Journal of Empirical Finance 34 (2015) 1–14

Contents lists available at ScienceDirect

Journal of Empirical Finance

journal homepage: www.elsevier.com/locate/jempfin

Significance testing in empirical finance: A critical review and assessment☆

Jae H. Kim[a], Philip Inyeob Ji[b,*]

[a] Department of Economics and Finance, La Trobe University, Bundoora, VIC 3086, Australia
[b] Department of Economics, Dongguk University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 9 June 2015
Received in revised form 5 August 2015
Accepted 6 August 2015
Available online 28 August 2015

JEL classification:
C12
G12

Keywords:
Level of significance
Lindley paradox
Massive sample size
Meehl's conjecture
Publication bias
Spurious statistical significance

ABSTRACT

This paper critically reviews the practice of significance testing in modern finance research. Employing a survey of recently published articles in four top-tier finance journals, we find that the conventional significance levels are exclusively used with little consideration of the key factors such as the sample size, power of the test, and expected losses. We also find that statistically significant results reported in many surveyed papers become questionable, if Bayesian method or revised standards for evidence were instead used. We observe strong evidence of publication bias in favour of statistical significance. We propose that substantial changes be made to the current practice of significance testing in finance research, in order to improve research credibility and integrity.

1. Introduction

Significance testing is widely and extensively conducted in finance research. It is an essential tool for establishing statistical evidence of an association or relationship between financial variables of interest. In academic research, significance testing plays an important role in testing the empirical validity of financial theories or hypotheses, both new and old. In business and government, outcomes of significance testing aid the key stakeholders in their corporate and policy decisions. Hence, the way in which significance testing is conducted has a huge bearing on knowledge advancement and social welfare. For example, controversy has arisen that the

☆ We would like to thank Sandra Andraszewicz, Geoff Cumming, Steve Easton, Paul Embrechts, Tom Engsted, Raymond Hubbard, Edward Leamer, Edward Podolski-Boczar, Benjamin Scheibehenne, Enrico Schumann, Abul Shamsuddin, Mervyn Silvapaulle, Tom Stanley, Eric-Jan Wangenmakers, Mark Wohar, and Xiangkang Yin for their constructive comments. An earlier version of the paper was presented at the University of Newcastle, Australia; and at the World Finance Conference (Venice, Italy) in 2014.

* Corresponding author.
E-mail address: philipji0422@dongguk.edu (P.I. Ji).

http://dx.doi.org/10.1016/j.jempfin.2015.08.006
0927-5398/© 2015 Elsevier B.V. All rights reserved.

recent global financial crisis is partly attributable to rating agencies which incorrectly over-rated sub-prime mortgages, based on deficient mathematical models or faulty statistical models with inadequate historical data (Donnelly and Embrechts, 2010).[1] Since significance testing is a building block for statistical or mathematical models, we should carefully conduct it, with mindful regard to the potential consequences of making incorrect decisions.

For years, abuse and misuse of significance testing have been a subject of criticism in many disciplines of science (see Morrison and Henkel, 1970; Ziliak and McCloskey, 2008, p. 57). In medical research, Ioannidis (2005) expresses concerns that most current published findings are false, partly because researchers merely chase statistical significance. From psychology, Cumming (2013) calls for substantial changes in the way that statistical research and significance testing are being conducted. Keuzenkamp and Magnus (1995) and McCloskey and Ziliak (1996) critically review the practice of significance testing in applied economics. Their main criticisms include: (i) arbitrary choice of the level of significance; (ii) little consideration of the power (or Type II error) of test; (iii) confusion between statistical and substantive importance (economic significance); and (iv) the practice of "sign econometrics" and "asterisk econometrics" with little attention paid to effect size. Despite these continuing criticisms, it appears that the practice of significance testing has not improved noticeably. For example, in their updated survey of articles published in the *American Economic Review*, Ziliak and McCloskey (2004) report little improvement has been made since the publication of their earlier survey in 1996, although Hoover and Siegler (2008) and Engsted (2009) strongly criticize and refute this claim. Recently, Ioannidis and Doucouliagos (2013) question the credibility of empirical research in economics and business studies, and discuss a range of key parameters affecting the credibility including sample size, effect size, and replicability.

