American Journal of Economics and Business Innovation (AJEBI)

ISSN: 2831-5588 (ONLINE), 2832-4862 (PRINT)

VOLUME 3 ISSUE 1 (2024)

PUBLISHED BY

E-PALLI PUBLISHERS, DELAWARE, USA

American Journal of Economics and Business Innovation (AJEBI)

Volume 3 Issue 1, Year 2024
ISSN: 2831-5588 (Online), 2832-4862 (Print)
DOI: https://doi.org/10.54536/ajebi.v3i1.2349
https://journals.e-palli.com/home/index.php/ajebi

The Scientology of Hypothesis Testing in Empirical Research: Emphasizing Economic Significance

Moses Mayondi[1*], Richard Mulenga[2]

Article Information

Received: December 19, 2023

Accepted: January 24, 2024

Published: January 27, 2024

Keywords

Scientology, Bayesian, Hypothesis Testing, Economic Significance, Statistical Significance

ABSTRACT

Over the decades, scientists across various disciplines have cautioned against the practice of over-emphasizing statistical significance at the expense of economic or practical significance. It seems that empirical researchers in the 21st century have not heeded the caution because meeting statistical significance targets continue to take precedence over the wider discovery and publication of scientifically objective findings. Statistically insignificant results are rarely reported, in part due to publication bias towards statistically significant findings. The survey finds that although large sample sizes are widely used in empirical economics and finance research, none of the surveyed papers adopted alternative methods of hypothesis testing, such as Bayesian methods. All of them explicitly or implicitly used the classical Fisher's hypothesis testing methods. This study finds that discussions on economic significance in nearly all papers (almost 97% of papers) only wrote one sentence or two regarding the magnitude of the effect of the regressors and a declaration that the findings were economically significant. We recommend that to enhance research credibility, other methods of hypothesis testing such as Bayesian methods, should be adopted. Journal article publishers should encourage the publication of statistically insignificant empirical findings. Economic or practical significance should be emphasized and comprehensively discussed.

INTRODUCTION

This study examines the extant scientology or 'cult' of classical hypothesis testing, popularly called statistical significance in empirical research conducted in social sciences in general, and in economics and finance in particular. Following the seminal publication of hypothesis testing by Sir R.A Fisher in 1925 and its subsequent widespread, extensive use in empirical literature over the years, there has been intense scholarly debate regarding the misapplication and misuse of statistical significance at the expense of the wider scientific objectivity or economic feasibility (See, for example, Bakan, 1966; Leamer, 1978; Ziliak and McCloskey, 2008; Cohen,1990; Cumming; 2013; Terry *et al.*, 2022; Ohlson, 2023). The American Statistical Association (2016) contended that statistical significance does not measure the size of an impact or the practical (or economic) importance of empirical research findings (Wasserstein, 2016). Ziliak and McCloskey (2009:2303) succinctly explain that the overuse and/or misapplication of statistical significance in empirical research is a "diversion from the proper objects of scientific study. Fit is not the same thing as importance, and statistical significance is not the same thing as economic or scientific sense." However, the scholars who defend the use of statistical significance in empirical studies, and largely to the total exclusion of economic significance, contest that focusing on statistical significance targets is a measurable scientific benchmark that is not only a focal point of academic performance but is also tied to academic and managerial career performance in terms of tenure and promotions (Peden & Sprenger, 2021; Mitton, 2023). Critics of statistical significance assert that the publication bias of academic journals towards the publication of statistically significant papers reinforces the frequent abuse and arbitrary application of significance testing (Ioannidis, 2005; Kim & Ji, 2015; Mitton, 2023; Ohlson, 2023). Proponents of significance testing cannot be blamed in totality because their careers, including promotions and tenure, are tied to the publication of the popular 'statistically significant' research papers (Keuzenkamp and Magnus in 1995; Mitton, 2023). The corollary argument is, like academic empirical researchers; corporate managers oftentimes also become overly concerned with the tenure and career progression tied to meeting periodic financial targets such as earning targets, and consequently, they consciously sacrifice economic value (Graham *et al.*, 2005; Ziliak & McCloskey,1996; Ohlson, 2023). However, the arbitrary application and misuse of statistical significance have huge implications in that this unprofessional practice in empirical studies often leads to poor decision-making in governments, underperformance in the corporate and business world, causes loss of human lives in the medical fraternity and stifles the objectivity of scientific research. Overall, the arbitrary use and misapplication of statistical significance not only compromise the integrity and credence of the empirical research but also tend to violate research ethics (Ioannidis, 2005; Ziliak & McCloskey, 2009; Kim & Ji, 2015; Ohlson, 2023).

A review of relatively recent related literature revealed that the bulk of extant literature on the Scientology or "Cult" of statistical significance in the empirical literature seems to be skewed towards finance (e.g., Kim & Ji, 2015; Michaelis, 2021), accounting (e.g., Graham

[1] Provincial Planning Unit, Ministry of Finance and National Planning, Solwezi, Zambia
[2] Faculty of Economics, ZCAS University, Dedan Kimathi Road, Box 35243, Lusaka, Zambia
[*] Corresponding author's e-mail: mayondi.moses@gmail.com

et. al, 2005; Mitton, 2023; Ohlson, 2022), Economics and Econometrics (See for example, Terry *et al.*, 2022; Peden & Sprenger, 2021; McCloskey & Ziliak,1996), Mathematics and Statistics (e.g., Ziliak and McCloskey, 2008; Andrews & Kasy, 2019) and comparative studies regarding statistical and economic significance (e.g., Berchicci & King, 2022; McShane *et al.*, 2019; Kewei, et. al., 2019; Engsted, 2009). In addition, given the increasing availability and use of large data sets in financial and economic studies, we need to evaluate the effect of sample size on classical hypothesis testing. So far, it appears that the examination of the scientology of statistical significance in empirical economic and financial research remains an open question. Also, it seems that few empirical studies have taken interest in evaluating the effect of sample size in classical hypothesis testing (See, for example, Ellis, 2010; Cumming; 2013; Kim & Ji, 2015; Ioannidis & Doucoulias, 2013). Therefore, this study seeks to fill this knowledge gap by providing a reality check and suggesting changes that need to be made with respect to significance testing and economic significance in empirical economic and financial studies.

