Faculty Scholarship Collection

Large Sample Size Bias in Empirical Finance

Michael Michaelides[†*]
[†]Allegheny College,
Department of Business
and Economics,
Quigley Hall, Box 20,
Meadville, PA 16335, USA.
E-mail address: mmichaelides@allegheny.edu.

Published in *Finance Research Letters*, Volume 41, July 2021, 101835

Submitted: 3 June 2020

Article accepted for publication: 31 October 2020

Abstract

The vast majority of empirical studies in finance employ large enough sample sizes and use the conventional thresholds for statistical significance. This routine practice can potentially lead to spurious statistically significant results. The primary aim of this paper is to present a rule of thumb that can be used to determine the appropriate thresholds for statistical significance for a given sample size. The paper argues that the list of statistically significant findings in the broader finance literature is likely to be much shorter after accounting for large sample size bias.

Key words: large sample size; high statistical power; spurious statistical significance; appropriate significance thresholds; methodological crisis; publication bias.

JEL: C12, C18, G0, G12

*Corresponding author: Michael Michaelides, 520 N. Main Street, Box 20, Department of Business and Economics, Allegheny College, Meadville, PA 16335, USA; phone: +1 (814) 332-3346; e-mail address: mmichaelides@allegheny.edu

1 Introduction

During the past 50 years, the focus of finance research has shifted largely from theoretical to empirical. There are several reasons for this shift, including the increase in the availability of computer power, easy to use analytics software, and accurate comprehensive data. Recently, however, there has been an increasing concern that most published research findings are false (Ioannidis (2005)), primarily because many statistical tools are misused by empirical researchers. Currently, many scientific fields, including financial economics, face a methodological crisis concerning the credibility of their findings (De Prado (2015); Harvey (2017); Christensen and Miguel (2018)).

In the flood of recent finance literature, there has been a growing concern about the methodological crisis. Ferson et al. (2003) and Welch and Goyal (2008) suggested that many regressions and prominent variables from earlier finance research may be spurious due to model misspecification. Kim and Ji (2015) found that the conventional significance levels are exclusively used in empirical finance, and observed strong evidence of publication bias in favor of statistical significance. Harvey et al. (2016) and Linnainmaa and Roberts (2018) argued that the majority of historically discovered factors are most likely an artifact of data mining and p-hacking (see also Black (1993)). Hou et al. (2020) showed that most anomalies fail to replicate, suggesting that there exists a replication crisis in finance research.

The aim of this paper is to bring to attention how sample sizes might interact with spurious empirical findings. In empirical finance, the employed sample sizes are almost always *large enough* (i.e., sample sizes of $n > 100$) to violate the most commonly used level of statistical significance of 0.05 (or t critical value of ± 1.96) suggested by Ronald A. Fisher (Fisher (1925)). In the era of large and massive data sets, it is completely inappropriate to exclusively use the conventional decision

thresholds for declaring statistical significance. This was clearly articulated by Fisher (1956): "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"; and later pointed out by other scholars such as Lehmann (1958), Arrow (1960), Leamer (1978), and Good (1982), who argued that the significance level should be adjusted as a decreasing function of sample size.

Despite widespread and long-standing attention to the issue, it is clear from the recent statements from the American Statistical Association (ASA) that statistical significance is "misunderstood and misused in the broader research community" (Wasserstein and Lazar (2016); Wasserstein et al. (2019)). In similar context, recent statements from the American Finance Association (AFA) affirm that empirical research in finance "relies too much on p-values", and this incentivizes researchers to "produce 'significant' results" and "engage in data mining and p-hacking" (Harvey (2017)).

This paper is closely related to the article of Kim and Ji (2015), where the authors provide an excellent background of significance testing in the context of finance research, and present enlightened Monte-Carlo illustrations and empirical applications. The current paper complements Kim and Ji (2015) by providing a convenient rule of thumb that can be used to determine the appropriate statistical significance thresholds for a given sample size. This rule of thumb can be used to revisit the significance of previously published findings, as well as to provide new thresholds for future research. The paper argues that the misuse of significance thresholds contributes greatly to the ongoing methodological crisis in financial economics. If large sample size bias is taken into account, the list of statistically significant findings in the literature is likely to be much shorter.