In empirical finance to date, the practice of significance testing has not been given the proper attention that it deserves.[2] The purpose of this paper is to fill this gap. We conduct a survey of recently published papers in top-tier finance journals to shed light on the current practice of significance testing in finance. We have identified the following salient features. First, the use of large or massive sample size is prevalent. As Neal (1987) points out, there is a danger that statistical significance can be over-stated or even spurious in this case. Second, the conventional levels of significance (1%, 5%, and 10%) are exclusively used, with little consideration given to the key factors such as sample size, power of the test, or expected losses from incorrect decisions. There is no reason to believe that these conventional levels are optimal or should be preferred. Third, we find clear evidence of publication bias in favour of statistical significance, where the proportion of the studies with statistical significance is unreasonably high.

In finance, noting the effect of large sample size on significance testing, Connolly (1989, 1991) proposes using the Bayesian method of hypothesis testing previously developed by Leamer (1978) and Zellner and Siow (1979). However, it has been largely ignored in modern finance research. Specifically, Leamer (1978) recommends that the level of significance be adjusted as a decreasing function of sample size, which is not generally followed in finance (and neither in other areas). Recently, from a survey of the studies published in psychological journals, Johnson (2013) argues that the level of significance be set at 0.001 or 0.005 as a revised standard for evidence by reconciling the Bayesian and classical methods. In this paper, we find that the outcomes of significance testing reported in many studies included in our survey are reversed if the Bayesian alternatives were instead used or a much lower level of significance than the conventional ones was adopted as revised standard for evidence.

We also discuss the methods of selecting the optimal level of significance by explicitly considering sample size, power (or Type II error) or expected losses from incorrect decisions, following Leamer (1978). We provide Monte Carlo evidence and an empirical application that the optimal choice of level of significance based on Leamer's (1978) line of enlightened judgement can provide substantially different inferential outcomes from those based on conventional levels. With a large or massive sample size, we recommend using the Bayesian method of Zellner and Siow (1979) or setting the level of significance at a much lower level following Johnson (2013). Conversely, when the small sample size is small and the power is low, the level of significance should be set at a level much higher than conventional ones, for a sensible balance between Type I and II errors. We also find that, while the use of robust standard error estimators is widespread in finance research, little effort is being made to identify the error structure and conduct a more efficient estimation of effect size for improved performance of significance testing.

We conclude that finance researchers should rethink the way they conduct significance testing in their statistical research. In particular, mindless use of the conventional level of significance should be avoided. As mentioned earlier, grave concerns have been raised about the credibility of published studies in many fields of science. With big data sets becoming more and more accessible to finance researchers, a new method of significance testing should be in place, in order to maintain the research credibility in finance research. Ioannidis and Doucouliagos (2013) provide a detailed discussion on how research credibility can be improved; while Ellis (2010) and Cumming (2013) propose guidelines for improved statistical research, which are highly suggestive to empirical finance. The rest of the paper is organized as follows. In the next section, we discuss the background of statistical methods, including the Bayesian methods of hypothesis testing and the methods of choosing the optimal level of significance. Section 3 provides the details and a summary of our survey. In Section 4, we present the analysis of our survey results, with additional Monte Carlo evidence and an empirical example. Section 5 presents further discussions, and Section 6 concludes the paper.

[1] See also p. 32 of Financial Stability Forum Report available at http://www.financialstabilityboard.org/publications/r_0804.pdf.

[2] Petersen (2009) reports a survey on the use of robust standard error estimators in finance; and Engsted (2009) provides a selected review of the studies in empirical finance in relation to discussion of economic significance, in response to the criticisms made by Ziliak and McCloskey (2004).