At this point, it is important to outline the difference between statistical and economic significance. The term "statistical significance" describes the process of performing a statistical test on a sample with the goal of identifying any significant departure or deviation from the given null hypothesis (Fisher, 1925, p. 36; American Statistical Association, 2016). After the statistical significance test, a decision is made to reject or not to reject the null hypothesis. The understanding of the differences between statistical and economic significance is vital because some statistical findings may appear noteworthy on paper or in theory but may have little practical bearing on the economy. This can be so especially were the choice of the null and alternative hypotheses are made after the data are seen. To decide on a hypothesis as a result of the data is to introduce a bias into the procedure, invalidating any conclusion that might be drawn from it. This means that while there may be some statistical significance, the deviation from the predicted value may not be necessarily economically feasible . Put differently, the findings may carry statistical weight (that is, findings may be statistically significant and replicable) yet may lack socio-economic impact or relevance (Graham et. al., 2005; Cohen,1990; Terry *et al.*, 2022). We observe that the distinction between statistical and economic significance in the empirical literature (albeit even in conventional economics and econometrics textbooks) is not clearly discussed. Even the literature that seems to critique the mindless use and inappropriate interpretation of statistical significance also seems to fail to give a categorical difference between statistical and economic significance (Sneed, 2016; Rommel & Weltin, 2021). It is not uncommon for researchers to use statistical significance and economic significance interchangeably in empirical research.[4] Figure 1 illustrates conceptually, the difference between economic and statistical significance. Our view is that statistical significance forms a very tiny part of economic or practical significance.

Figure 1 shows that Economic significance encompasses

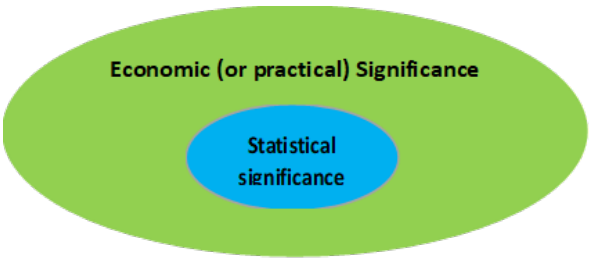

Figure 1: Conceptual differences between Economic and Statistical significance
Source: Adapted from Sneed (2016)

both the statistical and economic effects embedded in the decision arrived at by the researcher after data analysis and testing (Michaelis, 2021). Economic significance and its interpretations go beyond the interpretation of statistical significance in that it considers both the economic theory and practical relevance of the empirical research findings.

Theoretical Underpinnings of Classical Hypothesis Testing

Econometric and economic inference, overall, follow Sir R.A Fisher's theory of evidential approach (Fisher, 1956; Peden & Sprenger, 2021), while the formalisms of decision-making theory are generally linked to the rational choice theory (Reichenbach,1938). Now, both the rational choice theory and Fisher's theory are rooted in the traditions of interpreting statistical significance procedures based on two competing, but contrasting hypotheses popularly referred to as the null, denoted; H_0 and the alternative, denoted; H_1.

According to Sir Fisher (1956), the goal of statistical analysis entails examining the relation of the null hypothesis (H_0) to the metrics of data observations. The convention is that the null represents the absence of causal relationships between the variables. In simpler terms, the null hypothesis according to Sir Fisher, means there is no impact or effect of interest (Cohen, 1994; Peden & Sprenger, 2021). An illustration helps to make this clearer. Assuming that we conducted a simple linear regression analysis modelled as shown in equation 1:

$$y_i=\beta_0+\beta x_i+\varepsilon_i \quad (1)$$

Where the data points x_i and y_i are paired realizations of the variables (or quantities) of interest denoted X and Y, ε_i are error terms that are independently and identically distributed (i.i.d). In this context, a null hypothesis H_0: $\beta=0$, claims that there is no systematic relationship or effect between the variables, X and Y. Conversely, the alternative, H_1: $\beta\neq0$, is a claim that there is a systematic relationship between the variables, X and Y. Conducting the null hypothesis test implies testing for compatibility with data (often called a Null Hypothesis Significance Test, NHST). The rationale for conducting the NHST

is that the results that fall into the extreme tails of the probability distribution as claimed by the null hypothesis compromise its acceptability. That is, either the theory is not true, or an exceptionally rare chance or event has occurred (Fisher, 1956, p39). Fisher argued that the goal of an experiment or empirical research in our context, is to allow the "facts" or give data a chance to disprove the null hypothesis. Fisher and proponents of NHST contend that accepting the null hypothesis does not give conclusive positive evidence for the tested null hypothesis (Fisher, 1956; Kim &Ji, 2015; Peden & Sprenger, 2021).