2 Large sample size bias

Suppose that we wish to test the hypothesis that a regression coefficient, β_j, for any $j = 0, 1, 2, ..., k$, equals a hypothesized value, say β_{j0}. The appropriate null (H_0) and two-sided alternative hypotheses (H_1) are as follows:

$$H_0 : \beta_j = \beta_{j0} \quad \text{against} \quad H_1 : \beta_j \neq \beta_{j0}. \tag{1}$$

The testing of the above hypotheses is based on the following test statistic (known as the t-statistic):

$$t_{\widehat{\beta}_j} = \frac{\left(\widehat{\beta}_j - \beta_{j0}\right)}{\frac{s_{\widehat{\beta}_j}}{\sqrt{n}}} = \frac{\sqrt{n}\left(\widehat{\beta}_j - \beta_{j0}\right)}{s_{\widehat{\beta}_j}} \overset{H_0}{\sim} \mathbf{St}(n-k-1), \tag{2}$$

where n denotes the number of observations (sample size), $\widehat{\beta}_j$ is the point estimate of the unknown population mean (β_j), β_{j0} denotes the hypothesized value of β_j, $s_{\widehat{\beta}_j}$ is the point estimate of the unknown population standard deviation ($\sigma_{\widehat{\beta}_j}$), and k denotes the number of explanatory variables. The "$\overset{H_0}{\sim} \mathbf{St}(n-k-1)$" reads "under H_0 is distributed as Student's t with $(n-k-1)$ degrees of freedom".

In the general context of regression hypothesis testing, one of the researcher's primary concerns is the attainment of point estimates that are as close as possible to the true population parameters. Hence, it is common practice to employ large enough sample sizes as a direct way of improving the accuracy of point estimates. In general, larger sample sizes decrease the likelihood of a Type II error (probability of failing to reject a false null hypothesis) or, equivalently, increase the statistical power (probability of correctly rejecting a false null hypothesis) of a hypothesis test.

The conventional "large sample size - high statistical power" approach, however, can often lead to misleading results. This is because large sample sizes often lead to artificially large t-statistics of the estimated coefficients; this can be easily seen by recalling that the test statistic in (2) is a monotonically increasing function of the

sample size, n. As a consequence, even if the discrepancy between $\widehat{\beta}_j$ and β_{j0} is tiny, the researcher will almost certainly reject the null hypothesis more frequently, even if the null hypothesis is true.

This issue may seem trivial at first since there is no way to know whether the null hypothesis is true or false prior to conducting the hypothesis test on the estimated regression coefficients. Worse, there is no way to know whether the true null hypothesis is incorrectly rejected posterior the hypothesis test. Yet, this methodological issue stems from the Neyman-Pearson formulation of hypothesis testing (Neyman and Pearson (1928a, b)) in which the two types of error probabilities (Type I and II) are traded off against each other (see also Lehmann and Romano (2005, p. 57)). Therefore, even though larger sample sizes successfully decrease the likelihood of a Type II error (or increase statistical power), they also increase the likelihood of a Type I error (probability of rejecting a true null hypothesis). Type I is, of course, much more serious than Type II because it increases the likelihood of obtaining spurious statistically significant results.

2.1 The "large-high crisis"

There is a widespread belief that the larger the sample size the stronger the evidence for a conclusion. This is because *large* sample sizes ensure *high* statistical power. However, the routine practice of employing large sample sizes without appropriately adjusting the thresholds for statistical significance can potentially increase the risk of spurious statistical significance associated with the elevated Type I error. The misuse of significance thresholds contributes greatly to the ongoing methodological crisis in financial economics.

