(In)appropriate Significance Levels

Michael Michaelides*
†Cyprus University of Technology
Limassol, Cyprus

Abstract

Statistical significance remains the primary standard for evaluating empirical evidence in finance, yet its interpretation depends critically on sample size. Using more than 200,000 regression specifications from leading finance journals, this paper documents that sample sizes have increased substantially and become more dispersed over time, while the thresholds used to assess statistical significance have remained largely unchanged. To examine the implications of this misalignment, a simulation-based frequentist framework is developed to derive sample-size-dependent critical values for t-statistics. The results show that fixed thresholds increasingly classify economically negligible effects as statistically significant as sample size grows, undermining both the reliability of empirical findings and the comparability of evidence across studies. This suggests a role for significance thresholds that vary with sample size, rather than remaining fixed across studies.

Key words: statistical significance; hypothesis testing; significance thresholds; sample-size-dependent inference; t-statistics; replication; false discoveries; empirical finance.

JEL: C12, C15, C18, G10

*Michael Michaelides, Department of Finance, Accounting and Management Science, Cyprus University of Technology, 30 Arch. Kyprianos Str., Limassol 3036, Cyprus. Email: michael.michaelidis@cut.ac.cy

1 Introduction

A central concern in empirical finance, as in many research fields, is the reliability of statistical evidence and the prevalence of findings that fail to replicate. This issue has been extensively documented in the broader scientific literature (e.g., Ioannidis, 2005; Open Science Collaboration, 2015) and has received considerable attention in finance (e.g., Harvey, Liu, & Zhu, 2016; Harvey, 2017; Hou, Xue, & Zhang, 2020). One recognized source of the problem is the mechanical application of conventional thresholds for statistical significance without regard to context, typically by declaring results significant whenever p-values fall below 0.05 or absolute t-statistics exceed 1.96. The American Statistical Association (ASA), in its statement on statistical significance (Wasserstein & Lazar, 2016), warns against treating the 0.05 threshold as a universal standard for scientific conclusions. Among the concerns emphasized by the ASA is the dependence of statistical significance on sample size: "[a]ny effect, no matter how tiny, can produce a small p-value if the sample size ... is high enough" (p. 132). This paper examines the implications of this dependence for statistical inference and develops an alternative approach to significance thresholds for empirical finance.

The motivation for the paper comes from the evidence presented in Section 2, which surveys more than 200,000 regression specifications from leading finance journals. The survey documents three main findings: (i) regression sample sizes have increased dramatically over time; (ii) they have become increasingly heterogeneous, both across and within studies; and (iii) statistical reporting practices have remained largely unchanged, with statistical significance continuing to be assessed almost exclusively using conventional thresholds. Taken together, these findings suggest that uniform significance thresholds may no longer provide an appropriate basis for statistical inference in empirical finance. Because statistical significance depends directly on sample size, identical thresholds can make economically negligible effects appear highly significant in sufficiently large samples, while also hindering comparability across studies with substantially different sample sizes. This raises a natural ques-

tion: should the same significance threshold apply regardless of sample size?

This paper argues that the answer is no and proposes that each study should employ significance thresholds that are appropriate for its sample size. Using a simulation-based frequentist framework, the paper derives sample-size-dependent critical values for t-statistics – the workhorse of regression-based inference in empirical finance – under a conservative definition of an economically negligible effect. The resulting critical values become more stringent as sample size increases, preserving the standard of statistical evidence associated with the conventional threshold of 0.05. Consequently, these thresholds mitigate the tendency of economically negligible effects to appear statistically significant in large samples, helping to reduce false discoveries and improve comparability across studies. The proposed approach is straightforward to implement, requires no prior information, and can be readily applied to standard test statistics, with immediate implications for empirical practice, replication, and the interpretation of statistical evidence.

Concerns about the dependence of p-values on sample size date back at least to Berkson (1938), who observed that "extremely large" samples can produce p-values "beyond any usual limit of significance" (p. 526).[1] The issue later received a formal Bayesian treatment in the work of Jeffreys (1939/1961) and Lindley (1957), giving rise to what is now known as the Jeffreys-Lindley paradox. The paradox highlights a fundamental divergence between frequentist and Bayesian approaches to hypothesis testing: as sample size grows, p-values can indicate increasingly strong evidence against the null hypothesis even as Bayesian evidence, measured by Bayes factors, increasingly favors it. Lindley attributed this divergence to the unjustified "practice of keeping the significance level fixed" (p. 190), emphasizing that "5%

[1] More specifically, Berkson noted that "if the number of observations is extremely large – for instance, on the order of 200,000 – the chi-square P [p-value] will be small beyond any usual limit of significance." Berkson's characterization of approximately 200,000 observations as an "extremely large" sample illustrates how dramatically empirical datasets have expanded over time, as a sample size that was extraordinary in 1938 is comparatively modest by contemporary standards, making conventional significance thresholds increasingly problematic in modern empirical settings.

in to-day's [today's] small sample does not mean the same as 5% in to-morrow's [tomorrow's] large one" (p. 189).[2]

Over the years, many prominent statisticians and econometricians have criticized the use of conventional significance thresholds in large samples, arguing that this practice can produce spurious findings of statistical significance (see, e.g., Lehmann, 1958; Arrow, 1960; Leamer, 1978, pp. 105–106; Good, 1982a, 1988; DeGroot, 1986, pp. 449–451, 496–497; Keuzenkamp & Magnus, 1995; Royall, 1997, pp. 68–71).[3] Several of these authors went further by proposing that significance thresholds should become more stringent as sample size increases.[4,5] Despite these longstanding warnings, conventional thresholds remain deeply embedded in empirical practice across finance and most other research fields. In fact, this continued reliance goes far beyond what the original architects of modern hypothesis testing intended. As documented in Section 3, conventional significance thresholds emerged as heuristic guidelines in an era dominated by small-sample experiments and were never intended to function as universal standards applied irrespective of context.

This paper contributes to a cross-disciplinary methodological literature calling for a revision of conventional significance thresholds. This literature is motivated by concerns that conventional thresholds generate excessive false positives and con-

[2]Jeffreys (1939/1961, p. 435) made a related observation, attributing the divergence to the frequentist "omission ... to give the increase of the critical values for large n [sample size]."

[3]For a striking example, see Royall (1997, p. 81), who shows how an economically negligible effect in a very large sample can generate an extremely small p-value of 0.00003, producing a spurious finding of statistical significance under conventional thresholds.

[4]For instance, DeGroot (1986, p. 497) suggested that "a level of significance much smaller than the traditional value of 0.05 or 0.01 is appropriate for a problem with a large sample size," while Leamer (1978, p. 106) stated that "[t]he cure ... is obvious – the significance level must be made a decreasing function of sample size."

[5]Notably, the idea that significance thresholds should become more stringent as sample size increases appears to have been implicitly anticipated by Egon Pearson (one of the architects of modern hypothesis testing) in his discussion of Lindley's (1953) paper (see pp. 68–69). In discussing Lindley's illustrations, Pearson noted that maintaining a fixed significance level across different sample sizes leads to a "rapidly increasing chance of establishing that a difference is significant" and asked whether one might "sacrifice some of this additional power" by lowering the significance level as sample size increased. The idea was later echoed by Lindley (1957), who discussed Pearson's computations to show that aligning frequentist and Bayesian approaches requires adjusting the significance level with sample size.

tribute to poor replicability, as well as by the tension between p-values and Bayes factors highlighted by the Jeffreys-Lindley paradox. In response, the literature has proposed Bayesian calibrations of p-values that align significance thresholds with at least "substantial" Bayesian evidence. Prominent examples include Johnson (2013), who proposed thresholds of 0.005 and 0.001 for "significant" and "highly significant" findings, respectively, and Benjamin et al. (2018)[6], who proposed a 0.005 threshold for claims of new discoveries and recommended that findings falling between 0.005 and 0.05 be treated as "suggestive." The justification for such calibrations is not straightforward, however, for at least two reasons. First, Bayesian calibrations of p-values are not uniquely determined, as the same p-value can correspond to different Bayes factors.[7] Second, the rationale for aligning frequentist significance thresholds with particular levels of Bayesian evidence remains unclear (Lakens et al., 2018).

Rather than relying on Bayesian calibrations of p-values, the present paper derives revised significance thresholds within the frequentist paradigm. More fundamentally, it departs from the assumption, shared by existing proposals, that significance thresholds should be fixed and instead allows them to vary with sample size. The proposed approach draws on the logic of interval hypothesis testing, although it does not formally adopt an interval-testing framework. Classical hypothesis testing is built around point (sharp) null hypotheses that assess whether a parameter equals an exact value, such as whether a regression coefficient is exactly zero. However, point nulls are rarely plausible in empirical applications, as parameters are unlikely to be

[6]Benjamin et al. (2018) was co-authored by 72 researchers from a wide range of fields, including statistics, medicine, biology, economics (among them Guido Imbens), political science, philosophy, psychology, and sociology, reflecting broad interdisciplinary support for the proposed 0.005 threshold.