The p-value is integral to both classical and 21st century statistically significant testing in empirical studies. We illustrate this concept using a two-sided test problem. If we desire to know or to draw an inference whether the mean,μ, of an unknown population distribution is statistically significantly different from the null value H_0: $\mu=\mu_0$ given the observed data, x: $=(x_1,x_2'\ldots,x_N)$ that corresponds to Ni.i.d experiment with an unknown population mean, μ, and a known variance, σ^2. It follows that we can measure the deviation in the observed data, x, with respect to the mean value, μ_0 , by employing the standard test statistic:

$$z(x):=(1/N \sum^{N}_{(i=1)} x_i-\mu_0)/(\sqrt{N}.\sigma^2) \quad (2)$$

In words, equation 2 can be re-written as shown in equation 3:

z=(The observed effect-the Hypothesized effect)/(the standard error) (3)

Equations 2 and 3 imply that given the null hypothesis, the p-value depends on the probability distribution of z expressed as shown in equation 4:

$$p: = pH_0 (|z(X)|\geq z(x)|) \quad (4)$$

The p-value explains the probability of observing a more extreme deviation or discrepancy under the null hypothesis as opposed to the actual observation. Consequently, the lower the p-value, the more the observed effect diverges from the postulated or claims of the null hypothesis. In other words, the lower the p-value the less the likelihood that the null hypothesis explains the observed data or phenomena. These p-values, which represent the observed 'significant levels', are used extensively in empirical research because they act as indicators of whether the results of an empirical study or experiment are 'noteworthy' or statistically significant and thus worth publishing (Ioannidis & Doucoulias, 2013; Kim & Ji, 2015; Peden & Sprenger, 2021). Additionally, it is a standard practice to classify levels of significance and annotate them into correlation tables wherein statistically significant entries are marked with asterisks. It is conventional to set the level of significance at 0.01 (1%), 0.05 (5%) and 0.10 (10%).

The notation p< .01(1%) implies "very highly statistically significant" (three asterisks, ***) ",p< .05 " ("5%"), (two asterisks, **) is "highly statistically significant", and p<.10 (10%) is deemed "statistically significant" (one asterisk, *). It is this arbitrary 'suggestive annotation practice or asterisks econometrics' that has attracted a plethora of criticisms over the years because it not only fails to distinguish the differences between statistical significance and economic or practical significance, but also randomly sets the significance levels without rendering any explanation for the choice (Kim & Ji, 2015; Peden & Sprenger, 2021; Mitton,2022; Ohlson, 2023).

LITERATURE REVIEW

A review of relatively recent related literature revealed that the bulk of extant literature on the Scientology or cult of statistical significance in the empirical literature seems to be skewed towards finance (e.g., Kim & Ji, 2015; Michaelis, 2021), accounting (e.g., Graham et. al, 2005; Mitton, 2023; Ohlson, 2022), Economics and Econometrics (See for example, Terry *et al.*, 2022; Peden & Sprenger, 2021;), mathematics and Statistics (e.g., McCloskey & Ziliak,1996; Ziliak and McCloskey, 2008; Alexander & Kasy, 2018) and comparative studies regarding statistical and economic significance (e.g., Engsted, 2009; McShane *et al.*, 2019; Kewei, et. al., 2019; Berchicci & King, 2022). In addition, given the increasing availability and use of large data sets in empirical financial and economic studies, we need to evaluate the effect of sample size on classical hypothesis testing.

For many decades, the overuse, abuse, misapplication and the arbitrary choice of statistic level and use of statistical significance in empirical research in many fields have been debated. For instance, Cumming (2013), a psychologist, advocates for significant adjustments to the way statistical research and significance tests are carried out. In the field of Medicine, Ioannidis (2005) contended that published empirical research was false because researchers and publishers were biased towards statistically significant research outcomes with less consideration for the practical implications of the research outcomes. Ziliak and McCloskey (1996) and Keuzenkamp and Magnus in 1995 critically reviewed the practice of significant testing in applied economics, also observed the publication bias towards statistically significant empirical research results. Some of the collective criticisms levelled on the classical or Sir Fisher's hypothesis significant statistical testing methods include, among others, the confusion and disregard for economic significance or practical feasibility, and the arbitrary choice of the levels of significance (These are conventionally set to 1%, 5% and 10% respectively) without considering the power of the test, often called Type II error.

Kim & Ji (2015) conducted a survey of 320 published articles in four top-tier finance journals. Their critical reviews found that the conventional statistical significance testing methods were adopted to the exclusion of key factors like sample size and power of the test. They contended further that the credibility of reported results in the published papers in the survey were questionable if revised standards for evidence such as Bayesian methods were applied instead of the Classical hypothesis testing methods.

Mitton (2022) evaluated over 900 regression studies published in seven top finance journals. He found that

although researchers applied varied methods in the empirical finance studies, 93% of the surveyed finance research publications often chose specification methods which favored statistically significant results, with short discussion in a sentence or two declaring that the results were statistically significant, a comment on the size and the impact of the dependent variable and the economic significance of the result.

Mitton (2023) in his study titled; "De-emphasizing Statistical Significance" contends that many researchers of empirical studies devote time and efforts to defend statistical significance of their results using a varied combination of difficult and complex methodological approaches. He advises that a relatively productive approach is to put less emphasis on statistical significance and place "greater emphasis on the economic or practical significance of empirical results. After all, researchers should have a greater interest in the implications of their findings for the real world than detecting statistical significance." (Mitton, 2023, p4).

Rommel and Weltin (2021) reviewed the econometric practice in the American Economic Review (AER) and the American Journal of Agricultural Economics (AJAE) in their quest to conduct a reality check of "A Cult of Statistical Significance in Agricultural Economics." Employing a questionnaire-survey approach, the authors explained that the arbitrary practice of statistical significance testing in both the AER and AJAE is similar and increasing with less focus on the economic significance of empirical findings. Specifically, the power of the tests and discussions on the economic implications of type I or false-positive (that is, the probability of the researcher rejecting a null hypothesis that is true in the population) and type II or false-negative (the probability of the researcher accepting a false null hypothesis) were rarely discussed (only 2% of the respondents indicated discussing the economic significance).