The acceptable practice of using the conventional significance thresholds, without

having to "penalize" them for the sample size, incentivizes empirical researchers to abuse statistical power by extending sample sizes, using data at a higher frequency, etc. This form of p-hacking is typically done for the purpose of reaching the desired statistical significance (see Harvey (2017)) since statistically significant results are more likely to be published due to positive publication bias (see Harvey et al. (2016); Kim and Ji (2015)). Yet, statistical significance is often spurious; an artifact of large sample size bias. Thus, in the presence of publication bias, large sample size bias leads to the circulation of false-positive findings in the literature.

Moreover, large data sets are inherently complex to use and analyze, which increases the likelihood of model misspecification and sampling bias. Empirical researchers should not forget that models are only approximations of reality (Box (1976)), and hence carry estimation bias, which grows larger with increasing sample size (see Kaplan et al. (2014); for an example, see Harford (2014)). Besides, the complexity of large data sets increases the chances of encountering coding and other clerical errors, which is one cause of the current replication crisis (see Christensen and Miguel (2018); for an example, see Peng (2015)).

3 Deriving a sample size rule of thumb

To avoid large sample size bias, the choice of significance thresholds should take into account the effect of sample size on error probabilities (Type I and II) and statistical power (see also Kim and Ji (2015); Lehmann and Romano (2005, pp. 57-58); Mayo and Spanos (2006); McCloskey and Ziliak (1996)). That being said, this is rarely considered by empirical researchers in finance who routinely use the conventional significance levels of $0.1, 0.05$, and 0.01. A great deal of this "habit", of course, stems from the fact that analytics software, for convenience, use asterisks to denote values

that reached these prespecified levels of significance. Nonetheless, the conventional thresholds for statistical significance should not be used exclusively.

This section derives a rule of thumb to determine the appropriate significance level for a given sample size. As a basis for the sample size "penalization", the following 'standardized p-value' rule of thumb is used (Good (1982)):

$$\widetilde{p} = \min\left(0.5,\ \widehat{p}\sqrt{n/100}\right),\ n \geq 1. \tag{3}$$

This rule of thumb suggests that reported p-values, denoted by $\widehat{p}$, must be standardized to a sample size of 100, by replacing the reported p-values by the standardized by the sample size p-values, denoted as $\widetilde{p}$. For example, assuming a reported p-value is $\widehat{p} = 0.04$, the corresponding standardized by the sample size p-value, for a sample size of $n = 250$, would be $\widetilde{p} = \min(0.5,\ 0.04\sqrt{250/100}) = 0.06324$; leading to statistical insignificance, given the standard significance level of 0.05.

In its current form, the above rule of thumb is not convenient to use for two reasons. First, the researcher should use the estimated p-values to recalculate one by one the standardized p-values, which can be troublesome. Second, this rule of thumb does not allow to determine, prior the analysis, a single level of significance for the sample size in hand. Thus, the rule of thumb in (3) is modified here in a way as to calculate an appropriate significance level rather than standardized p-values.

Under the assumption that the appropriate level of significance for a sample size of $n = 100$ is $\alpha_{100} = 0.05$, the standardized by the sample size p-value ($\widetilde{p}$) is replaced by 0.05, and the reported p-value ($\widehat{p}$) is defined as the appropriate significance level for a sample size of n (α_n):

$$\longrightarrow \quad \alpha_n \quad = \quad \min\left(0.5,\ \frac{0.05}{\sqrt{n/100}}\right),\ n \geq 1. \tag{4}$$

The rule of thumb in (3) now takes the form of a more generalized rule of thumb:

$$\alpha_n \quad = \quad \min\left(0.5,\ \tfrac{0.5}{\sqrt{n}}\right),\ n \geq 1, \tag{5}$$

where the appropriate significance level is a decreasing function of the sample size. This rule of thumb allows one to calculate the appropriate significance level for any sample size. Table 1 reports significance levels and the corresponding t critical values for different sample sizes.