[7]For instance, Benjamin et al. (2018) evaluate the relationship between p-values and Bayes factors using a two-sided z-test under four distinct specifications of the alternative hypothesis (power, likelihood ratio bound, uniformly most powerful Bayesian test (UMPBT), and local-H_1 bound), and show that a p-value of 0.005 corresponds to Bayes factors between approximately 14 and 26 – a nearly twofold difference attributable solely to the specification of the alternative hypothesis. Johnson (2013), by contrast, relies exclusively on UMPBTs but applies them across several test types (z, t, χ^2, and binomial), reporting Bayes factors between 25 and 50 for the same p-value – a twofold difference attributable to test type alone, and substantially larger Bayes factors than those reported by Benjamin et al. (2018).

exactly equal to the value specified under the null (see, e.g., Meehl, 1967; Berger & Delampady, 1987; Good, 1988; Leamer, 1988; Cohen, 1990; De Long & Lang, 1992). As shown in Section 4, tests of point nulls readily detect even negligible departures from the null in large samples, making rejection effectively inevitable as sample size grows. Interval hypothesis testing overcomes this limitation by replacing the point null with a composite null defined over an equivalence (or indifference) region containing effects too small to matter, thereby shifting the inferential focus from exact to approximate validity (Hodges & Lehmann, 1954; Schuirmann, 1987).

Motivated by this logic, the present paper develops a simulation-based frequentist framework that considers benchmark values within a narrow neighborhood of the null value. These benchmark values represent economically negligible departures from the null and define the bounds of an equivalence-style tolerance region around it. Rather than replacing the point null with a composite null, the framework preserves conventional point-null testing while deriving significance thresholds from its behavior at nearby benchmark values. Implementing the framework requires specifying a tolerance parameter that determines the width of the tolerance region and thereby delineates economically negligible effects. While the choice of an appropriate tolerance level can often be anchored to established measurement scales and regulatory guidance in fields such as medicine (e.g., FDA benchmarks for clinically unimportant effects), it is much more difficult to determine in empirical finance, where researchers frequently study abstract and often latent constructs (e.g., risk preferences or market sentiment rather than blood pressure in mmHg or cholesterol in mg/dl) and where little consensus exists regarding what constitutes an economically negligible effect.

Although the framework accommodates any value of the tolerance parameter, its appropriate choice is inherently application specific, making a universal value difficult to justify. The simulations therefore adopt a tolerance value of 0.005 as a deliberately conservative reference point for the tolerance parameter rather than as a universally applicable definition of an economically negligible effect. This choice

defines an extremely narrow tolerance region around the null, so that only very small departures from it are treated as economically negligible. Because all variables are standardized in the simulation design, a tolerance value of 0.005 corresponds to a 0.005-standard-deviation change in the dependent variable associated with a one-standard-deviation change in the regressor. Relative to widely used standardized effect-size benchmarks (e.g., Cohen, 1988), this represents an extremely small effect.

The simulation results demonstrate that conventional significance testing becomes increasingly prone to rejecting economically negligible departures from the null hypothesis as sample size increases. For samples containing hundreds to a few thousand observations, the rejection probability remains close to the nominal 0.05 level associated with conventional two-sided significance tests. As sample size increases, however, the rejection probability rises rapidly, approaching one in samples containing approximately one million observations. To preserve the 0.05 rejection probability across sample sizes, the critical values must become increasingly stringent, rising from the conventional threshold of 1.96 to approximately 2.5 in samples with tens of thousands of observations, 3.5 in samples with hundreds of thousands of observations, and 7 in samples approaching one million observations. Because the adopted tolerance parameter is deliberately conservative, these critical values should generally be interpreted as conservative benchmarks for most empirical applications. Overall, the results suggest that fixed significance thresholds, conventional or more stringent, do not provide a consistent standard of statistical evidence across sample sizes.

The findings of this paper have implications for several active strands of the finance literature, where statistical significance is central both to the evaluation of empirical evidence and to publication decisions (see Harvey, 2017). First, the paper relates to a broader literature concerned with the reliability of empirical evidence in finance. One of the main concerns in this literature is that researcher degrees of freedom, including specification search (e.g., Mitton, 2022), multiple testing (e.g., Harvey et al., 2016), and related forms of p-hacking (e.g., Harvey, 2017), can inflate the incidence

of statistically significant results. The present paper identifies a complementary but largely overlooked source of inferential fragility arising from the role of sample size in determining statistical significance. As Kim and Ji (2015) note, significance testing in empirical finance is routinely conducted without explicit consideration of sample size. Although larger samples generally improve statistical precision, legitimate differences in sample construction, including sampling frequency, sample periods, or data availability, can substantially alter the probability of obtaining statistically significant results under conventional significance thresholds. As a result, statistical significance may reflect not only the magnitude of the underlying effect but also the sample size on which the estimate is based.

The paper also relates to a growing literature that examines the replicability of published findings as a means of assessing their reliability (e.g., McLean & Pontiff, 2016; Linnainmaa & Roberts, 2018; Hou et al., 2020; Cohn, Liu, & Wardlaw, 2022). A key question in this literature is whether published results remain statistically significant beyond the original study. This paper suggests that discrepancies in statistical significance between original and replication studies may arise not only from differences in the underlying effect but also from differences in sample size. For example, post-publication replications often rely on smaller samples and may therefore fail to achieve statistical significance despite replicating the same underlying effect, whereas extended-sample replications augment the original sample with newly available observations and may generate statistical significance for economically negligible effects simply because of the larger sample. Consequently, replication outcomes may partly reflect differences in sample size rather than differences in the underlying effect, complicating their interpretation.

Moreover, this paper relates to an emerging line of research arguing that conventional significance thresholds are too lenient in empirical finance because they do not adequately control the risk of false discoveries arising from multiple testing (e.g., Harvey et al., 2016; Chordia, Goyal, & Saretto, 2020; Heath et al., 2023; Harvey,

Sancetta, & Zhao, 2026). Accordingly, these studies advocate replacing the conventional threshold with more stringent fixed thresholds, with recommended values corresponding to t-statistics ranging from approximately 2.5 to 4.0. This paper examines whether any fixed significance threshold can provide a consistent standard of evidence across studies with widely differing sample sizes. The results suggest that, in addition to accounting for multiple testing and related concerns, significance thresholds should vary systematically with sample size to maintain a stable evidential standard across empirical studies.

Finally, the paper relates closely to Harvey's (2017) Presidential Address to the American Finance Association (AFA). Harvey proposes a "Bayesianized" p-value, defined as the posterior probability that the null hypothesis is true, obtained by combining a minimum Bayes factor with prior odds. He further applies this technique to answer the question: "[w]hat threshold of t-statistic do I need so that there is only a 5% chance that the null is true?" (p. 1399). The present paper addresses the same question but from a different perspective. In Harvey's framework, a given p-value implies a unique minimum Bayes factor and, for fixed prior odds, a unique posterior probability that the null is true. Thus, the required t-statistic to achieve a specified posterior probability does not depend on sample size. By contrast, this paper derives thresholds that become more stringent as sample size increases, thereby maintaining a consistent evidential standard across empirical studies.

The remainder of the paper proceeds as follows. Section 2 surveys sample sizes and statistical reporting practices in leading finance journals, documenting the rapid growth and increasing heterogeneity of regression samples alongside the continued reliance on conventional significance thresholds. Section 3 reviews the origins of conventional significance thresholds and the historical considerations that shaped their development and use. Section 4 derives the asymptotic relationship between sample size and the t-statistic, showing how large samples can produce statistically significant yet economically negligible results. Section 5 presents arguments for adjust-

ing significance thresholds upward as sample size increases. Section 6 develops the simulation-based framework, presents the simulation results, and discusses their implications for statistical significance testing in large samples. Section 7 concludes the paper.

2 Sample sizes and statistical reporting in finance

This section provides a comprehensive survey of articles published in the *Journal of Finance* (*JF*), the *Journal of Financial Economics* (*JFE*), and the *Review of Financial Studies* (*RFS*) over the 1988–2024 period. The start date was chosen to give all three journals a common sample window following *RFS*'s 1988 launch. After excluding editorials, comments, replies, and other non-research items, the final sample comprised 8,567 articles. Of these, 6,431 reported empirical results from at least one regression specification. A regression specification refers to a single column in a table of regression results, corresponding to one estimated model. The extraction process covered only the main body of each article, so regression specifications reported exclusively in appendix material were not considered.

For each specification within scope, the number of observations used in the regression was extracted from the article's tables and, where a table did not report this number explicitly, from the accompanying discussion in the main text. Extraction was performed using an automated parsing procedure, followed by manual review aimed at correcting errors and filling in missing values. Specifications for which sample sizes could not be reliably inferred by this procedure were excluded. In total, sample sizes were extracted for 215,978 reported regression specifications.

2.1 Sample size evolution

Panel A of Figure 1 traces the evolution of median and mean sample sizes per regression specification by journal and year, shown as solid and dashed lines, respectively.

