishna, and Miodrag M. Lovric. 2016. Testing Point Null Hypothesis of a Normal Mean and the Truth: 21st Century Perspective. *Journal of Modern Applied Statistical Methods* 15: 2–21. [CrossRef]

Robinson, Andrew P. 2016. Equivalence: Provides Tests and Graphics for Assessing Tests of Equivalence. R Package Version 0.7.2. Available online: https://cran.r-project.org/web/packages/equivalence/index.html (accessed on 31 May 2018).

Robinson, Andrew P. 2019. Testing Simulation Models Using Frequentist Statistics. In *Computer Simulation Validation: Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives*. Edited by Claus Beisbart and Nicole Saam. Berlin: Springer.

Robinson, Andrew P., and Robert E. Froese. 2004. Model validation using equivalence tests. *Ecological Modelling* 176: 349–58. [CrossRef]

Robinson, Andrew P., Remko A. Duursma, and John D. Marshall. 2005. A regression-based equivalence test for model validation: Shifting the burden of proof. *Tree Physiology* 25: 903–13. [CrossRef] [PubMed]

Saunders, Edward M. 1993. Stock prices and wall street weather. *American Economic Review* 83: 1337–45.

Schotman, Peter C., and Herman K. van Dijk. 1991. On Bayesian routes to unit roots. *Journal of Applied Econometrics* 6: 387–401. [CrossRef]

Shaman, Paul, and Robert A. Stine. 1988. The bias of autoregressive coefficient estimators. *Journal of the American Statistical Association* 83: 842–48. [CrossRef]

Spanos, Aris. 2017. Mis-specification testing in retrospect. *Journal of Economic Surveys* 32: 541–77. [CrossRef]

Startz, Richard. 2014. Choosing the More Likely Hypothesis. *Foundations and Trends in Econometrics* 7: 119–89. [CrossRef]

van der Laan, Mark, Jiann-Ping Hsu, Karl E. Peace, and Sherri Rose. 2010. Statistics ready for a revolution: Next generation of statisticians must build tools for massive data sets. *Amstat News* 399: 38–39.

Walker, Esteban, and Amy S. Nowacki. 2011. Understanding Equivalence and Noninferiority Testing. *Journal of General Internal Medicine* 26: 192–96. [CrossRef]

Wasserstein, Ronald L., and Nicole A. Lazar. 2016. The ASA's statement on p-values: Context, process, and purpose. *The American Statistician* 70: 129–33. [CrossRef]

Wellek, Stefan. 2010. *Testing Statistical Hypotheses of Equivalence and Noninferiority*, 2nd ed. New York: CRC Press.

Wellek, Stefan, and Peter Ziegler. 2017. EQUIVNONINF: Testing for Equivalence and Noninferiority. R Package Version 1.0. Available online: https://CRAN.R-project.org/package=EQUIVNONINF (accessed on 31 May 2018).

Winer, Ben J. 1962. *Statistical Principles in Experimental Design*. New York: McGraw-Hill.

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

econometrics

MDPI

Article

# Important Issues in Statistical Testing and Recommended Improvements in Accounting Research

**Thomas R. Dyckman [1,*] and Stephen A. Zeff [2]**

[1] Accounting Department, Cornell University, Ithaca, NY 14850, USA
[2] Accounting Department, Rice University, Houston, TX 77005, USA; sazeff@rice.edu
* Correspondence: trd2@cornell.edu

Received: 17 December 2018; Accepted: 26 April 2019; Published: 8 May 2019

**Abstract:** A great deal of the accounting research published in recent years has involved statistical tests. Our paper proposes improvements to both the quality and execution of such research. We address the following limitations in current research that appear to us to be ignored or used inappropriately: (1) unaddressed situational effects resulting from model limitations and what has been referred to as "data carpentry," (2) limitations and alternatives to winsorizing, (3) necessary improvements to relying on a study's calculated "*p*-values" instead of on the economic or behavioral importance of the results, and (4) the information loss incurred by under-valuing what can and cannot be learned from replications.

**Keywords:** model specification; model testing; reporting results (*p*-values); replications

## 1. Introduction

As professors of accounting for nearly 60 years and past presidents of the American Accounting Association, we are concerned about the quality of statistical research in accounting. This article is a call to our accounting colleagues, and also perhaps to those in other fields, to invest substantial time and effort toward improving their requisite knowledge and skill when conducting the appropriate statistical analysis. Involving expert statisticians may be helpful, as we all need to recognize the limitations in our own knowledge in order to tap into this expertise. Our heightened interest in improvements to the quality of statistical analysis in accounting research was in response to attending research presentations and reading the current literature.

Several years ago, we suggested several improvements to statistical testing and reporting (Dyckman and Zeff 2014). In that paper, we reviewed the 66 articles involving statistical testing that accounted for 90 percent of the research papers published between September 2012 and May 2013 in *The Accounting Review* and the *Journal of Accounting Research*, two leading journals in the field of accounting. Of these 66 papers, 90 percent relied on regression analysis. Our paper examined ways of improving the statistical analysis and the need to report the economic importance of the results.

An extension of these concerns was included in a commissioned paper included in the 50th anniversary of Abacus (Dyckman and Zeff 2015). We acknowledge several accounting academics who are also concerned with these issues, including Ohlson (2018), Kim et al. (2018), and Stone (2018), whose works we cite.

Concerns about statistical testing led to exploring the advantages of a Bayesian approach and abandoning null hypothesis tests (NHST) in favor of reporting confidence intervals. We also suggested the advantages—and limitations—of meta-analysis that would allow for the inclusion of replication studies in the assessment of evidence. This approach would replace the typical NHST process and its reliance on *p*-values (Dyckman 2016).

A fourth article which reviewed the first 30 years' history of the research journal, Accounting Horizons, continued our concern with the current applications of statistical testing to accounting research. An additional aspect of this article was the attention we gave to accounting researchers' seeming lack of interest in communicating with an audience of professionals beyond other like researchers, as if their only role as researchers was to enrich the research literature and not to contribute to the stock of accounting knowledge. We submit that accounting academics, because of the academic reward structure in their universities, tend to write for their peers. Accounting standard setters and accounting professionals, as well as those who make business and policy decisions, are all too often relegated to the sidelines. We argued that accounting research should, in the end, be relevant to important issues faced by accounting professionals, regulators and management, and that the research findings should be readable by individuals in this broader user community (Zeff and Dyckman 2018).

In the current paper, we expand on the statistical testing issues raised in our earlier papers, and we identify limitations often overlooked or ignored. Our experience suggests that many accounting professors, and perhaps those in other fields, are not familiar with, or equipped to, address them. We take up the following major topics: Model Specification and Data Carpentry, Testing the Model, Reporting Results, and Replication Studies, followed by A Critical Evaluation and A Way Forward.