Although a significant body of literature criticizes the practice or abuse of significant testing and promotes the interpretation of empirical findings in light of economic or practical significance, it seems the mindless practice of significant testing in empirical studies has not abated. Moreover, when carrying out empirical studies, researchers use a variety of methodological techniques which tend to influence the magnitude and significance of the empirical results. Critical discourses on the misuse or abuse of classical methods of hypothesis testing methods in empirical economics and finance rarely discuss the mechanism by which non-significant research findings are suppressed (also called file drawer effect) and how to mitigate the file drawer effect. This practice is complemented by p-value hacking (p-hacking). P-hacking refers to manipulating the p-values to outright significant or insignificant values. This also involves practices whereby researchers engage in selective reporting of results, eliminating outliers, adding more regressors with the view to obtaining statistically significant research outcomes (Ziliak & McCloskey, 2004; Kim & Ji, 2015; Michaelides, 2021; Ohlson, 2023).

The review of the related literature seems to suggest that so far, the examination of the scientology of statistical significance in empirical economic and financial research remains an open question. Also, it seems that few empirical studies have taken interest in evaluating the effect of sample size in classical hypothesis testing (See, for example, Ellis, 2010; Cumming; 2013; Kim & Ji, 2015; Ioannidis & Doucoulias, 2013). Therefore, this study seeks to fill this knowledge gap by providing a reality check and suggesting changes that need to be made with respect to significance testing and economic significance in empirical economic and financial studies. We believe these changes will enhance the integrity and credibility of research findings and improve the adherence to research ethics in 21st-century empirical studies, particularly in empirical economic and financial studies.

MATERIAL AND METHODS

In order to evaluate the Scientology or cult of statistical significance in empirical economic and financial research, we follow Rommel and Weltin (2021), Mitton (2022), and Kim & Ji (2015) survey methodological approaches. Our study differs from Rommel and Weltin (2021) in that Rommel and Weltin's (2021) survey focuses on the econometric practice in the American Economic Review (AER) and the American Journal of Agricultural Economics (AJAE) from 2018- 2020. Our study differs also from Mitton's (2022) study, which evaluated studies published only in top finance journals in 2020. Furthermore, this study is different from Kim and Ji (2015) in the sense that while Kim and Ji (2015) conducted a survey of only four (4) top finance journals of empirical finance publications done in 2012, our survey covers both financial and economic journal publications over a period of ten (10) years, from 2011 to 2021.

In addition, given some restrictions that top journals may impose in terms of subscriptions to access the articles which assumably also tends to reduce the number of readers accessing the articles, to circumvent this potential problem, our study surveyed empirical financial and economic journal publications in open-access economic and finance journals from the web-of- Science (WoS), Google Scholar, and Scopus databases. In recent years, many journals have moved away from print to online publication and made publications freely available to readers by adopting the open access policy. These practices are done on the assumption of increasing the number of readers accessing the published papers. Furthermore, whereas Kim and Ji (2015) only included empirical papers that employed linear regression methodological approaches, our inclusion and exclusion criteria excluded purely theoretical articles and included all the empirical financial and economic articles that employed various approaches including linear regressions, cross-sectional, panel and time-series as well as non-linear methods such as Bayesian, Generalized Method of Moments (GMM) and Probit or Logit Methods. To simplify the analysis,

the papers that did not provide relevant information such as the sample size, p-value or t-statistics were excluded from the analysis. Where only p-values were reported, we recalculated the t-statistics by inverting the cumulative distribution function (CDF) quantile functions and/or dividing the estimated coefficients by respective standard errors.

It is nearly 'standard practice' for empirical researchers to report regression results, sub-sample analysis, robust check, or sensitivity analysis. We focused mainly on baseline regression and/or the relatively most representative regressions reported in each paper to avoid surveying or evaluating qualitatively similar empirical studies.

The population of articles was 3, 254 with 1, 439 economics papers and 1, 815 finance papers. After excluding some papers that did not meet our inclusion criteria, the number of articles evaluated in this study reduced to 2, 907 comprising 1, 254 and 1, 653 empirical economics and finance research papers, respectively.