sample size	α_n	τ_n	sample size	α_n	τ_n
$n = 25$	0.1	1.70814	$n = 10,000$	0.005	2.80766
$n = 50$	0.07071	1.84672	$n = 25,000$	0.00316	2.95202
$n = 75$	0.05774	1.92722	$n = 50,000$	0.00224	3.05657
$n = 100$	0.05	1.98397	$n = 100,000$	0.00158	3.15966
$n = 250$	0.03162	2.16132	$n = 250,000$	0.001	3.29057
$n = 500$	0.02236	2.29132	$n = 500,000$	0.00071	3.38571
$n = 1,000$	0.01581	2.41740	$n = 1,000,000$	0.0005	3.48077
$n = 2,500$	0.01	2.57780	$n = 2,500,000$	0.00032	3.59855
$n = 5,000$	0.00707	2.69464	$n = 5,000,000$	0.00022	3.69487

Table 1. Significance thresholds for different sample sizes. *Notes*: The table reports two-sided significance levels (α_n) and the corresponding positive t critical values (τ_n) for different sample sizes (n). The significance levels are calculated using the rule of thumb in (5). All values are rounded to 5 places after the decimal point. *Example*: For a sample size of 500 observations, the significance level and t critical value are $\alpha_{500} = 0.02236$ and $\tau_{500} = 2.29132$, respectively. Thus, the sample size "penalization" is 0.02764 on the significance level and 0.30735 on the t critical value.

4 Large sample size bias in empirical finance

4.1 2020 midyear survey of finance research papers

A 2020 midyear survey of papers published in the top three finance journals (*Journal of Finance*, *Journal of Financial Economics*, *Review of Financial Studies*) has been conducted. The total number of published papers is 183. Papers with no empirical content and papers or regressions that do not clearly indicate sample sizes are excluded from the survey. The survey is confined to 114 studies with 5,943 regressions.

Table 2 reports summary statistics of the survey. The median sample size is 17,651, with the minimum and maximum being 20 and 120 million, respectively. All surveyed papers except one use the conventional significance thresholds, either explicitly or implicitly. Specifically, 100 papers use asterisks beside the estimates to indicate statistical significance, 13 papers do not report asterisks but focus their discussion on significant effects based on the conventional significance thresholds, and 1 paper uses the stricter thresholds for statistical significance suggested by Harvey et al. (2016). Out of a total of 5,943 regressions, 1,624 (27%) employ sample sizes of 2,500 observations or less. For these regressions, at least one of the conventional significance thresholds is appropriate (see Table 1). For the remainder 4,319 (73%) regressions, the conventional significance thresholds seem to be inappropriate.

# total papers	183	A. basic summary statistics (# obs.)	
# surveyed papers	114	Minimum	20
# surveyed regressions	5,943	Maximum	119,961,752
B. reported significance levels (# papers)		Median	17,651
*0.1, **0.05, ***0.01	89	Mean	516,425
*0.1, **0.05	6	C. employed sample size (# regressions)	
**0.05, ***0.01	2	≤ 25	9
**0.05	2	$25 < n \leq 100$	173
**0.05, ***0.01, ****0.001	1	$100 < n \leq 2,500$	1,442
no asterisks	14	$2,500 < n$	4,319

Table 2. Summary statistics of 2020 midyear survey. *Notes*: The table reports summary statistics of a 2020 midyear survey of papers published in the top three finance journals (*Journal of Finance, Journal of Financial Economics, Review of Financial Studies*). The total number of published papers is 183. Papers with no empirical content and papers or regressions that do not clearly indicate sample sizes are excluded from the survey. The survey is confined to 114 studies of empirical content. These 114 studies report a total of 5,943 regressions. Panel A reports the minimum, maximum, median, and mean number of observations for a sample of 5,943 regressions. Panel B groups the 114 papers according to their reported significance levels. Panel C reports the number of regressions that employ sample sizes of less than or equal to 25, greater than 25 and less than or equal to 100, greater than 100 and less than or equal to 2,500, and greater than 2,500 observations.