## 2. Model Specification and Data Carpentry

The choice of a topic and related theory established the basis for the hypotheses to be examined and the concepts that will constitute the independent variables. Accounting investigations often rest only on a story rather than on a theory. A major problem here is that a story, but not theory, can be changed or modified, which encourages data mining (Black 1993, p. 73). Establishing the appropriate relationships require an understanding of the actual decision-making environment. These ingredients, along with the research team's insights and abilities, are critical to designing the research testing program and the data collection and analysis process. Failure to take them into account in the data-selection decision process and analysis was discussed in detail in a recent paper by Gow et al. (2016). There, the authors provided a detailed example (pp. 502–14) of how the decision environment can reflect its own idiosyncratic differences that, in turn, influence the data. For example, even if the business context is essentially the same across companies, data limitations remain. First, the data will inevitably reflect different sets of decision makers and different organizations, different time periods, different information, and, at least, some differences in the definitions of the variables deemed to be relevant. The interactions between these variables, and with any relevant but excluded variables, will, as the authors showed, lead to questionable results. How the selected variables interact with each other—and with any excluded but relevant variables—depends on the nature of the contextual environment in which the relation arises. We note here that careful research designs up front can reduce interactions among the independent variables. Authors can and should describe the decision environment and differences, if any, that have a potential impact upon the analysis and conclusions. A thorough analysis and description of the decision environments is essential and endows additional credibility on the research.

Typically, a concept can be operationalized by more than one variable. For example, firm size may be proxied by the number of employees or by revenue. Furthermore, the choice of a measure is often made according to data availability. Even the topic selection may be determined by the availability of an interesting data set. Unfortunately, authors usually do not acknowledge the latter and may fail to justify the selected variable measure. Once the hypotheses have been modeled and the variables with their measure selected, the decisions must not be altered, expanded, or dropped without full disclosure. Yet, we have seldom seen these explicit limitations revealed, let alone discussed. Authors appear to ask readers to accept implicitly that such alterations have not occurred. Even a careful reading may not reveal the authors' reasons for their specific choices. Authors should not assume that their choices are transparent and elect not to address the choice process.

The choice of the data set for the variables included in the study is critical. We think of this as the data carpentry, during which the raw data are melded into the data set for analysis. This is when data snooping, data mining, and related inappropriate activities must be avoided. Furthermore, researchers should not unquestioningly adopt a data set used by previous authors without verifying its accuracy and applicability to the current issue addressed. (For a discussion of what can occur, see Zeff 2016). Authors should also be alert to data sets reflecting different time periods, locations, or information processes. Conditions can be very different for the same variable across these dimensions. An assumption that data obtained under such circumstances will lead to valid conclusions cannot be sustained. Moreover, if the data source, timing, processing, or availability changes, the research team is obliged to bring these changes to the attention of the reader, together with the resulting limitations imposed on the findings.

### *2.1. Assumption of Randomness*

The concept of hypothesis testing and its key elements, including test statistic, $p$-value, standard error, sampling distribution, significance level, rely on an implicit assumption of randomness. The investigation relies on the researchers obtaining a random sample from a well-defined population. Indeed, one of the purposes of hypothesis testing is to determine how big or small the random sampling error is with respect to the parameter value being tested under the null hypothesis. Accounting researchers, by their failure to address the issue, are taking this fundamental assumption for granted. This is unfortunate. Authors appear to be implicitly relying on Dunning (2012) assurance that randomness can be accepted if the reader can be assured that the researchers had no influence, intended or not, on the data process. Unfortunately, databases may be problematic in the context of random sampling. For example, these databases often cover the data for listed companies only, which can provide a biased sample if the research outcome is applied to non-listed companies. The decision to seek big data or even a large sample can lead to a similar problem (Harford 2014). Several examples with serious consequences are examined in this article.

A thorough defining of the population is essential, but is not easily accomplished, and often remains unspecified by the authors. An implicit assumption of randomness may be comforting, but it is not adequate. Authors are obliged to expend the necessary human capital to alert the reader to possible limitations in their data and how any such limitations could affect their results. An example is provided by big data (Boyd and Kate 2011). Unless the research design takes the sampling distribution into account, it becomes difficult to justify resampling and randomization. The authors recall no recent accounting papers, including those relying on big data, that have addressed this situation. The process of determining whether a subset of big data amenable to the theory of relevance could be identified is likely to doom any honest sampling process. Additionally, it would preclude replication.

### *2.2. Model Modifications*

Once the hypotheses have been modeled and the variables have been selected and measured, any changes must be justified with full disclosure. Yet we have seldom seen such changes revealed. Authors apparently expect readers to accept implicitly that such alterations have either not occurred or are appropriate. A new approach to reduce this problem is being explored that requires authors to describe their choices in advance of executing the research project and to communicate to the editors any changes thereafter (Bloomfield et al. 2018; Kupferschmidt 2018). However, there is no assurance that this requirement will always be met, because the action may occur before initial submission.

### *2.3. Winsorizing*

It is not uncommon to find accounting studies whose authors have winsorized their data and assumed their readers understand the process. By winsorizing, the authors are attempting to prevent what they regard as outliers in the data from unduly influencing the results. Authors using this approach apparently assume that the outliers do not belong to the set defined by the variable under consideration. Retaining these data points, if inappropriate, will bias the results. Each such data point also has a larger impact on the results. However, the adjustment process used is generally ad hoc.

We submit that data points omitted from the analysis require individual justification based on analysis. An omitted observation might even be the most interesting data point, were it to be investigated. Or it may be due to factors not associated with the other sample data, a possibility advanced by Belsley et al. (1980). There is no theory justifying winsorizing (or truncation). These methods also make replication decidedly more difficult.

Winsorizing is one example of inappropriate data manipulation practiced during data carpentry, together with data mining and data snooping. Other examples of inappropriate data activities involved in establishing the data set include omitting data obtained under different circumstances, such as from different companies, time periods, or locations. Data produced by different individuals operating under different procedures and dissimilar situations also should be assumed to be inappropriate. See Zeff (2016) for an example. Inclusion of any such data must be thoroughly vetted and disclosed, including its impact on the identified hypothesis.

Instead of winsorizing, we suggest the authors consider robust regression (RR) recommended by Leone et al. (2019). Using simulation, the authors find RR outperforms winsorization and truncation, which are largely ineffective. The authors suggest using approaches based on residuals for which RR is both theoretically appealing and easy to implement.

## 3. Testing the Model

An approach that some researchers are turning to—and which we encourage—is to apply the research model to alternative relevant data sets (Lindsay 1995). While this approach is time-consuming, the results, when thus confirmed, are more compelling. Additionally, tests might also be run on logical choices of subsets of the original data. An interesting alternative would be to test the model through one or more predictions, although we have not seen much enthusiasm for this option.