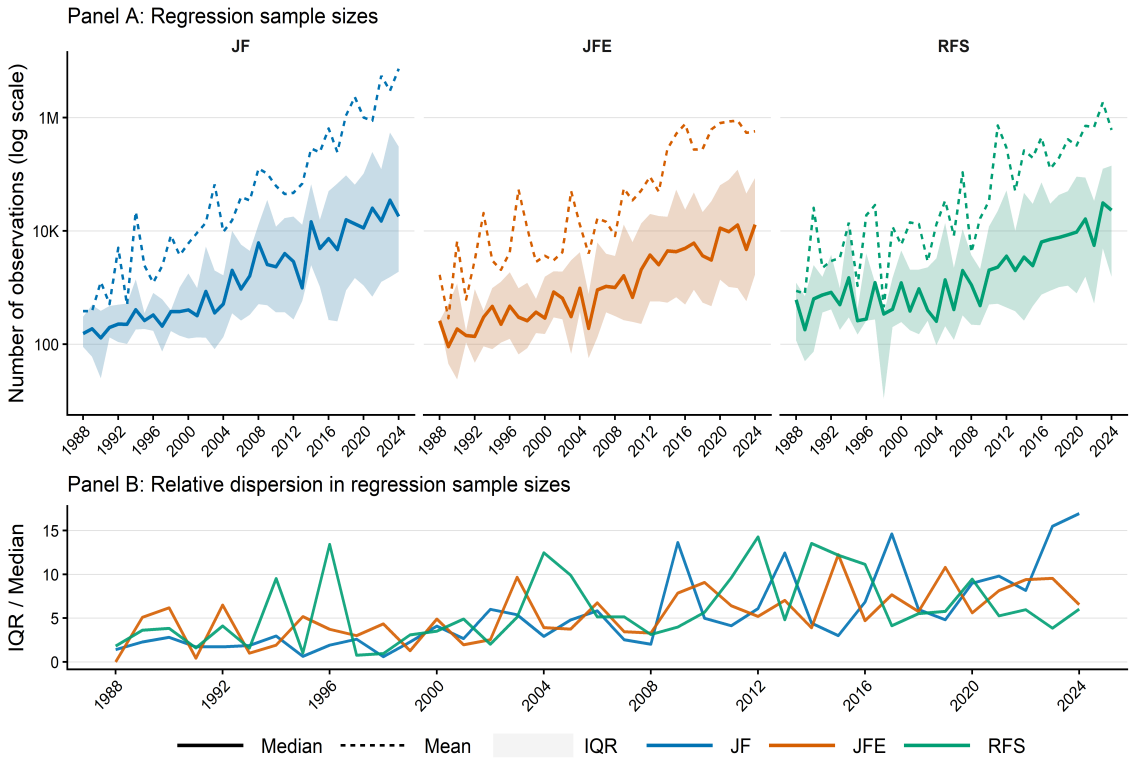

Figure 1: Evolution and dispersion of regression sample sizes. Panel A plots the evolution of regression sample sizes across articles published in the *Journal of Finance* (*JF*), *Journal of Financial Economics* (*JFE*), and *Review of Financial Studies* (*RFS*). Solid lines denote median sample sizes, dashed lines denote mean sample sizes, and shaded areas denote the interquartile range (IQR) of sample sizes. The vertical axis is shown on a logarithmic scale. Panel B plots relative dispersion in regression sample sizes, measured as the ratio of the IQR to the median sample size (IQR / Median). The sample period is 1988 to 2024.

Sample sizes are plotted on a logarithmic scale to accommodate the wide range of values.

Sample sizes increase substantially over time across all three journals, indicating a major expansion in the scale of empirical analysis. The median rises throughout the sample period, which suggests that this expansion is broad-based rather than driven by a small number of very large sample sizes. The rise becomes noticeably steeper in the early 2000s, pointing to a pronounced acceleration in the typical scale of empirical analysis over the past two decades. At the same time, the mean consistently exceeds the median, and the gap between the two widens over time. This pattern points to increasing right-skewness in the distribution of sample sizes, along with a growing prevalence of very large sample sizes.

2.2 Sample size dispersion

Figure 1 also reveals substantial changes in the dispersion of regression sample sizes over time. In Panel A, the shaded areas represent the interquartile range (IQR) of sample sizes. Panel B plots the ratio of the IQR to the median sample size (IQR / Median), which measures dispersion relative to the typical specification size.

Both panels show that dispersion increases markedly over the sample period. Through the early 2000s, the IQR remains comparatively narrow, indicating relatively limited variation in sample sizes across regression specifications. From the early 2000s onward, however, it expands steadily, reflecting a much wider range of sample sizes in more recent work. Relative dispersion follows a similar pattern. In the earlier years, the IQR / Median ratio typically ranges from 0 to 5, whereas after the early 2000s it rises substantially and generally fluctuates between 5 and 10. Taken together, the evidence indicates that regression sample sizes have become increasingly heterogeneous over time, both in absolute terms and relative to the typical specification size.

The extent of heterogeneity in regression sample sizes across the surveyed journals

is particularly striking in recent work.[8] A single volume of *JF*, for example, contains regression specifications whose sample sizes span more than six orders of magnitude: one study reports results based on as few as 20 observations (*JF*, Vol. 75, Issue 1, Table I), while another uses approximately 120 million observations (*JF*, Vol. 75, Issue 2, Table III). Similar contrasts are common throughout recent work. Among the smallest samples encountered are regressions based on 16 observations (*JF*, Vol. 79, Issue 3, Table IV), 24 observations (*JFE*, Vol. 144, Issue 2, Tables 1–2), and 9 observations (*RFS*, Vol. 33, Issue 9, Table 8). At the opposite extreme are specifications estimated using approximately 880 million observations (*JF*, Vol. 77, Issue 3, Table III), 450 million observations (*JFE*, Vol. 137, Issue 1, Tables 2–5), and 288 million observations (*RFS*, Vol. 35, Issue 6, Table 5).

Substantial variation in regression sample sizes is also common within individual studies. One of the papers referenced above (*JF*, Vol. 75, Issue 2) provides a clear illustration: different panels of Table III report specifications based on sample sizes ranging from roughly 50,000 to more than 500,000 and up to 120 million observations, while specifications in Tables IV–V are based on fewer than 5,000 observations. Similar within-study variation appears frequently across the surveyed journals.

2.3 Statistical reporting practices

The survey also reveals a high degree of uniformity in statistical reporting practices across the three journals. Frequentist hypothesis testing overwhelmingly dominates the finance literature. Virtually all empirical studies report estimated coefficients alongside standard errors or t-statistics in parentheses and denote statistical significance using one, two, and three asterisks corresponding to the conventional 0.10, 0.05, and 0.01 significance thresholds, respectively (see also Kim & Ji, 2015). These

[8] The examples provided below are identified by journal, volume, issue, and, where applicable, table numbers, rather than by author names. They are included solely to illustrate empirical patterns in the literature and are not intended as evaluations of individual studies. Moreover, the referenced specifications do not necessarily correspond to a paper's main results and should not be interpreted as implying that any paper's conclusions are incorrect.

conventions appear largely invariant to sample size. For example, both the small- and large-sample studies referenced above rely on the same conventional significance thresholds.

Departures from these reporting practices are relatively uncommon and tend to be idiosyncratic. Some studies omit significance asterisks altogether, although the accompanying discussion often continues to rely on conventional interpretations of statistical significance (e.g., *RFS*, Vol. 33, Issue 3). Other departures involve the adoption of more stringent significance thresholds. For example, some studies supplement the conventional thresholds with additional ones, such as 0.001 (e.g., *JFE*, Vol. 136, Issue 1), while others assess robustness using the stricter threshold proposed by Harvey et al. (2016) (e.g., *JF*, Vol. 76, Issue 2). In such cases, however, the rationale for adopting alternative thresholds is rarely made explicit, and it is often unclear whether the thresholds were chosen before or after the empirical results were observed.

Moreover, alternative reporting measures intended to complement conventional p-values remain exceedingly rare in the surveyed literature. Even the Bayesianized p-value, proposed in Harvey's (2017) widely cited paper on the credibility of statistically significant findings in empirical finance, appears in only five of the surveyed articles (*JF*, Vol. 76, Issue 2; *JFE*, Vol. 133, Issue 2; *JFE*, Vol. 141, Issue 1; *JFE*, Vol. 142, Issue 3; *RFS*, Vol. 33, Issue 10).

Overall, departures from conventional statistical reporting practices remain isolated exceptions rather than signs of a broader shift in empirical reporting conventions within the finance literature.

3 Origins of conventional significance thresholds

The foundations of modern hypothesis testing can be traced back to the work of four key figures: William S. Gosset (publishing under the pseudonym *Student*), Ronald

A. Fisher, Jerzy Neyman, and Egon S. Pearson. Student (1908) laid the groundwork for later developments in statistical inference. He derived the t-distribution, introduced the t-test, and provided the first tables of critical values for different degrees of freedom. Fisher (1925) greatly extended the applicability of the t-test to the testing of regression coefficients and the analysis of variance. He developed the logic of significance testing, formalizing the use of p-values and significance levels as measures of evidence against what would later be called the null hypothesis. Neyman and Pearson (1933) redefined hypothesis testing as a decision-theoretic framework, introducing the concepts of alternative hypotheses, Type I and II errors, statistical power, and the Neyman-Pearson lemma. The lemma established that the most powerful test minimizes the probability of a Type II error while holding the Type I error probability fixed, the latter corresponding to the test's significance level. Their framework emphasized prespecified significance levels and a formal procedure for accepting or rejecting hypotheses.

Modern hypothesis testing has become one of the most widely used quantitative methodologies, becoming central to empirical research across nearly all fields. Yet, despite a century of widespread use, two core aspects of the original frameworks remain frequently misunderstood or neglected in contemporary research. First, the early development of statistical testing methods focused on small samples and was never intended for the very large sample sizes common today. Gosset developed his methods under strict industrial constraints at the Guinness Brewery, where experiments rarely exceeded seven observations (Ziliak, 2011). The small sample sizes he faced motivated his development of the t-distribution, which enabled valid statistical inference when the population variance was unknown and had to be estimated from limited data. Fisher faced similar constraints in his agricultural experiments at Rothamsted Experimental Station, where each additional plot demanded substantial land and labor. For example, his famous Broadbalk experiment on soil fertility comprised only 13 wheat plots (Box, 1976).