The survey indicates that the use of large and massive sample sizes is widespread in empirical finance, while the conventional significance thresholds are routinely used

by empirical researchers (see also Kim and Ji (2015)). Besides, when the median and maximum are compared with those in Kim and Ji (2015, p. 6, 2012 survey: median of 4,719; maximum of 6.96 million), it is clear that there has been an increase in sample sizes over the past decade. Nevertheless, little effort has been made towards the direction of adjusting significance thresholds as functions of sample size. It is notable that Fisher's exact statistical significance test was proposed as a small sample test, and never intended for large or massive sample sizes (see Keuzenkamp and Magnus (1995, p. 12)).

4.2 Statistical significance is not routinely significant

As can be seen in Table 1, the appropriate thresholds for statistical significance vary significantly from smaller to larger sample sizes. Hence, there is no single threshold that would fit all empirical studies. Johnson (2013), for instance, recommended that the level of significance should be set at 0.001 or 0.005. Although these significance levels seem to be appropriate for empirical studies with sample sizes of tens or hundreds of thousands of observations (see Table 1), they may be conservative for smaller sample sizes or moderate for larger ones.

Fig. 1 illustrates how statistical significance may change when adjusting the significance level as a decreasing function of sample size using the rule of thumb in (5). Regs. 1-4 present 51 statistically significant p-values reported in regression tables of two studies. Reg. 1 employs a sample of 20 observations. The adjusted significance level for this small sample size is 0.11. In this case, the conventional significance levels seem appropriate. Reg. 2 employs a sample of 276 observations. Though not too large, this sample size is large enough to violate the commonly used significance levels of 0.05 and 0.10; the adjusted significance level for this sample size is 0.03. Reg. 3

and 4 employ sample sizes of hundreds of thousands and millions of observations, respectively. The adjusted significance levels for these sample sizes are as small as 0.0007 and 0.00005. For such large and massive sample sizes, the conventional significance levels seem inappropriate. In total, out of the 51 statistically significant p-values, 17 become insignificant after adjusting the significance level.