The data sets in accounting studies, other than experimental ones, tend to be large. Data sets over 10,000 are not uncommon. Small data sets, below 25, are unusual, except in behavioral accounting work, an area we are not explicitly targeting here. Accounting journals would not reject a small sample of, say, 25 if there were a compelling reason for the size and the results were clearly of interest. Multiple hypotheses based on a single data set are common, and the use of a data set to examine a different hypothesis by a different research team is not uncommon. The concern here, however, is that any data problem that is unresolved or miss-handled in the original research is likely to influence the new work. We believe that a borrowed data set must go through a thorough analysis before it can be presumed to be appropriate for testing a new hypothesis. This is appropriate for the original paper and even more so for a replication or the use of the sample by a new investigating team.

Our reading of the accounting literature indicates that some authors do rely on the same data set to test multiple hypotheses. Yet Floyd and List (2016, p. 454) observe, "When multiple hypotheses ... are considered together, the probability that at least some Type 1 errors are committed often increases dramatically with the number of hypotheses." See also Ohlson (2018). Fortunately, studies that perform multiple tests on a single data set are not difficult to identify. Authors should either alert readers to what amounts to over-testing or confine their analysis to the critical issue of the study. Additionally, it often happens that the ideal data set is impossible to obtain. When this is the case, the ideal data set should be acknowledged and its absence justified, including any change in the variables selected and their measures. Unfortunately, while applauding the changes in the reviewing process championed by Bloomfield et al. (2018) and others, we believe there is currently no way to assure that data tampering, including data mining, does not occur prior to the submission of a research paper to a journal.

### *Sample Size Concerns*

Sample size has an important effect on statistical tests. Many accounting studies involve very large samples, while small samples are rare except in situations involving behavioral experiments. Indeed, researchers appear to believe that very large samples are somehow superior or more likely to generate statistical significance. Yet what the researcher observes in the sample may not be true at the population level. This is particularly likely to happen when the sample data ultimately used in the test are substantially fewer than what are contained in the initial sample. This situation, while not common in accounting research, does occur. See Santanu et al. (2015). The authors reported a sample size of 11,262 after concluding that, for undisclosed reasons, 14,042 observations were rejected, a rejection level of 55% of the available data. This condition alone does not invalidate the research. Such a large reduction calls for an explanation, which was not given. Indeed, the authors were likely engaged in data snooping, which could lead to such a large reduction in sample size. Researchers should also not forget the Jeffery-Lindley paradox, which shows that, with a large enough sample size, a 0.05 significance result can correspond to assigning the null hypothesis a high probability (0.95). This result does not hold, however, for interval hypotheses. See also Ohlson (2015) on sample size.

## 4. Reporting Results

The most important point to be made here is not whether a reported significant *p*-value, say at the 0.01 level, has been obtained but rather the overall credibility of the work. Credibility depends on a myriad of factors. These factors include the accuracy and veracity of not only the model but also the variable choices and their measurement. For example, if the model were to omit an important explanatory variable, the effect of the omission may be subsumed under one or more of the other explanatory variables, causing it or them to appear more significant than would otherwise be the case. It remains the responsibility of the authors to consider the challenges that serious readers are likely to raise concerning the central model, the variable choices or omissions, and how they should be operationalized. Assuring that readers have an adequate description of the methodology, including the model, data set, and the computer protocol in order to permit, and indeed invite, a replication would be one template for ascertaining that the essential elements have been disclosed. Improperly executed research can, as pointed out most recently by Kim et al. (2018) and by Lindsay (1994, 1995), ultimately lead to poor decisions and may even inflict serious social harm.

### *4.1. Reporting p-Values*

First, it is useful to define what a *p*-value is. A *p*-value is the probability of observing a value of the test statistic that is as extreme as, or more extreme, than the value resulting from the sample, given that the null hypothesis is true. It is both a conditional probability and a statistic with a sampling distribution. Our view of *p*-values' contribution to the research in accounting is best captured by the following quotation: "Misinterpretations and abuse of statistical tests, confidence intervals, and statistical power have been decried for decades, yet remain rampant. A key problem is that there are no interpretations of these concepts that are at once simple, intuitive, correct, and foolproof. Instead, correct use and interpretation of these statistics requires an attention to detail which seems to tax the patience of working scientists. This high cognitive demand has led to an epidemic of shortcut definitions and interpretations that are simply wrong, sometimes disastrously so—and yet these misinterpretations dominate much of the scientific literature" (Greenland et al. 2016, p. 337). Authors of publications in accounting and related fields invariably report and rely on small *p*-values (0.01, 0.05, 0.10) as an indication of the importance of their work. Yet, there is no theoretical justification to guide researchers in selecting a specific *p*-value to justify the conclusion that significance has been in fact attained or, if so, whether it matters. The smaller the calculated *p*-value, the more comfortable the researcher may feel in rejecting the null hypothesis. However, this information alone considers neither the importance of the results nor the costs of an incorrect rejection. Johnstone and Lindley (1995) argue that significance at 0.05 is meaningless without knowing the sample size, the magnitude of the observed effect, and the operational importance of that effect. Alone, it fails to assure readers that the analysis has even uncovered a useful result. Indeed, the *p*-value, so endemic to much of accounting research, is useless by itself. Without some measure of the impact (size effect) or economic importance of the result, little if anything has been learned. Ziliak and McCloskey (2004), after reviewing the literature in several fields, found that nine out of 10 published articles make this mistake. If, in the accounting literature for example, the reported results suggest that a specific behavior reduces audit delays, little is gained unless a reasonably accurate determination of the economic impact of the revealed delay to an identified clientele is determined. Authors, then, must identify in advance the user of the research result who would find the size or impact important. We note that several distinguished journals in other fields, including Basic and Applied Social Psychology, have banned the use of *p*-values, while others, including PLOS Medicine and the Journal of Allergy and Clinical Immunology, actively discourage its use. The American Statistical Association has recently urged caution in relying on statistical significance at the traditional 0.05 level as a basis for claims (Wasserstein et al. 2019). We continue to be dismayed that editors or reviewers in our field appear to require a reported *p*-value of 0.10 or less as a necessary condition for publishing the results of a study relying on statistical research in accounting. Perhaps the recent final appeal to renounce relying on *p*-values was sounded by a recent paper published in The American Statistician in which Wasserstein et al. (2019, p. 1) state: "Don't conclude anything about scientific or practical importance based on statistical significance (or lack thereof)." The same issue of the American Statistician includes 43 additional papers that addresses statistical Inference in the twenty-first century.

An improvement would be to report the Bayes factor, as suggested in the context of accounting research by Kim et al. (2018). This is the ratio of the observed value, or a lesser value under the null hypothesis, to the probability of the observed value or a lesser value under the alternate hypothesis. The approach is a Bayesian concept and reflects the ratio of the new knowledge to what was previously presumed. This concept is consistent with using new information to update one's prior beliefs, a common accounting objective. The calculation necessitates that the researcher initially specifies the denominator of the likelihood ratio. Furthermore, reporting a confidence interval is an improvement over reporting a *p*-value and provides more information.