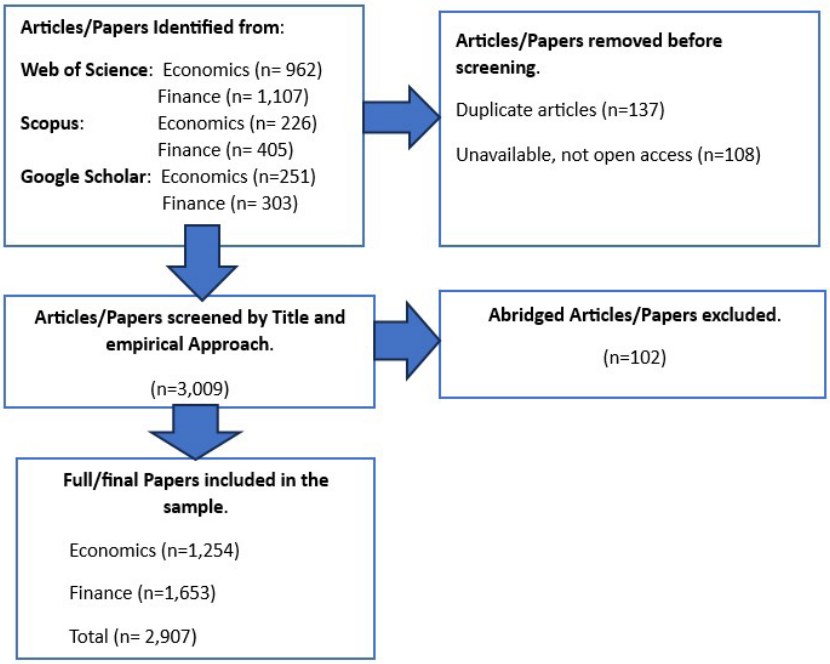

Figure 2: Flow chart of Article or Paper Selection Process

RESULTS AND DISCUSSIONS

The main findings of our survey revealed that the 'standard' or conventional significance levels of 0.01, 0.05 and 0.1 (that is, 1%, 5% and 10%), respectively, were used to the total exclusion of other methods of significance testing. We also observed that a relatively large sample size was extensively used. As can be easily observed in Table 1, the mean sample size hovered around 17,148, the maximum sample size was 33,195, and the minimum sample size was 613, with the median sample size standing at 17,188. In Economics, 56% used mean sample sizes above 11,000, whereas in empirical Finance 63% used mean sample sizes above 15,000. Overall, 76% of empirical researchers in our study used mean sample sizes above 17,000 between 2011 and 2021. All the papers in our sample used the classical null hypothesis testing approach, and only one recommended the use of Leamer's (1978) Bayesian methods. Nine papers briefly discussed economic significance without operationalizing it. The p-values or t-statistics were used by all papers for statistical inference. Only six (6) papers, with five in economics research papers, analyzed or reported confidence intervals. One hundred sixty-five (165) papers, with ninety-three in economics research papers, reported and published statistically insignificant results at 5% significance level. This represented 5.67 % of the sample. Seven papers of which five were in economics, and two were in empirical finance research discussed the potential losses from making incorrect decisions or spurious inferences. Thirteen percent (13%) of the empirical papers surveyed (all of them in economics) reported at least one diagnostic for endogeneity problem, heteroskedasticity and/or autocorrelation. Only three papers in economics reported confidence intervals. This means that there is

Table 1: Summary Statistics

	Mean Sample size	Max. Sample size	Min. Sample size	Median Sample size	Obs.
Economics papers	11,158	21,484	942	17,165	1,254
Finance papers	15,125	33,166	918	17,164	1,653
Both papers	17,148	33,195	613	17,188	2,907

Source: Authors elaboration on data from Google Scholar, Scopus, and Web-of-Science

widespread neglect of the effect of sample size among the papers in our sample. Critical discourses on the misuse or abuse of classical methods of hypothesis testing methods rarely discussed the mechanism used to suppress non-significant research findings (file drawer effect) even if they were methodologically sound, and how to mitigate the file drawer effect. This practice complemented by p-hacking (manipulating p-values to outright significant or insignificant values) practices with the view to obtaining statistically significant research outcomes is relatively common.

Figures 3 to 5 show the scatter plots of the economics,

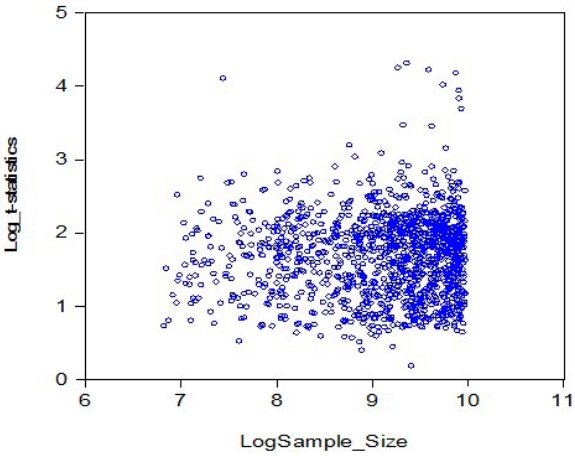

Figure 3: Scatter Plot for Economics sub-sample (n=1,254)

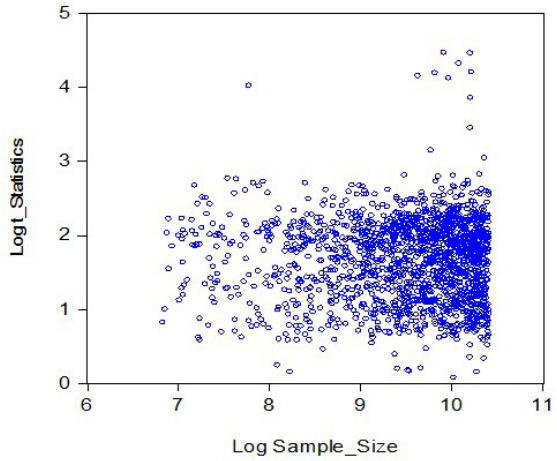

Figure 4: Scatter Plot for Finance sub-sample (n=1,653)

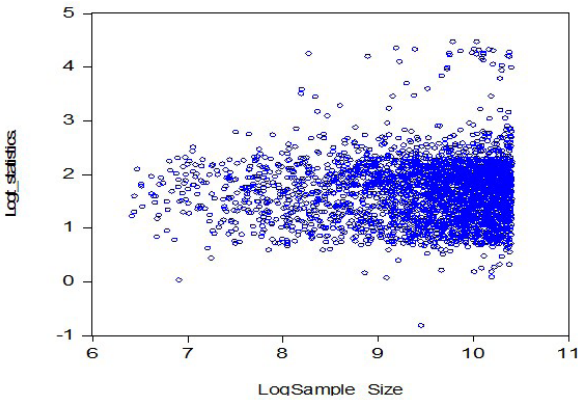

Figure 5: Scatter Plot-Overall Sample Size (n=2,907)

finance and overall sub-samples of t-statistics and their respective sample sizes.