Second, conventional significance thresholds arose as practical rules of thumb and were never intended to serve as universal or enduring standards. Fisher introduced the idea of assessing evidence against a null hypothesis using the p-value, proposing that a hypothesis could be rejected when the p-value was sufficiently small. Although he was the first to suggest the use of significance levels, Fisher emphasized these as heuristic guidelines rather than rigid rules, viewing the p-value as a graded measure of evidence. In his first mention of the 0.05 level, he described it as "convenient to take this point as a limit in judging whether a deviation is to be considered significant or not" (Fisher, 1925, p. 44). Throughout his writings, Fisher consistently referred to the 0.05 significance level as "convenient" (see, e.g., Fisher, 1926, p. 504; Fisher, 1935, p. 13), while also acknowledging that alternative levels such as 0.02 or 0.01 might be more appropriate depending on the context (Fisher, 1926, p. 504). He later reinforced this perspective, stating that "the calculation is absurdly academic, for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and ideas" (Fisher, 1956, p. 45).

While significance levels were prescriptive and required prespecification in Neyman and Pearson's (1933) decision-theoretic framework, they, like Fisher, did not endorse rigid or universal thresholds. As they emphasized, "how the balance [between Type I and II errors] should be struck must be left to the investigator" (p. 296). In their illustrative examples, they employed the 0.05 (p. 303) and 0.01 (p. 305) significance levels, but solely "for the purpose of illustration" (p. 305). These choices were clearly influenced by Fisher's earlier tables (see Fisher, 1925, Tables III and IV, pp. 112–113, 176), as Pearson (1962, p. 395) later acknowledged: "[h]is tables of 5% and 1% significance levels ... lent themselves to the idea of choice, in advance of experiment, of the risk of the 'first kind of error' [Type I error] which the experimenter was prepared to take."[9] Both Neyman and Pearson repeatedly stressed that significance

[9]Pearson was even more outspoken in an unpublished letter to Neyman, remarking that "[t]he concept of the control of 1st kind of error [Type I error] would not have come so readily" without

thresholds were not meant to be applied inflexibly. For instance, when discussing Yates's (1935) paper, Neyman noted that although the author applied a fixed 0.02 level, "in practice, however, it is likely that the level of significance will not be fixed so rigidly" (p. 239); see also fn. 5 for Pearson's related view.

4 The limits of conventional significance

Consider the classical linear regression model:

$$y = X\beta + u, \tag{1}$$

where X is the $n \times (k+1)$ design matrix and $\beta := (\beta_0, \beta_1, \ldots, \beta_k)^\top$ is the vector of coefficients. For a given coefficient β_j, $j = 0, 1, \ldots, k$, conventional significance testing is based on the hypotheses:

$$H_0 : \beta_j = \beta_j^{(0)} \qquad \text{versus} \qquad H_A : \beta_j \neq \beta_j^{(0)}, \tag{2}$$

where $\beta_j^{(0)}$ denotes the value of β_j under the null hypothesis, usually taken to be zero. Under the classical homoskedastic linear model, the associated t-statistic is:

$$t_j = \frac{\widehat{\beta}_j - \beta_j^{(0)}}{SE(\widehat{\beta}_j)}, \tag{3}$$

with estimated standard error:

$$SE(\widehat{\beta}_j) = \left\{ \widehat{\sigma}^2 \left[\left(X^\top X \right)^{-1} \right]_{jj} \right\}^{1/2}, \qquad \widehat{\sigma}^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} \widehat{u}_i^2. \tag{4}$$

Under standard regularity conditions, $\widehat{\sigma}^2 \xrightarrow{p} \sigma^2$, and if X has full column rank

Fisher's tables, adding that he and Neyman "must be grateful for those two tables" (see Lehmann, 1993, p. 1244).

and well-behaved second moments:

$$\frac{1}{n}X^\top X \xrightarrow{p} Q, \qquad Q \succ 0. \tag{5}$$

It follows that:

$$\left(X^\top X\right)^{-1} = \frac{1}{n}Q^{-1} + o_p(n^{-1}). \tag{6}$$

Substituting the asymptotic representation in (6) into (4) yields:

$$SE(\widehat{\beta}_j) = \frac{c_j}{\sqrt{n}} + o_p(n^{-1/2}), \qquad c_j := \left[\sigma^2 \left(Q^{-1}\right)_{jj}\right]^{1/2}, \tag{7}$$

where c_j is a positive constant independent of n.

Since $\widehat{\beta}_j \xrightarrow{p} \beta_j$, under any fixed alternative $\beta_j \neq \beta_j^{(0)}$:

$$t_j = \frac{\left(\beta_j - \beta_j^{(0)}\right)\sqrt{n}}{c_j} + O_p(1). \tag{8}$$

Therefore, under any fixed alternative, $|t_j|$ diverges at the rate $\sqrt{n}$. Consequently, as the sample size increases, even arbitrarily small but fixed departures from the null hypothesis eventually exceed any fixed critical value with probability approaching one.[10]

5 Why sample-size-dependent significance thresholds?

Several factors motivate the use of sample-size-dependent significance thresholds in empirical finance. First, and most importantly, empirical finance exhibits substantial heterogeneity in sample sizes, ranging from samples of only a few observations to hun-

[10]The derivation is presented for the classical homoskedastic linear regression model for expositional simplicity. The same argument extends to regular estimators converging at the usual $n^{-1/2}$ rate and to the associated large-sample hypothesis tests.

dreds of millions, yet conventional practice applies a uniform significance threshold across studies. This uniformity is problematic because the same p-value can represent very different evidential content depending on sample size. As Royall (1997, p. 71) argues, "[i]f two experiments that are identical except for their sample sizes produce results with the same p-value, these results do not necessarily represent equally strong evidence against the null hypothesis – the evidence is stronger in the smaller experiment." Significance thresholds should therefore vary with sample size rather than remain fixed uniformly across studies.

Second, despite growing interest in Bayesian methods, the field remains predominantly frequentist. The choice between frameworks may depend on factors such as data quality, computational resources, research objectives, and, most importantly, the availability of reliable prior information and researcher familiarity with each framework. While Bayesian approaches can offer conceptual advantages in certain contexts, their broader adoption is limited by the subjectivity of prior specification (see Royall, 1997, pp. 169–172) and the scarcity of reliable priors in many financial applications.[11] Given these constraints and the strong frequentist tradition in finance, a substantial shift toward Bayesian inference seems unlikely in the near future. Revising significance thresholds within the frequentist framework therefore offers a pragmatic and readily implementable path to improving inferential standards.

Third, unlike experimental settings such as clinical trials, where sample sizes are typically prespecified through formal power calculations and researchers are often required to justify sample-size decisions,[12] empirical finance relies primarily on observational data that are comparatively less costly to obtain and can be continuously updated as new information is recorded. Consequently, sample sizes are determined

[11] These limitations are acknowledged even by committed Bayesians. Good (1982b, p. 342), for example, concedes: "logically there is something wrong with the use of tail-area probabilities [p-values], but I still find them useful because of the well-known difficulties about priors in the Bayesian position."

[12] For example, the CONSORT checklist for reporting randomised controlled trials requires authors to report "[h]ow sample size was determined, including all assumptions supporting the sample size calculation." CONSORT is widely endorsed and often required by leading medical journals.

by data availability, time coverage, and cross-sectional scope rather than by design, and researchers cannot realistically predefine key components of the testing procedure before the final sample is realized. The significance threshold should therefore be determined based on the observed sample rather than fixed in advance, departing from the strict prespecification principle of the Neyman-Pearson framework (Neyman & Pearson, 1933) and its continued emphasis in contemporary methodological discussions (e.g., Lakens et al., 2018).

Fourth, while recent proposals to adopt uniformly stricter thresholds (e.g., Benjamin et al., 2018) aim to reduce false positives, they remain fixed regardless of sample size and therefore do not resolve, and may actually exacerbate, the distortions arising from sample-size heterogeneity. A fixed threshold of 0.005, for instance, can be too conservative in small samples, where limited power makes it difficult to detect economically meaningful effects (Kish, 1959), while remaining insufficiently stringent in very large samples, where economically negligible effects easily attain statistical significance. Such thresholds therefore mitigate the Jeffreys-Lindley paradox only for an intermediate range of sample sizes, arguably introducing a reverse paradox in small samples and failing to address the original paradox in very large samples. Uniform tightening thus does not resolve the fundamental problem that the evidential meaning of a p-value depends on sample size.

Fifth, the absence of any accepted standard for choosing significance thresholds raises concerns for research integrity and transparency. Without such guidance, researchers retain substantial discretion in this choice, creating clear incentives to adjust them after observing results, a practice that is difficult to detect and likely compounded by publication bias (see Kim and Ji, 2015; Harvey, 2017). As documented in Section 2, deviations from conventional thresholds are rare, yet when they occur it is seldom clear whether they reflect deliberate prior choices or post hoc adjustments intended to strengthen the appearance of significance. Clear principles for choosing significance thresholds based on factors such as sample size would leave less room for

this kind of discretionary modification. This would be consistent with the broader shift toward formal statistical reporting standards already adopted in other fields, such as medicine (see, e.g., *New England Journal of Medicine*'s guidelines; Harrington et al., 2019).

Sixth, significance thresholds remain attractive precisely because of their simplicity and practicality. Compared with more radical alternatives, such as abandoning the use of any threshold to dichotomize results as significant or not (e.g., Amrhein, Greenland, & McShane, 2019; Wasserstein, Schirm, & Lazar, 2019; McShane et al., 2019), significance thresholds remain easy to implement and interpret. They provide a clear decision structure valuable for pedagogy, research evaluation (including editorial screening and replication), communication with non-technical audiences, and policy and applied decision-making, where explicit thresholds offer guidance under resource and information constraints. Crucially, these advantages stem from the use of a significance threshold itself, not from fixing it at a conventional value. Adjusting thresholds to account for sample size therefore preserves the practical benefits of conventional testing while eliminating the distortions of a fixed threshold applied uniformly across studies.