Yet, it is wrong to conclude that *p*-values are useless. A *p*-value from a well-executed study could provide useful information. For example, such a value could indicate that there is something unusual or interesting in the analysis, justifying further study. Alternatively, a review of the process may indicate a flaw in the analysis. Perhaps the data or the data-carpentry activity was faulty. The model may be inappropriate. There could have been an error in the computer program. And well-done studies could suggest further potential regardless of the resulting *p*-value.

### *4.2. Effect Size (ES) or Economic Importance (EI)*

Determining the effect size (Cohen 1990; Stone 2018) or the economic importance (Basu 2012; Dyckman 2016) of the results should be the sought-after objective of research. One way of presenting the result is to use a confidence interval measure to capture the impact on the specific costs or benefits suggested by the research. Yet we could locate very few articles in the accounting literature that have rcct size. Judd et al. (2017, p. 34) provide an example that does addresses this issue. "In terms of economic significance, we find on average, a one standard deviation increase in CEO narcissism [proxied by the CEO's picture size in the annual report and the CEO's relative cash and non-cash pay] is associated with a 2.4 percent to 3.3 percent increase in external audit fees, which equates to approximately $116,497 to $160,183 for our sample mean firms." We note here that the authors elect, not to emphasize the economic significance by reporting their findings in either the Synopsis or Conclusion. Their approach may be explained in part by the studies limitations described in their footnote two.

Irani et al. (2015, p. 847) provide a recent example of explicitly addressing the statistical significance/economic importance issue. They state, "We recognize the small magnitude of the univariate market reaction, which, although *statistically significant*, is arguably not *economically significant*" [emphasis added]. It is essential to identify the importance of a study's results and not just rely on whether one or more hypotheses are statistically significant at a specific reported *p*-value. Furthermore, as noted earlier, not finding a variable to be statistically significant does not necessarily mean it is unimportant. Eshleman and Lawson (2017, p. 75) report that they "find a positive association between audit market concentration and audit fees." Their main conclusion was, "As a whole, our findings suggest that U.S. audit market concentration is associated with both higher audit fees and higher audit quality" (p. 76). Unfortunately, we are not informed of the importance of the impact that the concentration had on audit fees.

A more recent accounting study by Brown et al. (2018) recognizes the limitation of research that stops with the reporting of a significant *p*-value, yet the authors fail to deal with the economic importance.

## 5. Replication Studies

Researchers quite understandably seek to explore new questions in their research. Thus, it is not surprising that replication studies are rare. Yet, regardless of a study's results, whether it is an important finding or an unsuspected failure to support an expected finding, replication studies are relevant and important. Replications are, however, decidedly difficult to perform satisfactorily and are not welcomed by many journals across the accounting research landscape. Replications therefore must pass rigorous scrutiny. Several new journals have been launched, and a few existing outlets now do consider replication papers. New journals have been initiated recently that do consider replication papers. There are also existing journals that have published replications for some time. The *American Economic Review* and the *Journal of Applied Econometrics* are leaders in their field, and they publish about a 30 percent of the replications in economics (Reed 2018). The Replication Network is an excellent source for information on replication studies.

A few replication studies have begun to appear in accounting journals. The ability to fully replicate depends on an agreed theory. Stories do not provide ideal bases for replications. An early, well executed replication study in accounting that deserved and ultimately achieved publication was done by Bamber et al. (2000), who replicated Beaver (2000) Seminal Award-winning paper. It is interesting to note that the authors' work was published in *Accounting, Organizations and Society*, not in one of the journals noted for empirical/archival research. This is not where one would expect to find it, because it was rejected by the journal that published Beaver's article.

Mayo (2018, preface) sets a high bar when applied to any study, including replications. She advises that the results of any study need to be "severely tested." She writes, "The [severe] testing metaphor grows out of the idea that before we have evidence for a claim, it must have passed an analysis that could have found it flawed." In other words, as Mayo states, "a hypothesis must have passed an analysis that could have found it flawed" (Mayo 2018, preface). We are unable to locate an accounting paper that currently meets this or a similar rigorous standard. We would encourage researchers to consider applying this test.

The rewards for attempting replications are currently not enticing. Furthermore, precise replications have seldom been possible, in part because the necessary information to perform such studies is not ordinarily made available by authors. Nevertheless, we encourage replications because they provide the test that what has been found matters. This could be the case if a study were to identify a potential measurable and meaningful size effect. Unfortunately, in accounting, we are left with a sparse landscape of replication studies that provide confirmation of important results or which encourage the publication of synopses of important replications (particularly of effect size or economic importance) that are well-executed. Fortunately, journals do exist that consider replications. One relatively new journal is The International Journal for Re-Views in Empirical Economics (IREE).

A recent study by Peng (2015) concludes that a high proportion of published results across fields were not reproducible by replication. This does not reflect well on the academic community. Studies of reproducible results lend credence to the value of the exercise. On the other hand, failed replications cast doubt on the original results. Brodeur et al. (2018, abstract) report that in "Applying multiple methods to 13,440 hypothesis tests reported in 25 top economics journals in 2015, we show that selective publication and p-hacking is a substantial problem in research employing DID [differences-in-differences] and (in particular) IV [instrumental variables]." A large study reported in Science describes the results of 270 researchers replicating 100 experiments reported in papers published in 2008 in three high-ranking psychology journals (Aarts et al. 2015). The replications yielded the same results according to several criteria. They showed that only 39% of the original findings could be replicated unambiguously. This information is not encouraging.

Recently, we came across an announcement of a new e-journal, SURE (for The Series of Unsurprising Results in Economics), which commits to publishing high-quality research even with unsurprising findings. The journal emphasizes scientifically important and carefully executed studies with statistically insignificant results or otherwise unsurprising findings. Studies from all fields of economics will be considered. As a bonus, there are no submission fees.

An additional process that can increase our confidence in results, and one that merits consideration, is meta-analysis (Dyckman 2016; Hay and Knechel 2017; Stone 2018). An advantage of meta-analysis is that it suggests the integration of current and future investigations of a given phenomenon. Using this technique reflects a cumulative approach to a specified hypothesis by which a triangulation on the topic can lead to a better understanding of a common research objective. This approach allows competing explanations of a given phenomenon to be merged to produce a result depicted, for example, by a confidence interval. Opting for a meta-analysis approach provides an opportunity for researchers to reexamine and perhaps reinforce important past results. The adoption of meta-analysis in accounting remains exceedingly rare, perhaps partially reflecting editor reluctance to publish replications.

## 6. A Critical Evaluation and a Way Forward

We believe that a healthy skepticism should abide from the beginning and then remain with the authors throughout the process. This attitude must extend to the apparent confirmation of any basic hypotheses. Asking why and how the authors could be wrong, or whether they have missed an important influence on the analysis, should accompany the entire investigation. We fear, however, that authors may not expend the time and human capital to critique their work sufficiently. Readers are often not sufficiently familiar with the authors' subject to discover a study's short-comings. Indeed, the authors themselves are the most likely to be conscious of their study's limitations and potential extensions. They should advise readers on where additional analysis would be most fruitful, including the known or suspected limitations to their own work. Essentially all studies have limitations, and that alone provides ample reason for encouraging disclosure, and in important situations, replications. Access to all that went into the original analysis should be available, on request if necessary.