It is easy to observe from the scatter plot in Figures 3 to 5 that the relationship between the natural log values of t-statistics and sample sizes is largely positive. However, in Figure 5, we observe that four (4) studies show a negative relationship. Notice that the variation of t-statistics is larger with increases in sample sizes. We also found that only one hundred sixty-five (165) papers, translating to about 5.67% of 2 907 surveyed papers, reported and published statistically insignificant research findings. This implies that 94.33% reported their empirical research findings with statistical significance. This means that 94.33% of the time the null hypothesis was false for all the empirical economics and finance studies we examined. These findings point to the publication bias exhibited by both article reviewers and publishing editors in favor of statistically significant results. The downside of publication bias in favour of statistically significant research results is that the wider scientific discovery of new and important findings is suppressed, and the over-arching scientific objective of the research is also sacrificed. These broader research values are sacrificed in defense of a narrower goal of statistical significance.

Relationship between Sample Size and Level of Significance

In this section, we briefly discuss ways of choosing or selecting the level of significance that is optimal given the sample size. Empirical literature asserts that sample size is a significant factor in determining the outcomes of significance testing. For example, Kish (1959, reprinted in Morrison & Henkel, 1970: 139) explained that in small samples, significant results may fail to appear statistically significant. However, increasing the sample size makes the statistically insignificant results appear significant. It is "universally" common practice or conventional to set significance levels at 1%, 5% and 10% (0.01, 0.05 and 0.1). The significance level, α, refers to the probability of rejecting the true null. The significance level set to 0.05, according to Sir R.A Fisher (1959), entails that one in twenty is "a reasonable criterion for unusual sampling outcome" (Ji & Kim, 2015, p. 3). However, this reasoning is contested on the grounds that Sir Fisher did not provide scientific reasons for choosing this significance level and that overall, significance levels set at 1%, 5% and 10% are arbitrarily chosen by researchers. Further, critics argue that Sir Fisher's classical theory of hypothesis testing was intended for small sample sizes (See, for example, Lehmann & Romano, 2005, p. 57; Keuzenkamp & Magnus, 1995, p. 20). Thus, it can be argued that, given the increasingly large sample sizes being used in empirical economics and finance studies, classical hypothesis testing is also becoming increasingly inappropriate in empirical research.

The increasing use of large sample sizes (due to the availability of big data sets) in empirical research implies that new or alternative methods to classical hypothesis

testing (or significance testing) should be used to enhance research credibility. Although Ioannidis and Doucouliagos (2013) and Cumming (2013) suggested ways on how to improve research credence and guidelines for empirical statistical research, Leamer in our view (1978) stands out in proposing the most appropriate alternative method of hypothesis testing. Leamer (1978) advised that the significance level should be adjusted as a decreasing function of sample size. Leamer (1978) and Connolly (1991) proposed Bayesian methods as an alternative to classical methods. This is premised on the understanding that, in the context of linear regressions, smaller models are easily rejected in large samples if fixed levels of significance were maintained.

The Bayesian approach of significance testing is rooted in the posterior odds ratio in favor of the alternative hypothesis (H_1) to the null hypothesis (H_0). Following Kim and Ji (2015) notations, the Bayesian method of significance testing is defined as follows:

$$P_{10} \equiv (P(H_1 \mid D))/(P(H_0 \mid D)) = (P(D \mid H_1)P(H_1))/(P(D \mid H_0)P(H_0)), \quad (5)$$

Where, $P(H_i)$ refers to the prior probability for H_i; D is the data; $P(H_i \mid D)$ refers to the posterior probability for H_i; $\beta_{10} \equiv (P(D \mid H_1))/(P(D \mid H_0))$ refers to the Bayes factor. The evidence supports H_1 against H_0 if $P_{10}>1$. Based on P_{10}, Leamer (1978) derived the Bayesian critical value. This critical value increases with sample size. This means that the evidence of significance testing favours the alternative, H_1, when:

$$F>(T-K-1)/P\,(T^{(p|T)}-1), \quad (6)$$

Where, T refers to the sample size and $p \leq k$.

To derive a proper posterior odds ratio that may not favour smaller models, Zeller and Siow (1979) advised the researchers to employ proper diffuse prior, defined as follows:

$$p_{10} = \left[\frac{\pi^{0.5} \qquad (0.5_{V_1})^{K_1}}{\Gamma[0.5(K_1+1)][1+(k_{1/V_1})F]^{0.5(V_1-1)}}\right]^{-1} \quad (7)$$

Where $\Gamma(\,)$ refers to the gamma function, $V_1=T-K_0-K_1-1$ with K_0 and K_1 referring to the numbers of X variables under the null, H_0 and the alternative, H_1, respectively.

Please note that the posterior odds ratios derived under the prior odds of one where $P(H_0)=P(H_1)$ in equation 6, and P_{10} in equation 7 is identical to B_{10}. If $P(H_0) \neq P(H_1)$, Connolly (1991, p.64) advised that the value of P_{10} may be adjusted in accordance with the researcher's prior beliefs.

