Finance and Economics Discussion Series
Divisions of Research & Statistics and Monetary Affairs
Federal Reserve Board, Washington, D.C.

The Limits of p-Hacking: a Thought Experiment

Andrew Y Chen

2019-016

Please cite this paper as:
Chen, Andrew Y. (2019). "The Limits of p-Hacking: a Thought Experiment," Finance and Economics Discussion Series 2019-016. Washington: Board of Governors of the Federal Reserve System, https://doi.org/10.17016/FEDS.2019.016.

The Limits of p-Hacking: a Thought Experiment

Andrew Y. Chen

Federal Reserve Board

andrew.y.chen@frb.gov

January 2019*

Abstract

Suppose that asset pricing factors are just p-hacked noise. How much p-hacking is required to produce the 300 factors documented by academics? I show that, if 10,000 academics generate 1 factor every minute, it takes 15 million years of p-hacking. This absurd conclusion comes from applying the p-hacking theory to published data. To fit the fat right tail of published t-stats, the p-hacking theory requires that the probability of publishing t-stats < 6.0 is infinitesimal. Thus it takes a ridiculous amount of p-hacking to publish a single t-stat. These results show that p-hacking alone cannot explain the factor zoo.

*I thank Preston Harry for excellent research assistance and Steve Sharpe for helpful comments. The views expressed herein are those of the authors and do not necessarily reflect the position of the Board of Governors of the Federal Reserve or the Federal Reserve System.

1. Introduction

> *There is a well-known solution to every human problem—neat, plausible, and wrong.*
>
> — H.L. Mencken (1920), Prejudices: Second Series.

Academics have documented more than 300 factors that explain expected stock returns.[1] This enormous set of factors begs for an economic explanation, yet there is little consensus on their origin.[2]

p-hacking (a.k.a. data-snooping, data-mining) offers a neat and plausible solution (Harvey, Liu, and Zhu 2016, Chordia, Goyal, and Saretto 2017, Hou, Xue, and Zhang 2017, Linnainmaa and Roberts 2018, among others). This cynical explanation begins by noting that the cross-sectional literature uses statistical tests that are only valid under the assumptions of classical single hypothesis testing. These assumptions are clearly violated in practice, as each published factor is drawn from multiple unpublished tests. In this well-known explanation, the factor zoo consists of factors that performed well by pure chance.

In this short paper, I follow the p-hacking explanation to its logical conclusion. To rigorously pursue the p-hacking theory, I write down a statistical model in which factors have no explanatory power, but published t-stats are large because the probability of publishing a t-stat t_i follows an increasing function $p(t_i)$. I estimate $p(t_i)$ by fitting the model to the distribution of published t-stats in Harvey, Liu, and Zhu (2016) and Chen and Zimmermann (2018). The p-hacking story is powerful: The model fits either dataset very well.

Though p-hacking fits the data, following its logic further leads to absurd conclusions. In particular, the pure p-hacking model predicts that the ratio of unpublished factors to published factors is ridiculously large, at about 100 *trillion* to 1. To put this number in perspective, suppose that 10,000 economists mine the data for 8 hours per day, 365 days per year. And suppose that each economist

[1] I use the term "factor" to refer to any variable that helps explain expected returns, following Harvey, Liu, and Zhu (2016).

[2] Cochrane (2017) provides a macro-finance perspective on predictability. Barberis (2018) provides a psychological perspective. Recent explicit factor models based on q-theory, the present value relation, and mispricing are given by Hou, Xue, and Zhang (2015), Fama and French (2015), and Stambaugh and Yuan (2016), respectively. Rigorous statistical explanations for cross-sectional predictability are proposed by Kozak, Nagel, and Santosh (2017), Kelly, Pruitt, and Su (2017), and Lettau and Pelger (2018).

finds 1 predictor every *minute*. Even with this intense p-hacking, it would take 15 million years to find the 316 factors in the Harvey, Liu, and Zhu (2016) dataset.