The process begins with the selection of an important question or issue. A relevant and available data source will need to be identified or created in order to proceed. Once these are determined, we suggest that the investigation concentrate on operationalizing the dependent variable, identifying the independent variables, including their interrelationships, and how they can best be operationalized. This approach then allows the research team to craft the model. The research team should keep a record of all assumptions and decisions made in this process. Attention will need to be given to variable interactions, and whether the conditions affecting the observations are different in a way, or ways, that could have an impact upon the findings.

The primary objective is to reveal an important result, one that is based on an important economic or behavioral impact. If no such impact was revealed, the authors should take what has been learned, pack their bags and move on to a new project. If the process has been appropriate, the authors should take what has been learned and seek a new topic worthy of investigation. There is no reason to attempt to resurrect a deceased patient.

Authors should avoid placing reliance on $p$-values, concentrating instead on the economic or behavioral implications of the work. If the result is controversial and the analysis is well done, so much the better. Reporting confidence intervals instead of $p$-values should be the common practice. (Dyckman 2016; Stone 2018).

Our discipline, and others as well, will benefit from applying new approaches to establishing the importance of the phenomena being studied. Stone (2018, p. 113), has recently suggested exploring triangulation, a complementary approach that could, for example, combine a quantitative and a behavioral approach to a problem. Also see Jick 1979. Such studies can provide insights not otherwise apparent. Combining methodologies has limitations, one being that replications are extremely difficult to execute. The adoption or reliance on new methodologies or pirating them from other disciplines is also to be encouraged.

Thus, we are in accord with Johnstone (1990) and with Kim et al. (2018, p. 14) that a Bayesian approach to statistical hypothesis testing, which recognizes the importance of the power of the test, offers a means of dealing with the inherent bias introduced by the conventional hypothesis testing currently prevalent in accounting. Furthermore, we encourage authors, as a few have done, to consider areas and methodologies from sister disciplines, including medicine and even philosophy. In this paper, we have relied on medicine (Ioannidis 2005), epidemiology (Greenland et al. 2016), and on philosophy (Mayo 2018). We maintain that there is much to be learned from these and other disciplines.

The ultimate importance of a study is the economic or behavioral consequences of the research findings, not the statistical significance as reflected in a calculated $p$-value. Investigators should be looking for a size or economic Importance measure. The $p$-value may provide some information, as described above. However, it should not be considered the study's goal or a measure of its contribution.

Finally, we would encourage accounting professors to improve their statistical knowledge. The resources are immediately available. In addition, universities hold excellent summer programs. One current example is an August program under the auspicious of Northwestern and Duke Universities.

## 7. Conclusions

Our purpose in this paper has been to encourage authors to consider ways to improve their research. Our paper began by emphasizing the importance of careful model building, data selection, and carpentry as well as the disclosures essential to assuring the integrity and reproducibility of the work. The choice of topic should be based on its relevance to practice or on improving research techniques.

Researchers must understand the limitation of relying on $p$-values as the sole or even primary support for their findings. Confidence intervals should replace reporting $p$-values. Until the reader is informed of the importance of the results, framed by the economic or behavioral consequences of the findings, the research objective has not been achieved. The issues addressed in this paper are advanced as a means of accomplishing this objective. To assist in the mission, we strongly encourage researchers to read the paper by Greenland et al. 2016, which offers a tutorial on statistical tests, $p$-values, confidence intervals, and power. (Greenland et al. 2016 is also available through Dave Giles' Blog of 2 May 2019 (Giles 2019) and Bob Jensen's Blog 2 September 2019 (Jensen 2018). The Jensen blog includes the article).

**Author Contributions:** Only the categories that apply to our paper are covered here. Unless indicated, the authors contributed equally. Data curation was mainly obtained from prior research papers of the authors and papers authored by others in and outside our field. The original draft was primarily undertaken by the first author (T.R.D.) while the second author (S.A.Z.) took the lead on project review and editing. The project was under our own supervision and there was no outside project administration. The project was entirely funded and written by the authors, T.R.D. and S.A.Z.

**Funding:** The project received no outside funding.

**Conflicts of Interest:** There were no conflicts of interest.

## References

Aarts, Alexander A., Joanna E. Anderson, Christopher J. Anderson, Peter Attridge, Angela Attwood, Jordan Axt, Molly Babel, Štěpán Bahník, Erica Baranski, Michael Barnett-Cowan, and et al. 2015. Estimating the reproducibility of psychological science. *Science* 349: 6251.

Bamber, Linda Smith, Theodore E. Christensen, and Kenneth M. Gaver. 2000. Do we really "know" what we think we know? A case study of seminal research and its subsequent overgeneralization. *Accounting Organizations and Society* 25: 103–29. [CrossRef]

Basu, Sudipta. 2012. How can accounting researchers become more innovative? *Accounting Horizons* 26: 851–70. [CrossRef]

Beaver, William H. 1968. The information content of annual earnings announcements. *Empirical Research in Accounting, Selected Studies 1968. Supplement to Journal of Accounting Research* 6: 67–92. [CrossRef]

Belsley, David A., Edwin Kuh, and Roy E. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley.

Black, Fischer. 1993. Beta and return. *Journal of Portfolio Management* 20: 8–18. [CrossRef]

Bloomfield, Robert, Kristina Rennekamp, and Blake Steenhoven. 2018. No system is perfect: Understanding how registration-based editorial processes affect reproducibility and investment in research quality. *Journal of Accounting Research* 56: 313–62. [CrossRef]

Boyd, Danah, and Crawford Kate. 2011. Six Provocations for Big Data. A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, September 21. Available online: https://ssrn.com/abstract=1926431 (accessed on 14 March 2019).

Brodeur, Abel, Nikolai Cook, and Anthony G. Heyes. 2018. Methods Matter: P-Hacking and Causal Influence in Economics. *Dated August 2018*. Available online: https://drive.google.com/file/d/10an9l3ndpjIfBVy1q5tC-9YGrVzPvmfg/view (accessed on 15 March 2019).

Brown, Jason P., Dayton M. Lambert, and Timothy R. Wojan. 2018. At the intersection of null findings and replication. *The Replication Network*. August 23. Available online: https://replicationnetwork.com/2018/08/23/brown-lambert-wojan-at-the-intersection-of-null-findings-and-replication/ (accessed on 23 August 2018).

Cohen, Jacob. 1990. Things I have learned (so far). *The American Psychologist* 45: 1304–12. [CrossRef]

Dunning, Thad. 2012. *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge: Cambridge University Press.