Reconciling Classical and Bayesian Methods of Hypothesis Testing

Empirical studies have repeatedly demonstrated the conflicting inferential decisions between the Bayesian and the Classical (Fisher's) methods of hypothesis testing (Lindley, 1957; Neal; 1957; Connolly, 1991; Selleke *et al.*, 2001; Johnson, 2013). The conflicting inferential outcomes between Bayesian and Classical hypothesis testing approaches arose from the fact that the classical methods of hypothesis testing were done using fixed levels of significance (usually, 1%, 5% and 10%) whereas the Bayesian methods uses critical values as an increasing function of sample size as demonstrated in equations 6 and 7. By employing the Bayesian methods in conducting the hypothesis of 800 p-values and t-test of 800 psychology papers with respective Bayes factors, Johnson(2013) found that p-value of 0.005(0.5%) and 0.001(0.1%) support or correspond, respectively, with strong and very strong evidence against the null hypothesis, H_0, whereas the p-values in the region of 0.05(5%) and 0.01(1%) indicate modest evidence. Similarly, Selleke *et al* (2001)'s application of the Bayesian p-values found that the p-values in the range of 0.05 seemed not to indicate strong evidence against the null hypothesis. Based on this, Johnson (2013) contended that the standards used for evidence against the null (accepting or rejecting the null) used in the Fisher or Classical method of hypothesis testing were rigid and that the conventional levels (1%, 5% and 10%) were too lenient as a standard for evidence. To reconcile and mitigate the conflicting inferential decisions between the Bayesian and Classical methods of hypothesis testing, Johnson (2013), Selleke *et al* (2002), Gill (2002), for example, advised that the classical method of hypothesis testing should be revised, and the significance levels be scaled down to 0.005 or 0.001 whenever sample sizes exceeded 500 bounds. We are inclined to agree with these reforms in empirical research, particularly in economics and financial research given the increasing access of 21st Century researchers to relatively large databases and frequent extensive use of large sample sizes. None of the 2, 907 empirical researchers in our survey used Bayesian methods. This suggests that in as far as empirical economics and finance studies were concerned, Bayesian methods were seldom employed between 2011 and 2021.

Implications of the Study

The implications of this study are predicated on the findings of this survey. Although a significant body of literature criticizes the abuse of significant testing and promotes the interpretation of empirical findings considering economic or practical significance, it seems the mindless practice of significant testing in empirical studies has not abated. This implies that publication bias of statistically significant results remains a challenge in empirical studies particularly in economics and finance. Empirical literature asserts that sample size is a significant factor in determining the outcomes of significance testing. For example, Kish (1959, reprinted in Morrison & Henkel, 1970: 139) explained that in small samples, significant results may fail to appear statistically significant. However, increasing the sample size makes the statistically insignificant results appear significant. Further, critics argue that Sir Fisher's classical theory of hypothesis testing was intended for small sample sizes (See, for example, Lehmann & Romano, 2005, p. 57; Keuzenkamp & Magnus, 1995, p. 20). The implication is that, given the increasingly large sample sizes being used

in empirical economics and finance studies, classical or Fisher's hypothesis testing methods are correspondingly also becoming increasingly inappropriate in empirical research. Alternative methods such as the Bayesian methods which adjust with increasing sample sizes should be used in empirical studies (Leamer, 1978; Conolly, 1991; Kim & Ji 2015). Publication of statistically insignificant research findings must be encouraged by journal publishers in various fields. This will mitigate the file-drawer effect and p-hacking practices. Adopting the alternative methods of hypothesis testing will also help in harmonizing the theoretical and methodological conflicts that exist between the classical rigid approaches and other flexible methods such as the Bayesian approaches.

CONCLUSIONS

This study examined the extant scientology or cult of classical hypothesis testing in empirical research conducted in economics and finance studies between 2011 and 2021. The cult or scientology of misapplication and abuse of statistical significance at the expense of the wider scientific objectivity or economic feasibility is still prevalent among the 21st Century researchers in empirical economics and finance. Additionally, this study found that statistically insignificant results were rarely reported, in part due to publication bias towards statistically significant findings. Although large sample sizes were widely used in empirical economics and finance research (on average, the sample size in the surveyed papers was above 17,000) due to the increasing availability of massive data sets, none of the surveyed papers adopted alternative methods of hypothesis testing, such as Bayesian methods. Critics of the misuse or abuse of classical methods of hypothesis testing methods rarely discussed the mechanism of suppressing non-significant research findings (file drawer effect) and p-hacking (manipulating p-values to outright significant or insignificant value).

In sum, the cult of Classical hypothesis testing continues to promote a culture of mindless use of statistical inference without due consideration of economic and (practical) implications. Given the foregoing research findings, we recommend the following:

- Researchers in empirical economics and finance should prioritize economic or practical significance over statistical significance because the overarching goal of any empirical research is not to produce statistically significant outcomes but rather to provide economically/practically meaningful results that would effectively inform policy decisions. For example, researchers should select study designs where the parameter estimates are most closely associated with the relevant economic (or practical) magnitude because a theory is supported by significant economic magnitudes rather than by a parameter's statistical significance.
- Given the increasing access to massive data sets and use of large sample sizes in empirical economics and finance studies (mean sample size in our survey is 17,000), we recommend that the classical method hypothesis significance levels should be revised. Whenever the sample size exceeds 500 bounds, the significance levels should be scaled down from the conventional 0.05 to 0.005, and from 0.01 to 0.001 respectively. This practice would reconcile the Bayesian methods with the Classical methods.
- Journal editors and reviewers should encourage the publication of statistically insignificant empirical findings. This would help mitigate the rampant and unethical practice of publishing biased statistically significant empirical results. From the reserachers'side, it is necessary for them to be open to discuss the theoretical farmeworks that guide their work that yielded statistically insignificant results. This could entail creating new theories as necessary to account for unexpected or statistically insignificant findings or relying on pre-existing theories to help explain their results. By doing this, researchers can guarantee, not only the credence of the empirical research, but also that the empirical findings were interpreted and applied in a meaningful and scientific impactful way, in addition to continuing to advance knowledge in their respective fields.
- By and large, the main finding of this study was that economic or practical impact cannot be easily determined from the conventional statistical significance. Going forward, more studies on economic significance in empirical studies should be conducted to foster research integrity in empirical studies in general, and empirical research in economics and finance in particular.