6 Simulation framework

This section presents a simulation-based framework to examine how conventional significance testing behaves when benchmark coefficients differ from the population coefficient by an economically negligible amount. The analysis focuses on two related questions. First, how often does conventional two-sided 5% significance testing reject such near-null benchmark coefficients as sample size increases? Second, what critical values are required to maintain a constant two-sided 5% rejection probability for these benchmark coefficients across sample sizes? The results are then used to compare the implied thresholds with existing proposals in the literature and to discuss the

implications of significance testing in massive sample sizes.

6.1 Simulation Design

6.1.1 Data-generating process (DGP)

Let $Z_i = \left(y_i, X_i^\top\right)^\top$ denote a $(1+k)$-dimensional random vector, where y_i is the dependent variable and $X_i = (x_{i1}, x_{i2}, \ldots, x_{ik})^\top$ is the vector of regressors. The sample $\{Z_i\}_{i=1}^n$ consists of independent and identically distributed (i.i.d.) draws from a mean-zero multivariate normal distribution:

$$Z_i \backsim \mathcal{N}_{1+k}\left(0, \Sigma_z(\rho)\right), \qquad i = 1, 2, \ldots, n.$$

The covariance matrix follows a compound symmetry (equicorrelation) structure:

$$\Sigma_z(\rho) = (1-\rho)\, I_{1+k} + \rho J_{1+k},$$

where I_{1+k} denotes the identity matrix and J_{1+k} denotes the matrix of ones. Consequently, all variables are standardized with unit variance, while every pair of distinct variables – both among regressors and between the dependent variable and each regressor – shares the common correlation coefficient ρ.

Positive definiteness of $\Sigma_z(\rho)$ requires $\rho \in (-1/k, 1)$. The simulation restricts attention to nonnegative values of ρ, allowing the correlation structure to vary from independence ($\rho = 0$) to strong positive correlation while remaining within the admissible parameter space.

6.1.2 Data generation

For each design point (n, k, ρ) and replication $r = 1, 2, \ldots, R$, an $n \times (1+k)$ random sample is generated using the Cholesky decomposition of $\Sigma_z(\rho)$.

Let $Z_0^{(r)}$ denote a matrix with i.i.d. rows drawn from $\mathcal{N}_{1+k}(0, I_{1+k})$, and let L

satisfy:

$$\Sigma_z(\rho) = LL^\top.$$

The simulated data matrix is then constructed as:

$$Z^{(r)} = Z_0^{(r)} L^\top.$$

6.1.3 Estimation and t-statistics

For each replication, the linear regression model:

$$y_i^{(r)} = \alpha^{(r)} + X_i^{(r)\top}\beta^{(r)} + \varepsilon_i^{(r)}, \qquad i = 1, 2, ..., n,$$

is estimated by ordinary least squares (OLS). Although the DGP implies $\alpha^* = 0$, an intercept is included to account for finite-sample variation in sample means.

Let $\widehat{\beta}_j^{(r)}$ denote the estimated coefficient on regressor $j = 1, 2, ..., k$, and let $SE(\widehat{\beta}_j^{(r)})$ denote its associated standard error. The conventional t-statistic is defined as:

$$t_j^{(r)} = \frac{\widehat{\beta}_j^{(r)} - \beta_j^*}{SE(\widehat{\beta}_j^{(r)})},$$

where β_j^* denotes the population regression coefficient implied by the DGP.

6.1.4 Population regression coefficients

The population coefficients are defined by the linear projection of y on X:

$$\beta^* = \Sigma_{XX}^{-1}\Sigma_{Xy},$$

where $\Sigma_{XX} = Var(X_i)$ and $\Sigma_{Xy} = Cov(X_i, y_i)$. Under the equicorrelation structure:

$$\Sigma_{XX} = (1-\rho)\, I_k + \rho J_k, \qquad \Sigma_{Xy} = \rho 1_k,$$

where 1_k is a k-dimensional vector of ones. Since the covariance structure is symmetric across regressors, all population coefficients are identical. Using the inverse of the compound-symmetry matrix gives:

$$\beta_j^* = \frac{\rho}{1 + (k-1)\rho}, \qquad j = 1, 2, ..., k.$$

Thus, each regressor has the same population coefficient, and the sampling distribution of the corresponding OLS estimator is identical across regressors. The analysis can therefore focus on a single representative coefficient without loss of generality.

6.1.5 Near-null deviations and shifted t-statistics

Standard inference evaluates the point null hypothesis $H_0 : \beta_j = \beta_j^*$. In many empirical applications, however, exact equality is of limited economic relevance, as sufficiently small deviations from benchmark values may be economically negligible.

To capture this idea, benchmark values of the form $\beta_j^* \pm \delta$ are considered, where $\delta > 0$ is a fixed tolerance parameter. These benchmark values define a narrow neighborhood around the population coefficient within which departures are regarded as economically negligible. The approach follows the logic of interval hypothesis testing by introducing an equivalence-style tolerance region while continuing to evaluate conventional point-null hypotheses centered at nearby benchmark values.

For each replication, the corresponding shifted t-statistics are defined as:

$$t_{j,+\delta}^{(r)} = \frac{\widehat{\beta}_j^{(r)} - (\beta_j^* + \delta)}{SE(\widehat{\beta}_j^{(r)})}, \qquad t_{j,-\delta}^{(r)} = \frac{\widehat{\beta}_j^{(r)} - (\beta_j^* - \delta)}{SE(\widehat{\beta}_j^{(r)})}.$$

6.1.6 Rejection probabilities under near-null deviations

The first quantity of interest is the probability that shifted point-null hypotheses are rejected under conventional significance thresholds. This examines how often standard inference rejects benchmark values that differ from the population coefficient

by only an economically negligible amount.

Under the conventional two-sided 5% testing rule based on the asymptotic critical value of 1.96, rejection probabilities are computed as:

$$\pi_{+\delta} = \frac{1}{R}\sum_{r=1}^{R} 1\left\{\left|t_{j,+\delta}^{(r)}\right| > 1.96\right\}, \qquad \pi_{-\delta} = \frac{1}{R}\sum_{r=1}^{R} 1\left\{\left|t_{j,-\delta}^{(r)}\right| > 1.96\right\},$$

where $1\{\cdot\}$ denotes the indicator function.

Because the benchmark values $\beta_j^* \pm \delta$ are symmetric around the population coefficient and the rejection rule is two-sided, the corresponding rejection probabilities satisfy $\pi_{+\delta} = \pi_{-\delta}$ in theory.

6.1.7 Sample-size-dependent critical values

The second quantity of interest is a set of sample-size-dependent critical values constructed to maintain stable rejection probabilities under near-null deviations. The objective is to adjust rejection thresholds across sample sizes so that economically negligible departures from the population coefficient are not increasingly rejected as statistical precision improves.

Under a two-sided 5% testing rule, the critical values satisfy:

$$\frac{1}{R}\sum_{r=1}^{R} 1\left\{t_{j,+\delta}^{(r)} < c_L\right\} = 0.025, \qquad \frac{1}{R}\sum_{r=1}^{R} 1\left\{t_{j,-\delta}^{(r)} > c_R\right\} = 0.025.$$

Because the benchmark values $\beta_j^* \pm \delta$ are symmetric around the population coefficient and the rejection rule is two-sided, the resulting critical values satisfy $c_R = -c_L$ in theory. The rejection rule can therefore be summarized by symmetric critical values $\pm c(n, k, \rho)$, replacing the conventional asymptotic critical values of ± 1.96.

6.1.8 Choice of the tolerance parameter

As with the choice of an equivalence margin in equivalence testing (see Schuirmann, 1987), the appropriate value of the tolerance parameter is inherently context-dependent

and may vary across applications. To ensure broad applicability, the analysis adopts a deliberately conservative value of $\delta = 0.005$. Since all variables in the DGP are standardized, this value corresponds to a change of 0.005 standard deviations in the dependent variable associated with a one-standard-deviation change in the regressor, a magnitude that would be regarded as economically negligible in most empirical applications.

To place this value in context, it is useful to compare it with established effect-size benchmarks. In the special case $k = 1$, the standardized regression coefficient coincides with the correlation coefficient, providing a natural scale for assessing its economic magnitude. Cohen's (1988) widely adopted taxonomy classifies correlations of 0.10 as "small," a value broadly consistent with the small-effect benchmarks proposed in subsequent work (see, e.g., Gignac and Szodorai, 2016; Funder and Ozer, 2019; Lovakov and Agadullina, 2021). The selected value is approximately twenty times smaller than any of these benchmarks.

Hence, while the choice inevitably involves judgment, the selected value is sufficiently conservative and, if anything, understates rather than overstates economic negligibility.

6.1.9 Simulation grid and replications

The simulation considers the following parameter grid:

$$n \in \left\{10^2, 10^3, 10^4, 10^5, 10^6\right\}, \qquad k \in \{1, 2, 4, 6, 8, 10\}, \qquad \rho \in \{0, 0.1, 0.3, 0.5, 0.7, 0.9\}.$$

For each design point, $R = 10,000$ replications are performed.[13] The primary focus is on the role of sample size n, while k and ρ capture how model dimension-

[13] The random-number generator seed in R is deterministically indexed by the simulation design parameters (n, k, ρ). Specifically, the seed is constructed by concatenating the base value 10, the exponent of the sample size $n \in \left\{10^2, 10^3, 10^4, 10^5, 10^6\right\}$, the number of regressors k, and the correlation parameter ρ. For example, the design point $(n, k, \rho) = (10^4, 6, 0.3)$ generates the seed 104603. This procedure ensures exact reproducibility within the same computational environment and avoids nesting smaller simulated samples within larger ones under a common seed.

ality and dependence in the data influence near-null rejection probabilities and the resulting sample-size-dependent critical values.