Dyckman, Thomas R. 2016. Significance testing: We can do better. *Abacus* 52: 319–42. [CrossRef]

Dyckman, Thomas R., and Stephen A. Zeff. 2014. Some methodological deficiencies in empirical research articles in accounting. *Accounting Horizons* 28: 695–712. [CrossRef]

Dyckman, Thomas R., and Stephen A. Zeff. 2015. Accounting research: Past, present and future. *Abacus* 51: 511–24. [CrossRef]

Eshleman, John Daniel, and B. P. Lawson. 2017. Audit market structure and audit pricing. *Accounting Horizons* 31: 57–81. [CrossRef]

Floyd, Eric, and John A. List. 2016. Using field experiments in accounting and finance. *Journal of Accounting Research* 54: 437–75. [CrossRef]

Giles, David. 2019. Blog. Available online: https://davegiles.blogspot.com/ (accessed on 5 February 2019).

Gow, Ian D., David F. Larcker, and Peter C. Reiss. 2016. Causal inference in accounting research. *Journal of Accounting Research* 54: 477–523. [CrossRef]

Greenland, Sander, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman. 2016. Statistical tests, *p*-values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology* 31: 337–50. [CrossRef]

Harford, Tim. 2014. Big data: A big mistake? *Significance* 11: 14–19. [CrossRef]

Hay, David C., and W. Robert Knechel. 2017. Meta-regression in auditing research: Evaluating the evidence on the Big N audit firm premium. *Auditing: A Journal of Practice & Theory* 36: 133–59.

Ioannidis, John P. A. 2005. Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association* 294: 218–28. [CrossRef]

Irani, Afshad J., Stefanie L. Tate, and Le Xu. 2015. Restatements: Do they affect auditor reputation for quality? *Accounting Horizons* 29: 829–51. [CrossRef]

Jensen, Robert. 2018. Blog. Available online: http://faculty.trinity.edu/rjensen/ (accessed on 2 September 2018).

Jick, Todd D. 1979. Mixing qualitative and quantitative methods: Triangulation in action. *Administrative Science Quarterly* 24: 602–11. [CrossRef]

Johnstone, David J. 1990. Sample size and the strength of Evidence: A bayesian interpretation of binomial tests of the information content of qualified audit reports. *Abacus* 26: 17–35. [CrossRef]

Johnstone, David J., and D. V. Lindley. 1995. Bayesian inference given data "significant at α": Tests of point hypotheses. *Theory and Decision* 38: 51–60. [CrossRef]

Judd, J. Scott, Kari Joseph Olsen, and James Stekelberg. 2017. How do auditors respond to CEO narcissism? Evidence from external audit fees. *Accounting Horizons* 31: 33–52. [CrossRef]

Kim, Jae H., Kamran Ahmed, and Philip Inyeob Ji. 2018. Significance testing in accounting research: A critical evaluation based on evidence. *Abacus: A Journal of Accounting, Finance and Business Studies* 54: 524–46. [CrossRef]

Kupferschmidt, Kai. 2018. A recipe for rigor. *Science* 361: 1192–93. [CrossRef] [PubMed]

Leone, Andrew J., Miguel Minutti-Meza, and Charles E. Wasley. 2019. Influential observations and inference in accounting research. *The Accounting Review*. forthcoming. [CrossRef]

Lindsay, R. Murray. 1994. Publication system biases associated with the statistical testing paradigm. *Contemporary Accounting Research* 11: 33–57. [CrossRef]

Lindsay, R. Murray. 1995. Reconsidering the status of tests of significance: an alternate criterion of Adequacy. *Accounting Organizations and Society* 20: 35–53. [CrossRef]

Mayo, Deborah G. 2018. *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge: Cambridge University Press.

Ohlson, James A. 2015. Accounting research and common sense. *Abacus* 51: 525–35. [CrossRef]

Ohlson, James A. 2018. Researchers' Data Analysis Choices: An Excess of False Positives? Available online: https://ssrn.com/abstract=3089571 (accessed on 6 January 2019).

Peng, Roger. 2015. The reproducibility crisis in science: A statistical counterattack. *Significance* 12: 30–32. [CrossRef]

Reed, Robert. 2018. An Update on Progress of Replications in Economics. Available online: https://replicationnetwork.com/2018/10/31/reed-an-update-on-the-progress-of-replications-in-economics/ (accessed on 5 January 2018).

Santanu, Mitra, Hakjoon Song, and Joon Sun Yang. 2015. The effect of Auditing Standard No. 5 on audit report lags. *Accounting Horizons* 29: 507–27.

Stone, Dan N. 2018. The "new statistics" and nullifying the null: Twelve actions for improving quantitative accounting research quality and integrity. *Accounting Horizons* 32: 105–20. [CrossRef]

Wasserstein, Ronald L., Allen. L. Schirm, and Nicole. A. Lazar. 2019. Moving to a world beyond "p > 0.05". *The American Statistician* 73: 1–19. [CrossRef]

Zeff, Stephen A. 2016. "In the literature" but wrong: Switzerland and the adoption of IFRS. *Journal of Accounting and Public Policy* 35: 1–2. [CrossRef]

Zeff, Stephen A., and Thomas R. Dyckman. 2018. A historical study of the first 30 years of Accounting Horizons. *Accounting Historians Journal* 45: 115–31. [CrossRef]

Ziliak, Stephen T., and Deirdre N. McCloskey. 2004. Size matters: The standard error of regressions in the American Economic Review. *Journal of Socio-Economics* 33: 527–46. [CrossRef]

econometrics

Article

# Not *p*-Values, Said a Little Bit Differently

**Richard Startz**

Department of Economics, University of California, Santa Barbara, CA 93106, USA; startz@ucsb.edu

Received: 14 December 2018; Accepted: 7 March 2019; Published: 13 March 2019

**Abstract:** As a contribution toward the ongoing discussion about the use and mis-use of *p*-values, numerical examples are presented demonstrating that a *p*-value can, as a practical matter, give you a really different answer than the one that you want.

**Keywords:** *p*-value; Bayesian

**JEL Classification:** C0; C4

## 1. Introduction

The American Statistical Association statement on "Statistical Significance and *P*-values" (Wasserstein and Lazar 2016) aimed at reminding the statistics community about a number of pitfalls that are commonly fallen into, in the everyday use of *p*-values. The statement and accompanying introduction also pointed to the rich history of the statisticians who have articulated the issues, providing a long list of references. This raises a question: If no one has listened before, will they be swayed by this latest exhortation? Perhaps a numerical example might be more convincing—an example that illustrates that the issue is less that the common use of *p*-values is philosophically misguided, and more that the numbers can just be completely wrong (for evidence that these issues are of real, applied importance in economics and finance see Kim and Ji (2015).

One way in which to understand the misuse of the *p*-value is as a misapplication of the *modus tollens* argument. Suppose we had data that would prove a null hypothesis to be true or false, with certainty. If the null hypothesis is true, then the data would support the null with certainty. So if the data did not support the null, we would know that the null is false. However, such logic does not apply to statistical reasoning, where the data does not give answers with certainty. If the null is true, then a small *p*-value is unlikely. The fallacy of applying *modus tollens* is that it may be that if the null is false, then a small *p*-value is also unlikely.