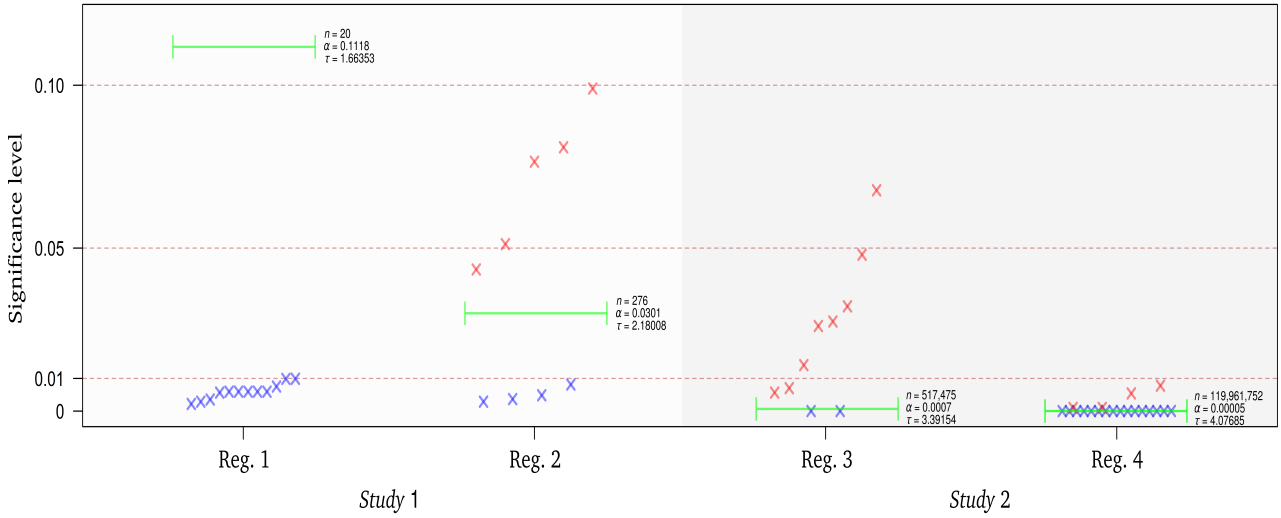

Fig. 1. Illustrations of how statistical significance may change when adjusting the significance level as a decreasing function of sample size. *Notes*: The figure illustrates how statistical significance may change when the significance level is adjusted as a decreasing function of sample size using the rule of thumb in (5). Study 1 and 2 are the articles of Chen et al. (2020) and Kostovetsky and Warner (2020), respectively. Reg. 1 presents the 12 statistically significant p-values in Chen et al. (2020, Table I, Panel B, Columns 1-6; n=20). Reg. 2 presents the 9 statistically significant p-values in Chen et al. (2020, Table I, Panel A, Columns 1-3; n=276). Reg. 3 presents the 10 statistically significant p-values in Kostovetsky and Warner (2020, Table III, Panel A, Columns 1-6; n=517,475). Reg. 4 presents the 20 statistically significant p-values in Kostovetsky and Warner (2020, Table III, Panels C-D, Columns 1-4,6; n=119,961,752). All p-values are indicated by cross markers. The green lines denote the appropriate significance levels calculated using the rule of thumb in (5). Blue cross markers below the green lines denote p-values that remain statistically significant. Red cross markers above the green lines denote p-values that become statistically insignificant. Reg. 1: p-values (12), blue (12), red (0); Reg. 2: p-values (9), blue (4), red (5); Reg. 3: p-values (10), blue (2), red (8); Reg. 4: p-values (20), blue (16), red (4). The figure is used for illustrative purposes only and is not intended to suggest that the research findings of the two studies are false.

The illustrations in fig. 1 suggests that the conventional thresholds for statistical significance should not be routinely used by empirical researchers. Instead, more appropriate significance thresholds should be chosen so as to take into consideration the effect of sample size on error probabilities and statistical power. In empirical

finance, where the use of large and massive data sets is widespread, the appropriate thresholds are expected to be fairly conservative. Yet, the choice should be made on a case-by-case basis because sample sizes vary greatly from one study to another, as well as from one study's regression to the next.

4.3 Factor discovery studies

The conventional thresholds for statistical significance were used exclusively in factor discovery studies until recently. The first study to raise concern of this issue was Harvey et al. (2016), where the authors introduced a new multiple testing framework, and provided historical and future t critical values to account for data mining. This framework, however, does not take into consideration large sample size bias. To account for both data mining and large sample size bias, the rule of thumb in (5) can be combined with the multiple testing framework of Harvey et al. (2016):

$$\tau(HLZ_{y,n}) = \underbrace{1.98}_{\tau_{100}} + \overbrace{[\tau(HLZ_y) - 1.98]}^{\text{data mining "penalization"}} + \underbrace{\left[\tau\left[\min\left(0.5,\ \frac{0.5}{\sqrt{n}}\right)\right] - 1.98\right]}_{\text{sample size "penalization"}}, \qquad (6)$$

where $\tau(HLZ_y)$ denotes the t critical value[1] for year y (see Harvey et al. (2016, p. 25)) and $\tau\left[\min\left(0.5,\ \frac{0.5}{\sqrt{n}}\right)\right]$ denotes the adjusted t critical value for sample size n (see rule of thumb in (5)). The "$\tau(HLZ_{y,n})$" reads "the 'penalized' t critical value for year y and sample size n".

After accounting for large sample size bias, the t critical values are expected to be higher (and significance levels lower) than the ones provided by Harvey et al. (2016). For example, the appropriate t critical value for a factor to be discovered today, assuming a sample size of $n = 678$ (July 1963 to December 2019), would be about

[1]Harvey et al. (2016) derived t critical values but one can easily convert those values into significance levels. Hence, the equation in (6) can also be expressed in terms of significance levels.

3.85 (significance level of 0.00013); that is, a data mining "penalization" of about 1.5 and a sample size "penalization" of 0.35.