Acknowledgements

The authors wish to thank the two anonymous reviewers for their helpful comments on the earlier versions of this article.

REFERENCES

Andrew I., and Kasy, M (2019). Identification of and Correction for Publication Bias. *American Economic Review 109*(8), 2766–2794 https://doi.org/10.1257/aer.20180310

Bakan, D. (1966). The test of significance in Psychological Research. *Psychological Bulletin, 66*, 423-437

Berchicci, L., and King, A. (2022). Corporate sustainability: A Model uncertainty analysis of materiality. *Journal of Financial Reporting, 7*, 43–74

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

Cohen J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304- 1312.

Cumming, G. (2013). The New Statistics: Why and How, Psychological Science. http://dx.doi.org/10.1177/0956797613504966.

Ellis, P. D. (2010). Effect sizes and the interpretation of research results in international business. *Journal of International Business Studies, 41*, 1581–1588

Engsted, T. (2009). Statistical vs. Economic Significance in economics and econometrics: Further comments on McCloskey and Ziliak. *Journal of Economic*

Methodology,16(4), 393–408.

Fisher, R. (1925). Statistical tests of agreement between observation and hypothesis. *Economica,* (8), 139-147.

Fisher, R. A. (1955). Statistical Methods and Scientific Induction. *Journal of the Royal Statistical Society, Series B (Methodological), 17*(1), 69-78.

Fisher, R. A. (1959). Statistical Methods and Scientific Inference. New York: Hafner. Second edition.

Graham, J., Campbell, H., and Rajgopal, S. (2005). The economic implications of corporate financial reporting. *Journal of Accounting and Economics, 40*, 3–73

Gill J. (2002). Bayesian methods: a social and behavioural sciences approach. London: Chapman Hall/CRC.

Ioannidis, J. P. A. (2005). Why most published research findings are false. PLoS Med. 2, e124.

Ioannidis, J. P. A. and Doucouliagos, C. (2013). What's to know about the credibility of empirical economics? *Journal of Economic. Survey, 27*(5), 997–1004.

Johnson, V. E. (2013). Revised standards for statistical evidence. Proceedings of the National Academy of Sciences http://dx.doi.org/10.1073/pnas.1313476110.

Johnson, V. E. (2013). Revised standards for statistical evidence. Proceedings of the National Academy of Sciences http://dx.doi.org/10.1073/pnas.1313476110.

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

Kewei, H., Chen X., and Lu Z. (2020). Replicating anomalies. *The Review of Financial Studies 33*(5), 2019–2133.

Keuzenkamp, H. A. and Magnus, J. (1995). On tests and significance in econometrics. *Journal of Economics. 67*(1), 103–128.

Leamer, E. (1978). Specification Searches: Ad Hoc Inference with Nonexperimental Data. Wiley, New York.

Lindley, D. V. (1957). *A Statistical Paradox. Biometrika 44*, 187–192.

McShane, B. B., David G., Andrew G., Christian R., and Jennifer L. T. (2019). Abandon Statistical Significance. *American Statistician 73*(S1), 235–45.

Michaelides, M. (2021). Large sample size bias in empirical finance. *Finance Research Letters, 41*, 101835.

Mitton, T. (2022). Methodological variation in empirical corporate finance. *Review of Financial Studies, 35*, 527–575

Mitton, T. (2023). Economic significance in corporate finance. Review of Corporate Finance Studies, forthcoming

Ohlson, J. A. (2022/2023). Empirical accounting seminars: Elephants in the room. *Accounting, Economics, and Law: A convivium,* (0).

Peden, W., & Sprenger, J. (2021). Statistical Significance Testing in Economics in Conrad Heilmann & Julian Reiss (eds.), The Routledge Handbook of the Philosophy of Economics (2021).

Reichenbach, R. (1938). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*, 638-641.

Rommel, J., & Weltin, M. (2021). Is There a Cult of Statistical Significance in Agricultural Economics? *Applied Economic Perspectives and Policy, 43*(3), 1176-1191. https://doi.org/10.1002/aepp.13050

Selleke, T., Bayarri, M. J. and Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *American Statistician, 55*(1), 62–71.

Sneed, D. M. (2016) The Significance of Economic Significance, *Oglethorpe Journal of Undergraduate Research, 6*(1), Article 2. https://digitalcommons.kennesaw.edu/ojur/vol6/iss1/2.

Terry, S. J., Toni M. W., and Anastasia A. Z. (2022). Information versus investment. Technical Report, National Bureau of Economic Research (NBER).

Wasserstein, R. L., and Nicole A. L. (2016). The ASA Statement on p-Values: Context, Process, and Purpose. *American Statistician, 70*(2), 129–33.

Zellner, A. and Siow, A. (1979). Posterior Odds Ratio of Selected Regression Hypotheses. http://dmle.cindoc.csic.es/pdf/TESTOP_1980_31_00_38.pdf

Ziliak, S. and McCloskey, D. (1996). The standard error of regressions. *Journal of Economic Literature, 34*(1), 97-114.

Ziliak, Stephen T., and Deirdre N. McCloskey. (2008a). The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives: Ann Arbor: University of Michigan Press.

Ziliak, S. T., McCloskey, D. N. (2004). Size Matters the standard error of regressions in the American Economic Review. *Journal of Socioeconomics, 33*, 527–546.

Ziliak, S. and McCloskey, D. (996). The standard error of regressions. *Journal of Economic Literature, 34*, 97–114.