6.2 Preliminary checks

Before presenting the simulation results, a series of diagnostic checks are conducted to verify that the simulation design behaves as expected and provides a reliable basis for evaluating rejection probabilities under near-null deviations and for deriving sample-size-dependent critical values.

First, the conventional t-statistic is examined under the point null hypothesis. When the benchmark value coincides with the population coefficient (i.e., when $\delta = 0$), the rejection probability associated with the conventional two-sided 5% testing rule should be close to its nominal level. Accordingly, the proportion of simulated t-statistics exceeding the asymptotic critical values of 1.96 is computed across all (n, k, ρ) combinations. The resulting rejection probabilities are uniformly close to 5%, with any minor deviations attributable to simulation error arising from the finite number of replications.

Second, the unbiasedness and consistency of the OLS estimator are assessed by examining the empirical distribution of the estimated coefficients, $\widehat{\beta}_j$, relative to their population values, β_j^*, across replications. Under the simulation design, the estimator is unbiased, so this distribution should be centered on the population value. It should also become increasingly concentrated around it as the sample size increases, reflecting consistency. The simulation results confirm both properties, as the distribution is centered on the population value and narrows around it as the sample size grows.

Third, for small sample sizes (i.e., $n < 100$), the simulated rejection probabilities and critical values obtained under the near-null deviations are compared with those implied by the theoretical Student's t distribution under the point null. Because sampling noise is greatest in this range, this comparison provides a test of whether the tolerance parameter is negligible, since a substantively large deviation would

still produce detectable departures from the nominal rejection rate. The similarity between the simulated and theoretical values therefore confirms that 0.005 represents an economically negligible effect.

6.3 Simulation results and discussion

6.3.1 Rejection probabilities and sample-size-dependent thresholds

Tables 1 and 2 report the near-null rejection probabilities and the corresponding sample-size-dependent critical values for the left- and right-shifted benchmark coefficients $\beta_j^* \pm \delta$, respectively. The reported rejection probabilities are based on conventional two-sided 5% significance testing using the asymptotic critical value of 1.96, while the adjusted critical values are constructed to preserve the corresponding two-sided 5% near-null rejection probabilities across sample sizes. Results are reported for alternative combinations of sample size (n), number of regressors (k), and the equicorrelation parameter (ρ). The reported results correspond to the first regression coefficient; symmetry of the covariance structure implies identical behavior for the remaining coefficients. Table 1 reports the absolute value of the left-tail critical value, $|c_L|$, alongside the right-tail critical value c_R in Table 2, since symmetry of the benchmark shifts and the testing procedure implies these should be theoretically identical; the small numerical differences between them reflect simulation error arising from the finite number of replications.

The reported rejection probabilities increase monotonically and substantially with sample size. For sample sizes in the hundreds and low thousands, rejection probabilities remain close to the nominal 5% level implied by conventional two-sided significance testing, indicating that benchmark coefficients differing only negligibly from the population coefficient are rarely rejected in smaller sample sizes. As sample size increases into the tens and hundreds of thousands, however, rejection probabilities rise sharply, eventually approaching one at a sample size of one million observations. The results therefore imply that a coefficient differing from its true value by only an

economically negligible amount becomes almost certain to be rejected under conventional 5% significance testing in sufficiently large samples.

The constructed critical values show how significance thresholds must adjust to preserve the nominal 5% near-null rejection probability as sample size increases. In samples containing hundreds or a few thousand observations, these critical values remain close to the conventional asymptotic threshold of 1.96, suggesting that conventional 5% significance testing remains appropriate in smaller sample sizes. Once sample sizes reach the tens of thousands, however, the required threshold rises noticeably, reaching around 2.5, which corresponds to a significance level near 0.01. By one hundred thousand observations it reaches approximately 3.5, near a significance level of 0.0005, and by one million observations it reaches about 7, corresponding to a significance level on the order of 10^{-12}. Thus, maintaining a stable evidentiary standard requires critical values that grow substantially more stringent as sample size increases.

The results also show how model dimensionality and correlation jointly affect rejection probabilities and critical values. Correlation with the dependent variable improves the precision with which a coefficient is estimated, while correlation among regressors reduces it by making individual effects harder to isolate. Panel B isolates the first effect alone by fixing the number of regressors at one, so that no correlation among regressors is possible. This shows that correlation with the dependent variable, left unopposed, substantially increases both rejection probabilities and critical values. This setting is nonetheless of limited empirical relevance, since applications typically involve several potentially correlated regressors. Panel C shows that, absent any correlation, increasing the number of regressors alone has little effect on the results, which remain close to those in Panel A. Panel D shows that once multiple regressors are correlated with one another, the precision gains from correlation with the dependent variable are increasingly offset by the loss of precision caused by correlation among the regressors themselves. This offsetting effect strengthens as the

number of regressors grows, so that rejection probabilities and critical values in Panel D are nearly identical to those in Panels A and C.

6.3.2 Comparison with existing proposals

The simulation results can be compared with recent proposals for more stringent significance thresholds, including lowering the significance level to 0.005 (Benjamin et al., 2018; a two-sided critical value of approximately 2.80), to 0.001 (Johnson, 2013; approximately 3.30), and increasing critical values to between 2.5 and 4 (Harvey et al., 2016; Chordia et al., 2020; Heath et al., 2023; Harvey et al., 2026; two-sided significance levels from roughly 0.01 to 0.00005).

Although these proposals differ in their specific recommendations, each specifies a single fixed threshold intended to apply uniformly across studies. The constructed critical values suggest that this uniformity is difficult to sustain across different sample sizes. For sample sizes in the tens of thousands, the constructed critical values are broadly comparable to existing proposals. Outside this range, however, they diverge sharply in opposite directions. For smaller samples, the constructed critical values remain close to the conventional threshold of 1.96, well below every proposal in the literature, although this value might be too conservative once concerns such as multiple testing are taken into account. For samples of several hundred thousand observations or more, they instead exceed every proposed threshold and continue rising.

6.3.3 Significance testing for massive samples

The simulations are conducted for samples up to one million observations, primarily because of computational constraints. Even at this sample size, which is routinely encountered in modern empirical finance, the derived critical values rise far above conventional thresholds, reaching approximately 7. Extrapolating beyond the simulated range suggests that, in samples of 10 million, 100 million, and one billion observations, maintaining stable rejection probabilities would require critical values

of approximately 18, 52, and 160, corresponding to significance levels of approximately 10^{-72}, 10^{-588}, and 10^{-5560}, respectively. To put this in perspective, with a sample of one billion observations (a scale that is likely become common in empirical finance in the coming years), a 0.000001-standard-deviation change in the dependent variable associated with a one-standard-deviation change in the regressor is sufficient to produce a statistically significant result under conventional significance thresholds.

These results suggest that, although sample-size-dependent significance thresholds are useful over a broad range of sample sizes, maintaining a constant evidentiary standard in extremely large samples would eventually require thresholds so stringent as to be impractical. In such settings, statistical significance alone provides a less informative basis for evaluating empirical evidence and should therefore be interpreted with greater caution. As datasets of this scale become increasingly common in finance, greater emphasis should be placed on complementing p-values with measures that convey additional information about the strength and relevance of the evidence. This recommendation aligns with broader concerns about the interpretation of statistical significance in general (e.g., Harvey, 2017; Imbens, 2021), though the simulation results suggest that it becomes especially important in settings involving massive sample sizes.

The results also suggest a potential role for subsample analyses as a robustness diagnostic in settings involving massive sample sizes. Rather than relying solely on full-sample estimates, researchers could repeatedly estimate the model on smaller, yet still economically informative, subsamples and compare the resulting coefficient estimates with those obtained from the full sample. If the estimated effect remains broadly stable, the analysis could then assess whether it also remains statistically significant despite the substantial reduction in sample size. Subsampling methods (see Politis, Romano, & Wolf, 1999) provide a natural framework for this approach by repeatedly estimating the quantity of interest across subsets of the data. When feasible, such analyses should complement full-sample inference by revealing whether

statistical significance reflects a robust underlying effect or depends primarily on the exceptional statistical precision afforded by massive samples.

7 Conclusion

This paper argues that conventional significance thresholds become increasingly problematic as sample size grows, failing to provide a consistent standard of statistical evidence across studies. Drawing on evidence from more than 200,000 regression specifications published in leading finance journals, it documents substantial growth in empirical sample sizes and variation in sample sizes across studies, alongside continued reliance on conventional significance thresholds. The paper then develops a simulation-based frequentist framework for deriving sample-size-dependent significance thresholds. The simulation results show that conventional significance testing increasingly rejects economically negligible effects as sample size grows, implying that fixed thresholds cannot preserve a stable evidentiary standard across widely differing sample sizes. More broadly, the findings reinforce the view held even by the architects of modern hypothesis testing that the appropriate significance threshold depends on context rather than reflecting a universally applicable standard. As empirical datasets continue to expand, significance thresholds should therefore evolve accordingly and be complemented by richer reporting of economic magnitude and uncertainty, as well as robustness assessments, including analyses based on smaller subsamples where appropriate.