This absurd conclusion comes from the fact that the right tail in published t-stats is extremely fat compared a t-distribution with many degrees of freedom. 10% of t-stats in Harvey, Liu, and Zhu (2016) are larger than 6.34, while the corresponding p-value of the t-distribution with 200 degrees of freedom is 0.00000007%. Thus, to account for the fat tail in the data, authors and journals must have an extremely strong preference for very large t-stats: t-stats less than 4.0 have at most a 10^{-10} probability of being published, while t-stats larger than 8.0 are published with a probability of 0.9997. While it is hard to place reasonable limits on the preference for large t-stats, logistical and physical constraints imply that the power of p-hacking is limited, far too limited to account for the literature on asset pricing factors.

This thought experiment demonstrates that assigning the entire factor zoo to p-hacking is wrong. Though the p-hacking story appears logical, following its logic rigorously leads to implausible conclusions, disproving the theory by contradiction. Thus, my thought experiment supports the idea that publication bias in the cross-section of stock returns is relatively minor (Green, Hand, and Zhang 2014, McLean and Pontiff 2016, Jacobs and Müller 2017, Chen and Zimmermann 2018, Chen 2018). Papers that argue that publication bias is dominant include Harvey, Liu, and Zhu (2016), Chordia, Goyal, and Saretto (2017), Hou, Xue, and Zhang (2017), and Linnainmaa and Roberts (2018). In this literature, my paper is unique in its rigorous analysis of the p-hacking story.

2. Model, Estimation Method, Data

This section presents a rigorous version of the p-hacking story and describes how I fit it to data. Estimation results and absurd implications are found in Section 3.

2.1. Model

The distribution of all t-stats (published and not) is standard normal

$$t_i \sim N(0,1), \text{ i.i.d.} \tag{1}$$

This assumption formalizes the notion that all factors are false: t-stats are just noise around the unobserved population return of 0.

Some readers may object to the independence assumption, noting that several well-known anomalies are related to value or momentum. Value- and momentum- related anomalies, however, comprise only a small portion of the total universe of published anomalies (Harvey, Liu, and Zhu 2016, McLean and Pontiff 2016). For example, in the Chen and Zimmermann (2018) dataset, predictors related to valuations represent only 8% of their 156 predictors. Momentum-related predictors represent only 6%.

Ultimately, the proper correlation should be measured from the data, and the data indicate close-to-zero correlation is appropriate. The average pairwise correlation between predictor returns is tiny, at 0.03 (McLean and Pontiff 2016, Chen and Zimmermann 2018). This tiny average correlation does *not* result from averaging across large positive and large-negative correlations. Indeed, Chen and Zimmermann (2018) find that 80% of correlations are between -0.36 and 0.43. Moreover, Chen and Zimmermann find that principal component analysis indicates that a large number of principal components are required to span the data.

Equation (1) also assumes normality. This assumption is justified by the fact that the numerator of the t-stat is the average of hundreds of monthly returns. Thus, by the central limit theorem, the sample mean return is approximately normal and the t-stat is approximately standard normal. Chen and Zimmermann (2018) show that this approximation holds very well for a 312 month sample of equal-weighted long-short quintile portfolios sorted on B/M. Equation (1) also assumes that performance is uncorrelated across predictors, consistent with the near-zero average pairwise correlation between monthly long-short returns of different published predictors (McLean and Pontiff 2016, Chen and Zimmermann 2018).

Though t-stats are on average zero, published t-stats are large due to authors' and journals' preferences for large t-stats. This preference is embodied in the function $p(t_i)$ which determines the probability that a t-stat t_i is published. I

assume a staircase (or step) function for $p(t_i)$:

$$p(t_i) = \begin{cases} p_1, & e_1 < t_i \leq e_2 \\ p_2, & e_2 < t_i \leq e_3 \\ ... & \\ p_K, & e_K < t_i \leq e_{K+1} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where the edges $\{e_1, e_2, ..., e_{K+1}\}$ and probabilities $\{p_1, p_2, ..., p_K\}$ are model parameters. In words, p_i is the probability of publishing a t-stat between e_i and e_{i+1}.

Equation (2) is a rigorous version of the p-hacking story. t-stats $< e_1$ are never published or observed by the public. The staircase functional form allows for the idea that larger t-stats are more likely to be published. The flexibility of the K step staircase allows the model to fit the data very closely and provides a tractable, closed-form estimation.