Problems with the misuse of *p*-values have been understand for a very long time—at least in principle. The purpose here is to provide a new-but-simple example of the disconnect between a *p*-value and the probability that a null hypothesis is true—adding to the long list of existing examples. Beginning with a quick review of what has been said in the past may be useful. There are a number of concerns with regard to *p*-values, which have been discussed at least as far back as Berkson (1942), and as recently as Wasserstein and Lazar (2016). The latter also includes many references. I focus here solely on the issue that a *p*-value is not designed to speak to the relative merits of a null hypothesis versus the alternative. Nickerson (2000) explains the problem, and gives many further references. Trafimow (2003) puts the matter succinctly, "although one can calculate the probability of obtaining a finding given that the null hypothesis is true, this is not equivalent to calculating the probability that the null hypothesis is true given that one has obtained a finding."[1] Trafimow (2005) is more pointed, writing,

[1] See also Trafimow (2015) and Trafimow and Marks (2015). Trafimow offers a bit of history and explanations that are suitable for students at https://www.youtube.com/watch?v=dsp_hSIsacQ.

"A $p$-value can be a dramatic overestimate or underestimate of the desired posterior probability of the null hypothesis depending on the prior probability of the null hypothesis and the probability of the finding given that the null hypothesis is not true."

Dickey (1977) points out that the area under the tail is not, in general, a good approximation to the Bayes factor. Berger and Sellke (1987) summarize the issue nicely, writing that "actual evidence against a null (as measured, say, by posterior probability or comparative likelihood) can differ by an order of magnitude from the $p$-value. ... The overall conclusion is that $p$-values can be highly misleading measures of the evidence provided by the data against the null hypothesis" Trafimow and Rice (2009) show that the $p$-value need not even be very highly correlated with the true probability of the null hypothesis.

The point addressed here is the ASA's second principle, "$P$-values do not measure the probability that the studied hypothesis is true ... " Or as Pearson (1938) wrote eight decades ago, "Gosset ... had a tremendous influence on the ... idea which has formed the basis of all the ... researches of Neyman and myself. It is the simple suggestion that the only valid reason for rejecting a statistical hypothesis is that some alternative hypothesis explains the events with a greater degree of probability." Hubbard and Bayarri (2003) explain the difference between Fisher's advocacy of the $p$-value and the idea of Neyman and Pearson to compare a null hypothesis to an alternative, offering historical perspectives as well. Robinson and Wainer (2002) discuss a number of issues with the use of $p$-values, including the point that " ... many users of NHST [Null Hypothesis Significance Testing] interpret the result as the probability of the null hypothesis based on the data observed. ... This error suggests that users really want to make a different kind of inference—a probabilistic statement of the likelihood of the hypothesis", which is the point that we pursue below.[2] Hubbard and Lindsay (2008) write that "*P-Values Exaggerate the Evidence Against the Null Hypothesis*". This is the most damning criticism of the $p$-value as a measure of evidence." (Emphasis in the original). We shall see, however, that it is also possible for a $p$-value to understate the evidence against the null.

## 2. The General Problem

Wasserstein and Lazar (2016) succinctly remind everyone that "Informally, a $p$-value is the probability under a specified statistical model that a statistical summary of the data ... would be equal to or more extreme than its observed value." Following Trafimow (2005), suppose that we call finding that probability to be equal to or more extreme to be the "finding", or simply, $F$. The "philosophical" problem is that $p$-values summarize $P(F|Hypothesis)$, while we are, with rare exception, interested in $P(Hypothesis|F)$. The two are related by Bayes' Theorem, but they are not the same. Pointedly, they need not even be close. As a reminder, the $p$-value is calculated by assuming that the null hypothesis is true, and then calculating the probability that some observed outcome would come about under the null hypothesis. The classic case is to calculate the probability that an estimated parameter should be as far or farther from a point that is specified by the null, as is the observed estimate. The $p$-value is $P(F|Hypothesis)$, but from Bayes' theorem:

$$P(Hypothesis|F) = \frac{P(F|Hypothesis) \times P(Hypothesis)}{P(F)} \tag{1}$$

The generic reason that the $p$-value need not be close to the conditional probability of the hypothesis is that the $p$-value is missing the other two elements in Equation (1). Since this is obvious, it is probably worth commenting on why the deployment of the $p$-value remains nearly pervasive. The requirement to specify $P(Hypothesis)$ is sometimes viewed as non-scientific, as it comes from something other than the data at hand. Also, the specification of $P(F)$ generally requires considerable

[2] Robinson and Wainer (2002) take a more sanguine view of the possible damage of conflating the Fisherian and Neyman–Pearson approaches than does Hubbard and Bayarri (2003).

information about the alternative hypothesis, certainly much more than merely the idea that the alternative is anything other than the null.

The notion that the *p*-value summarizes $P(F|Hypothesis)$ is an oversimplification, of course, as it ignores conditioning on the econometric specification of the entire estimate. Really, the *p*-value gives $P(F|Hypothesis; specification)$. Conditioning on specification carries through to the left side of Equation (1), but in what follows, I will omit it for the sake of brevity. In addition, when one applies Bayes' Theorem, the result is really conditioned on the prior specified, although this is traditionally omitted from the notation. It is also true that frequentists and Bayesians sometimes disagree over the entire nature of the statistical enterprise, including even the meaning of "probability." Nothing in what follows speaks to these deeper issues.

## 3. A Simple Example of the Problem

Consider the decision of whether a coin is fair or not, based on the number of heads, $h$, observed after $n$ tosses. If there are 26 heads out of 64 tosses, the *p*-value is 0.08 (so the null seems very unlikely). Though a bit short of the magic number 0.05, that's a sufficiently low *p*-value that a sympathetic editor might consider publication. Doing the Bayes' Theorem calculation requires some additional assumptions, but arguably innocuous assumptions suggest that the coin is more likely than not, fair, $P(fair|data) = 0.59$—a strikingly different conclusion ("Arguably innocuous" being taken here as the prior odds for the coin being fair being 50/50, and that if the coin is not fair, all we know is that the probability of a head is between zero and one). Note that since the posterior is not far from the prior, we would conclude that the data is not very informative, which is probably not the conclusion one would draw from looking at the 0.08 *p*-value.