Are the thresholds derived in this paper appropriate? Several considerations are relevant when interpreting them. First, the thresholds depend on the chosen tolerance parameter and should not be regarded as universally applicable. Although the value adopted here is intentionally conservative, some applications may regard effects deemed negligible under this criterion as economically meaningful. In such settings, a smaller tolerance parameter would be warranted, implying, in turn, less

stringent sample-size-dependent thresholds than those reported here. Second, the derived thresholds address only how evidentiary standards should vary with sample size; they do not incorporate other considerations that may be relevant in practice, such as the need to account for multiple testing. Where such concerns arise, the thresholds derived here may be viewed as relatively permissive across sample sizes, and stricter thresholds may be warranted. While reasonable researchers may disagree about the precise thresholds derived in this paper, the broader conclusion remains that fixed significance thresholds, whether conventional or otherwise, fail to provide a stable evidentiary standard across studies with vastly different sample sizes.

Do the proposed thresholds imply that a substantial fraction of empirical results in finance are false? Caution is warranted before drawing such a conclusion. The analysis does suggest, however, that statistically significant results obtained from samples containing hundreds of thousands or millions of observations are increasingly likely to reflect economically negligible effects. Sample-size-dependent thresholds can mitigate this problem by raising the evidentiary standard required for statistical significance. Yet the challenge extends beyond the choice of the threshold. Stricter standards would necessarily increase the incidence of insignificant results and, in the presence of publication bias, may strengthen incentives for p-hacking. Improving empirical practice therefore requires attention not only to how statistical evidence is evaluated, but also to the research and publication practices that determine which findings enter the literature. As Harvey (2017) argues, progress on this front depends on risk-taking by both authors and editors. Reassessing statistical significance to explicitly account for sample size is one step in that direction, as is creating greater space in the literature for carefully executed studies that yield insignificant results.

Table 1. Left-tail near-null rejection probabilities and critical values
The table reports left-tail near-null rejection probabilities (π_L) and sample-size-dependent critical values ($|c_L|$) for the fixed tolerance value $\delta = 0.005$. Rejection probabilities are computed using the conventional two-sided 5% asymptotic critical values ±1.96. The adjusted critical values are calibrated to stabilize the left-tail 2.5% rejection probability across sample sizes.

	π_L					$\lvert c_L \rvert$				
Panel A: $k = 1$ and $\rho = 0$										
$k, \rho \backslash n$	10^2	10^3	10^4	10^5	10^6	10^2	10^3	10^4	10^5	10^6
1, 0	0.052	0.051	0.075	0.352	0.999	2.029	2.068	2.434	3.533	6.967
Panel B: $k = 1$ and varying ρ										
$\rho \backslash n$	10^2	10^3	10^4	10^5	10^6	10^2	10^3	10^4	10^5	10^6
.1	0.055	0.054	0.078	0.359	0.999	2.077	2.098	2.472	3.536	7.003
.3	0.055	0.053	0.083	0.383	0.999	2.043	2.141	2.474	3.643	7.190
.5	0.057	0.052	0.085	0.445	1.000	2.069	2.126	2.514	3.767	7.750
.7	0.048	0.058	0.107	0.595	1.000	2.002	2.203	2.687	4.176	8.975
.9	0.054	0.066	0.205	0.955	1.000	2.077	2.337	3.120	5.584	13.396
Panel C: Varying k and $\rho = 0$										
$k \backslash n$	10^2	10^3	10^4	10^5	10^6	10^2	10^3	10^4	10^5	10^6
2	0.053	0.058	0.077	0.362	0.999	2.026	2.177	2.411	3.590	6.970
4	0.052	0.055	0.078	0.346	0.998	2.010	2.160	2.453	3.524	6.973
6	0.054	0.055	0.078	0.351	0.999	2.011	2.149	2.439	3.569	6.931
8	0.055	0.056	0.080	0.341	0.999	2.033	2.121	2.437	3.571	6.972
10	0.055	0.055	0.079	0.344	0.999	2.056	2.138	2.477	3.563	6.953
Panel D: Varying k and ρ										
$k, \rho \backslash n$	10^2	10^3	10^4	10^5	10^6	10^2	10^3	10^4	10^5	10^6
2, 0.1	0.050	0.055	0.080	0.355	0.999	2.024	2.132	2.482	3.522	6.943
2, 0.3	0.052	0.049	0.080	0.371	0.999	2.023	2.053	2.466	3.567	7.143
2, 0.5	0.051	0.051	0.082	0.379	0.999	2.008	2.123	2.484	3.620	7.297
2, 0.7	0.054	0.054	0.085	0.414	1.000	2.078	2.121	2.497	3.649	7.439
2, 0.9	0.054	0.051	0.092	0.435	1.000	2.024	2.104	2.534	3.766	7.637
4, 0.1	0.055	0.054	0.083	0.355	0.999	2.096	2.098	2.458	3.538	6.954
4, 0.3	0.053	0.054	0.081	0.359	1.000	2.046	2.093	2.477	3.544	7.014
4, 0.5	0.052	0.058	0.078	0.362	0.999	2.064	2.185	2.466	3.562	7.060
4, 0.7	0.050	0.053	0.085	0.370	0.999	2.011	2.112	2.467	3.601	7.064
4, 0.9	0.050	0.049	0.075	0.374	0.999	1.982	2.093	2.426	3.592	7.142
6, 0.1	0.052	0.057	0.078	0.357	0.999	1.987	2.125	2.420	3.534	6.946
6, 0.3	0.056	0.051	0.081	0.366	0.999	2.072	2.091	2.470	3.557	7.008
6, 0.5	0.049	0.053	0.076	0.363	0.999	2.000	2.112	2.475	3.542	7.014
6, 0.7	0.053	0.053	0.084	0.367	0.999	2.018	2.109	2.486	3.538	6.980
6, 0.9	0.051	0.057	0.080	0.345	0.999	1.987	2.130	2.469	3.554	7.028
8, 0.1	0.057	0.051	0.076	0.354	0.999	2.054	2.117	2.473	3.544	6.973
8, 0.3	0.053	0.049	0.078	0.354	0.999	2.031	2.060	2.489	3.536	7.011
8, 0.5	0.053	0.054	0.079	0.351	0.999	2.018	2.126	2.453	3.570	6.935
8, 0.7	0.053	0.052	0.074	0.359	0.999	2.034	2.126	2.426	3.534	6.955
8, 0.9	0.053	0.050	0.080	0.353	0.998	2.027	2.119	2.478	3.523	6.970
10, 0.1	0.052	0.049	0.077	0.352	0.999	1.987	2.071	2.440	3.566	6.905
10, 0.3	0.056	0.053	0.083	0.358	1.000	2.038	2.142	2.462	3.516	7.001
10, 0.5	0.051	0.052	0.077	0.356	0.999	2.003	2.121	2.476	3.526	6.970
10, 0.7	0.055	0.053	0.079	0.358	0.999	2.081	2.103	2.444	3.548	6.981
10, 0.9	0.054	0.048	0.078	0.353	0.999	2.066	2.066	2.494	3.513	7.043

Table 2. Right-tail near-null rejection probabilities and critical values
The table reports right-tail near-null rejection probabilities (π_R) and sample-size-dependent critical values (c_R) for the fixed tolerance value $\delta = 0.005$. Rejection probabilities are computed using the conventional two-sided 5% asymptotic critical values ± 1.96. The adjusted critical values are calibrated to stabilize the right-tail 2.5% rejection probability across sample sizes.