5 Concluding remarks

The discussion in the paper is intended to serve as a recommendation to the finance research community. In empirical finance, the conventional thresholds for statistical significance are used exclusively, with little consideration of the effect of sample size on error probabilities (Type I and II) and statistical power. These conventional thresholds should not be used routinely in order to mitigate the risk of producing spurious statistically significant results. Empirical researchers are advised to determine appropriate significance thresholds prior to conducting any empirical analysis. Similarly, journals are encouraged to ask authors to adopt more appropriate significance thresholds. The choice of such thresholds should be made on a case-by-case basis, by adjusting the level of significance as a decreasing function of the sample size under consideration. One way to calculate appropriate significance thresholds is to employ the rule of thumb presented in this paper. The adoption of more appropriate significance thresholds will ensure the credibility of empirical findings and prevent the publication of unreliable research.

Acknowledgements. I would like to thank the editor, anonymous referees, and Aris Spanos for valuable comments and suggestions on previous versions of the paper. All errors are my own.

References

[1] Arrow, K.J., 1960. Decision theory and the choice of a level of significance for the t-test, in Olkin, I., Ghurye, S.G., Hoefding, W., Madow, W.G., Mann, H.B.

(eds.), Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. Stanford University Press, Stanford, pp. 70-78.

[2] Black, F., 1993. Beta and return. *Journal of Portfolio Management*, 20 (1), 8-18.

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

[4] Chen, H., Michaux, M., Roussanov, N., 2020. Houses as ATMs: mortgage refinancing and macroeconomic uncertainty. *Journal of Finance*, 75 (1), 323-375.

[5] Christensen, G., Miguel, E., 2018. Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, 56 (3), 920-80.

[6] De Prado, M.L., 2015. The future of empirical finance. *Journal of Portfolio Management*, 41 (4), 140-144.

[7] Ferson, W.E., Sarkissian, S., Simin, T.T., 2003. Spurious regressions in financial economics?. *Journal of Finance*, 58 (4), 1393-1413.

[8] Fisher, R.A., 1925. Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh, U.K.

[9] Fisher, R.A., 1956. Statistical Methods and Scientific Inference. Hafner Publishing Co., Oxford, U.K.

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

[11] Ioannidis, J.P., 2005. Why most published research findings are false. *PLoS Medicine*, 2 (8), e124, 696-701.

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

[13] Harford, T., 2014. Big data: a big mistake?. *Significance*, 11 (5), 14-19.

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

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

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

[17] Kaplan, R.M., Chambers, D.A., Glasgow, R.E, 2014. Big data and large sample size: a cautionary note on the potential for bias. *Clinical and translational science*, 7 (4), 342-346.

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

[19] Kim, J.H., Ji, P.I., 2015. Significance testing in empirical finance: a critical review and assessment. *Journal of Empirical Finance*, 34, 1-14.

[20] Kostovetsky, L., Warner, J.B., 2020. Measuring innovation and product differentiation: evidence from mutual funds. *Journal of Finance*, 75 (2), 779-823.

[21] Leamer, E.E., 1978. Specification Searches: Ad Hoc Inference with Nonexperimental Data, Wiley, New York.

[22] Lehmann, E.L., Romano, J.P., 2005. Testing Statistical Hypotheses, third ed. Springer Science and Business Media, New York.

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

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

[25] Mayo, D.G., Spanos, A., 2006. Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *British Journal for the Philosophy of Science*, 57 (2), 323-357.

[26] McCloskey, D.N., Ziliak, S.T., 1996. The standard error of regressions. *Journal of Economic Literature*, 34 (1), 97-114.

[27] Neyman, J., Pearson, E.S., 1928a. On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, 20A (1-2), 175-240.

[28] Neyman, J., Pearson, E.S., 1928b. On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. *Biometrika*, 20A (3-4), 263-294.

[29] Peng, R., 2015. The reproducibility crisis in science: a statistical counterattack. *Significance*, 12 (3), 30-32.

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

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

[32] Welch, I., Goyal, A., 2008. A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21 (4), 1455-1508.