In this example, the studied hypothesis is that the mean chance of a head is $\mu = 0.5$, and that the *p*-value is $F_B(h, n\mu = 0.5) + (1 - F_B(n - h, n\mu = 0.5)$, where $F_B$ is the cdf (cumulative distribution function) of the binomial distribution. Bayes Theorem gives us $P(\mu = 0.5|h, n)$ as a function of the binomial probability mass function $P_B$, and a prior over $\mu$, $P(\mu)$.

$$P(u = 0.5|h, n) = \frac{P_B(h, n|\mu = 0.5) \times P(\mu = 0.5)}{\int_{-\infty}^{\infty} P_B(h, n|\mu) \times P(\mu) d\mu}$$

Unlike the formula for calculating the *p*-value, here, the answer requires some extra inputs. Most researchers would probably agree that the probability of a head is between zero and one, so that outside that range, $P(\mu)$ equals zero. Beyond that, we probably want to put some finite mass onto the studied hypothesis. $\pi \equiv P(\mu = 0.5) = 0.5$ might be thought of as neutral.[3] Also, we might be as ignorant as possible about alternative values, by spreading the rest of the mass uniformly between the limits, so between zero and one, $P(\mu) = 1$. This gives:

$$P(\mu = 0.5|H, n) = \frac{P_B(h, n|\mu = 0.5) \times \pi}{\int_0^1 P_B(h, n|\mu) d\mu \times (1 - \pi) + P_B(h, n|\mu = 0.5) \times \pi}$$

Using these assumptions gives us the $P(\mu = 0.5|h = 26, n = 64) = 0.59$ value given above. Of course, varying counts of heads give different probabilities and *p*-values, with the relation between the two for the 64 coin tosses shown in Figure 1. If the *p*-value gave the probability that the hypothesis were true, the plot would lie along the 45° line. However, it does not do so. Regardless, the more important lesson is that the curve is often very far from the 45° line, and indeed, it can lie either above or below.

[3] A dedicated Bayesian might point out that in the presence of prior information, a relatively non-informative prior would not be appropriate. An informative prior might lead to $P(\mu = 0.5|H, n)$, either closer to the *p*-value or farther away.

Different priors do give different probabilities for the studied hypothesis, so there may well be a prior for which the $p$-value does coincide with the correct probability for the studied hypothesis. In the example here, if we put a prior weight, $\pi$, on the fair coin that is equal to 0.04, we obtain a probability that is equal to the 0.08 $p$-value. Still, it seems likely that a $p$-valuista who rejects a fair coin did not intend to declare a prior of 96 percent against the coin being fair.

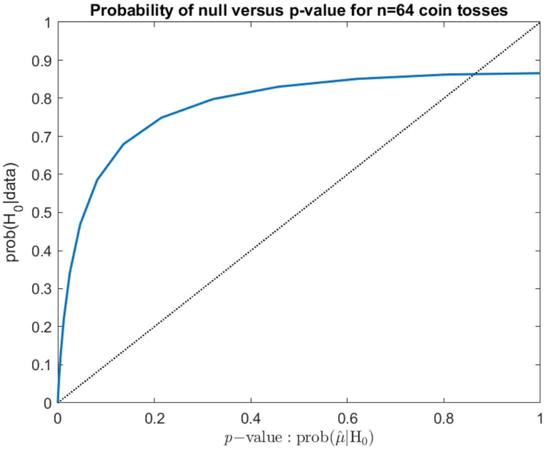


**Figure 1.** Relation between probability of the null and the $p$-value for various observed heads.

## 4. Summary

The point in the ASA statement is not that $p$-values give the wrong answer; the point is that $p$-values usually commit what (Raiffa (1968), attributing the idea to John Tukey) called "errors of the third kind: solving the wrong problem." Not always, of course. For example, in a capital punishment case, we might well be interested only in controlling for Type I error against a null of not guilty, as distinct from deciding whether the accused is innocent or guilty. But in most cases, we do care about what the data tells us with regard to the probability of the studied hypothesis. As a practical matter, the $p$-value cannot be expected to be a good guide for this probability.

**Funding:** This research received no external funding.

**Acknowledgments:** Elements of the arguments made here are drawn from Startz (2014). Helpful suggestions from the academic editor and referees are gratefully acknowledged.

**Conflicts of Interest:** The author declares no conflict of interest.

## References

Berger, James O., and Thomas Sellke. 1987. Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence (with comments). *Journal of the American Statistical Association* 82: 112–39. [CrossRef]

Berkson, Joseph. 1942. Tests of Significance Considered as Evidence. *Journal of the American Statistical Association* 37: 325–35. [CrossRef]

Dickey, James M. 1977. Is the Tail Area Useful as an Approximate Bayes Factor? *Journal of the American Statistical Association* 72: 138–42. [CrossRef]

Hubbard, Raymond, and María Jesús Bayarri. 2003. Confusion over Measures of Evidence ($p$'s) versus Errors ($\alpha$'s) in Classical Statistical Testing. *American Statistician* 57: 171–82. Comments by K. N. Berk and M. A. Carlton, and Rejoinder. [CrossRef]

Hubbard, Raymond, and R. Murray Lindsay. 2008. Why *P*-Values Are Not a Useful Measure of Evidence in Statistical Significance Testing. *Theory & Psychology* 18: 69–88.

Kim, Jae H., and Philip Inyeob Ji. 2015. Significance testing in empirical finance: A critical review and assessment. *Journal of Empirical Finance* 34: 1–14. [CrossRef]

Nickerson, Raymond S. 2000. Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods* 5: 241–301. [CrossRef] [PubMed]

Pearson, Egon S. 1938. "Student" as a statistician. *Biometrika* 30: 210–50. [CrossRef]

Raiffa, Howard. 1968. *Decision Analysis: Introductory Lectures on Choices under Uncertainty*. Reading: Addison-Wesley.

Robinson, Daniel H., and Howard Wainer. 2002. On the Past and Future of Null Hypothesis Significance Testing. *Journal of Wildlife Management* 66: 262–71. [CrossRef]

Startz, Richard. 2014. Choosing the More Likely Hypothesis. *Foundations and Trends in Econometrics* 7: 120–86. [CrossRef]

Trafimow, David. 2003. Hypothesis Testing and Theory Evaluation at the Boundaries: Surprising Insights From Bayes's Theorem. *Psychological Review* 110: 526–35. [CrossRef] [PubMed]

Trafimow, David. 2005. The ubiquitous Laplacian assumption: Reply to Lee and Wagenmakers. *Psychological Review* 112: 669–74. [CrossRef]

Trafimow, David. 2015. The benefits of applying Bayes' theorem in medicine. *American Research Journal of Humanities and Social Sciences* 1: 14–23.

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

Trafimow, David, and Stephen Rice. 2009. A test of the NHSTP correlation argument. *Journal of General Psychology* 136: 261–69. [CrossRef] [PubMed]

Wasserstein, Ronald L., and Nicole A. Lazar. 2016. The ASA's Statement on $p$-Values: Context, Process, and Purpose. *The American Statistician* 70: 129–33. [CrossRef]

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
Tel. +41 61 683 77 34
Fax +41 61 302 89 18
www.mdpi.com

*Econometrics* Editorial Office
E-mail: econometrics@mdpi.com
www.mdpi.com/journal/econometrics

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland

Tel: +41 61 683 77 34
Fax: +41 61 302 89 18

www.mdpi.com

ISBN 978-3-0365-0883-2