	π_R					c_R				
Panel A: $k = 1$ and $\rho = 0$										
$k, \rho \backslash n$	10^2	10^3	10^4	10^5	10^6	10^2	10^3	10^4	10^5	10^6
$1, 0$	0.053	0.054	0.079	0.352	0.999	2.026	2.157	2.445	3.542	6.952
Panel B: $k = 1$ and varying ρ										
$\rho \backslash n$	10^2	10^3	10^4	10^5	10^6	10^2	10^3	10^4	10^5	10^6
.1	0.056	0.056	0.079	0.352	0.998	2.025	2.154	2.441	3.579	7.015
.3	0.054	0.054	0.086	0.379	0.999	2.043	2.138	2.490	3.606	7.175
.5	0.058	0.052	0.091	0.453	1.000	2.083	2.116	2.530	3.817	7.716
.7	0.048	0.057	0.107	0.606	1.000	2.007	2.218	2.663	4.175	8.961
.9	0.054	0.065	0.209	0.952	1.000	2.099	2.318	3.132	5.546	13.436
Panel C: Varying k and $\rho = 0$										
$k \backslash n$	10^2	10^3	10^4	10^5	10^6	10^2	10^3	10^4	10^5	10^6
2	0.052	0.055	0.079	0.344	0.999	2.030	2.151	2.470	3.519	6.968
4	0.054	0.051	0.079	0.357	0.999	2.052	2.084	2.459	3.514	6.952
6	0.055	0.056	0.078	0.358	0.999	2.086	2.116	2.481	3.565	6.966
8	0.055	0.056	0.077	0.363	0.998	2.060	2.161	2.441	3.564	6.952
10	0.055	0.053	0.079	0.354	0.999	2.048	2.116	2.495	3.532	6.932
Panel D: Varying k and ρ										
$k, \rho \backslash n$	10^2	10^3	10^4	10^5	10^6	10^2	10^3	10^4	10^5	10^6
$2, 0.1$	0.051	0.057	0.080	0.360	0.999	1.991	2.141	2.437	3.547	7.016
$2, 0.3$	0.053	0.052	0.082	0.368	0.999	2.034	2.123	2.455	3.574	7.096
$2, 0.5$	0.052	0.052	0.083	0.393	0.999	2.031	2.126	2.480	3.622	7.247
$2, 0.7$	0.055	0.054	0.088	0.408	1.000	2.047	2.163	2.522	3.718	7.473
$2, 0.9$	0.053	0.055	0.093	0.436	1.000	2.070	2.127	2.558	3.756	7.608
$4, 0.1$	0.056	0.054	0.080	0.354	0.999	2.010	2.158	2.517	3.572	7.010
$4, 0.3$	0.053	0.053	0.081	0.360	0.999	1.999	2.153	2.487	3.582	7.032
$4, 0.5$	0.053	0.055	0.078	0.362	0.999	1.995	2.146	2.491	3.595	7.052
$4, 0.7$	0.049	0.053	0.079	0.368	1.000	1.984	2.130	2.482	3.589	7.086
$4, 0.9$	0.049	0.050	0.078	0.370	0.999	2.030	2.095	2.498	3.586	7.129
$6, 0.1$	0.054	0.054	0.074	0.345	0.999	2.074	2.172	2.436	3.527	6.968
$6, 0.3$	0.056	0.050	0.078	0.350	0.999	2.061	2.113	2.472	3.537	7.027
$6, 0.5$	0.048	0.052	0.080	0.364	0.999	1.984	2.115	2.401	3.538	7.011
$6, 0.7$	0.052	0.053	0.079	0.353	0.999	2.028	2.147	2.440	3.561	6.988
$6, 0.9$	0.052	0.056	0.081	0.370	0.999	2.062	2.164	2.468	3.578	7.018
$8, 0.1$	0.058	0.049	0.080	0.354	0.999	2.083	2.104	2.444	3.556	6.950
$8, 0.3$	0.053	0.049	0.084	0.360	0.999	2.036	2.112	2.470	3.575	6.988
$8, 0.5$	0.051	0.056	0.078	0.358	0.999	2.032	2.127	2.502	3.526	6.996
$8, 0.7$	0.053	0.051	0.074	0.358	0.999	2.045	2.098	2.402	3.543	6.925
$8, 0.9$	0.054	0.051	0.076	0.356	0.999	2.028	2.103	2.418	3.562	7.029
$10, 0.1$	0.051	0.050	0.079	0.359	0.999	2.046	2.119	2.452	3.549	6.989
$10, 0.3$	0.055	0.051	0.082	0.351	0.999	2.059	2.069	2.516	3.575	6.952
$10, 0.5$	0.051	0.057	0.084	0.358	0.998	2.035	2.137	2.467	3.588	6.951
$10, 0.7$	0.056	0.053	0.079	0.356	0.999	2.041	2.124	2.469	3.518	6.965
$10, 0.9$	0.055	0.049	0.078	0.365	0.999	2.036	2.089	2.452	3.566	7.053

References

[1] Amrhein, V., Greenland, S., & McShane, B. (2019). Retire statistical significance. *Nature*, 567(7748), 305–307.

[2] Arrow, K. J. (1960). Decision theory and the choice of a level of significance for the t-test. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (pp. 70–78), Stanford University Press.

[3] Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., ... & Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2(1), 6–10.

[4] Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, 2(3), 317–335.

[5] Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, 33(203), 526–536.

[6] Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356), 791–799.

[7] Chordia, T., Goyal, A., & Saretto, A. (2020). Anomalies and false rejections. *Review of Financial Studies*, 33(5), 2134–2179.

[8] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates.

[9] Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45(12), 1304–1312.

[10] Cohn, J. B., Liu, Z., & Wardlaw, M. I. (2022). Count (and count-like) data in finance. *Journal of Financial Economics*, 146(2), 529–551.

[11] De Long, J. B., & Lang, K. (1992). Are all economic hypotheses false?. *Journal of Political Economy*, 100(6), 1257–1272.

[12] DeGroot, M. H. (1986). *Probability and Statistics* (2nd ed.). Addison-Wesley.

[13] Fisher, R. A. (1925). *Statistical Methods for Research Workers* (14th ed.). In J. H. Bennett (Ed., 1990), *Statistical Methods, Experimental Design, and Scientific Inference: A re-issue of Statistical Methods for Research Workers, the Design of Experiments and Statistical Methods and Scientific Inference*. Oxford University Press.

[14] Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture*, 33, 503–515.

[15] Fisher, R. A. (1935). *The Design of Experiments* (8th ed.). In J. H. Bennett (Ed., 1990), *Statistical Methods, Experimental Design, and Scientific Inference: A re-issue of Statistical Methods for Research Workers, the Design of Experiments and Statistical Methods and Scientific Inference*. Oxford University Press.

[16] Fisher, R. A. (1956). *Statistical Methods and Scientific Inference* (3rd ed.). In J. H. Bennett (Ed., 1990), *Statistical Methods, Experimental Design, and Scientific Inference: A re-issue of Statistical Methods for Research Workers, the Design of Experiments and Statistical Methods and Scientific Inference*. Oxford University Press.

[17] Funder, D. C., & Ozer, D. J. (2019). Corrigendum: Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, 2(2), 156–168.

[18] Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, 102, 74–78.

[19] Good, I. J. (1982a). C140. Standardized tail-area probabilities. *Journal of Statistical Computation and Simulation*, 16(1), 65–66.

[20] Good, I. J. (1982b). Lindley's paradox: Comment. *Journal of the American Statistical Association*, 77(378), 342–344.

[21] Good, I. J. (1988). The interface between statistics and philosophy of science. *Statistical Science*, 3(4), 386–397.

[22] Harrington, D., D'Agostino Sr, R. B., Gatsonis, C., Hogan, J. W., Hunter, D. J., Normand, S. L. T., ... & Hamel, M. B. (2019). New guidelines for statistical reporting in the journal. *New England Journal of Medicine*, 381(3), 285–286.

[23] Harvey, C. R. (2017). Presidential address: The scientific outlook in financial economics. *Journal of Finance*, 72(4), 1399–1440.

[24] Harvey, C. R., Liu, Y., & Zhu, H. (2016). ... and the cross-section of expected returns. *Review of Financial Studies*, 29(1), 5–68.

[25] Harvey, C. R., Sancetta, A., & Zhao, Y. (2026). *What threshold should be applied to tests of factor models?* (NBER Working Paper No. 34898). National Bureau of Economic Research. http://www.nber.org/papers/w34898.

[26] Heath, D., Ringgenberg, M. C., Samadi, M., & Werner, I. M. (2023). Reusing natural experiments. *Journal of Finance*, 78(4), 2329–2364.

[27] Hodges Jr, J. L., & Lehmann, E. L. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society, Series B (Methodological)*, 16(2), 261–268.

[28] Hou, K., Xue, C., & Zhang, L. (2020). Replicating anomalies. *The Review of Financial Studies*, 33(5), 2019–2133.

[29] Imbens, G. W. (2021). Statistical significance, p-values, and the reporting of uncertainty. *Journal of Economic Perspectives*, 35(3), 157–174.

[30] Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.

[31] Jeffreys, H. (1961). *Theory of Probability* (3rd ed.). Oxford University Press.

[32] Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, 110(48), 19313–19317.

[33] Keuzenkamp, H. A., & Magnus, J. R. (1995). On tests and significance in econometrics. *Journal of Econometrics*, 67(1), 5–24.

[34] Kim, J. H., & Ji, P. I. (2015). Significance testing in empirical finance: A critical review and assessment. *Journal of Empirical Finance*, 34, 1–14.

[35] Kish, L. (1959). Some statistical problems in research design. *American Sociological Review*, 24(3), 328–338.

[36] Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., ... & Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, 2(3), 168–171.

[37] Leamer, E. E. (1978). *Specification Searches: Ad Hoc Inference With Nonexperimental Data*. John Wiley & Sons.

[38] Leamer, E. E. (1988). Things that bother me. *Economic Record*, 64(4), 331–335.

[39] Lehmann, E. L. (1958). Significance level and power. *The Annals of Mathematical Statistics*, 29(4), 1167–1176.

[40] Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two?. *Journal of the American Statistical Association*, 88(424), 1242–1249.

[41] Lindley, D. V. (1953). Statistical inference. *Journal of the Royal Statistical Society, Series B (Methodological)*, 15(1), 30–65.

[42] Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44(1/2), 187–192.

[43] Linnainmaa, J. T., & Roberts, M. R. (2018). The history of the cross-section of stock returns. *Review of Financial Studies*, 31(7), 2606–2649.

[44] Lovakov, A., & Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology*, 51(3), 485–504.

[45] McLean, R. D., & Pontiff, J. (2016). Does academic research destroy stock return predictability?. *Journal of Finance*, 71(1), 5–32.

[46] McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, 73(sup1), 235–245.

[47] Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34(2), 103–115.

[48] Mitton, T. (2022). Methodological variation in empirical corporate finance. *Review of Financial Studies*, 35(2), 527–575.

[49] Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, Series A, 231, 289–337.

[50] Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251).

[51] Pearson, E.S. (1962). Some thoughts on statistical inference. *Annals of Mathematical Statistics*, 33(2), 394–403.

[52] Politis, D. N., Romano, J. P., & Wolf, M. (1999). *Subsampling*. Springer.

[53] Royall, R. M. (1997). *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall.

[54] Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6), 657-680.

[55] Student. (1908). The probable error of a mean. Biometrika, 6(1), 1–25.

[56] Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2), 129–133.

[57] Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p$<$ 0.05". *The American Statistician*, 73(sup1), 1–19.

[58] Yates, F. (1935). Complex experiments. *Supplement to the Journal of the Royal Statistical Society*, 2(2), 181–247.

[59] Ziliak, S. T. (2011). W.S. Gosset and some neglected concepts in experimental statistics: Guinnessometrics II. *Journal of Wine Economics*, 6(2), 252–277.