2.2. Estimation

The model predicts that the fraction of published t-stats between e_i and e_{i+1} is

$$f_i^{\text{model}} = \frac{p_i \left[\Phi(e_{i+1}) - \Phi(e_i)\right]}{\sum_{j=1}^{K} p_j \left[\Phi(e_{j+1}) - \Phi(e_j)\right]} \quad \text{for } i = 1, ..., K \tag{3}$$

where $\Phi(\cdot)$ is the standard normal CDF. Equation (3) embodies the power of the p-hacking theory. It says that any number of t-stats can be observed. Even if it is unlikely to observe such a large t-stat by chance ($\Phi(e_{i+1}) - \Phi(e_i)$ is small), a large publication probability p_i can make it possible.

Equation (3) suggests an intuitive method-of-moments estimation. First choose a set of edges $\{e_1, e_2, ..., e_{K+1}\}$ that produces a histogram that describes the data well. Then, measure in the data the fraction of published t-stats between e_i and e_{i+1} and call this f_i^{data}. Finally, setting $f_i^{\text{model}} = f_i^{\text{data}}$ gives a set of

K equations to solve for the K probabilities $\hat{p}_1, ..., \hat{p}_K$. Specifically,[3]

$$\hat{p}_i \equiv \frac{1}{\kappa} \frac{f_i^{\text{data}}}{[\Phi(e_{i+1}) - \Phi(e_i)]} \tag{4}$$

where

$$\kappa \equiv \sum_{j=1}^{K} \frac{f_j^{\text{data}}}{\left[\Phi(e_{j+1}) - \Phi(e_j)\right]}. \tag{5}$$

The model is exactly identified, and thus Equation (4) does not provide any formal evaluation of the model. Instead, I discipline the model by examining a simple thought experiment in Section 3.2.

For histogram edges e_i I use $\{1, 2, 3, ..., 8, \infty\}$. Other edges lead to similar results.

2.3. Data on Published t-stats

I estimate the model on 2 datasets. The first is Chen and Zimmermann's (2018) replications of 156 equal-weighted long-short quintile portfolios. These portfolios are constructed from variables that have been shown to predict stock returns cross-sectionally and are published in finance, accounting and general interest economics journals. The majority are constructed using either accounting data or market prices, but about 1/3 use diverse data that include analyst forecasts, trading-related measures, and corporate events. The Chen and Zimmermann dataset allows for easy replication, as this data is publically available at http://sites.google.com/site/chenandrewy/code-and-data/.

I also consider the hand-collected t-stats for 316 factors in Harvey, Liu, and Zhu (2016). These factors include variables that predict cross-sectional returns, as well as other variables that broadly explain return patterns. Harvey et al do not make their data publically available, but in Table 5 (page 30) they provide parameter estimates for a model of the t-stats in their data. Using their model estimates, I can simulate their dataset. By design, this simulated data should match the moments in the original data. I use the parameter values from the first row of Table 5, but the other parameters lead to similar results.

[3] To see this, note that $f_i^{\text{data}} = \kappa \hat{p}_i [\Phi(e_{i+1}) - \Phi(e_i)]$. Then, noting that $\sum_i \hat{p}_i = 1$, we have $\kappa = \sum_j f^{\text{data}} / \left[\Phi(e_{j+1}) - \Phi(e_j)\right]$.

Table 1 summarizes the datasets. It shows the histogram counts for t-stats in percent. I use these counts as target moments in Equation (4). For comparison, the table also shows the histogram counts of the hand-collected data from Chen and Zimmermann (2018).

[Table 1 about here.]

All three datasets show a fat right tail in t-stats. About 50% of t-stats are between 2.0 and 4.0, and the remaining 50% are spread far out and to the right. At least 15% of t-stats are greater than 6.0 using any of the three datasets.

3. Results

3.1. Estimated Preference for Large t-stats

Figure 1 illustrates the model fit and estimation results. The figure plots the histogram of t-stats data (bars) and model (circle markers), along with the estimated preference for t-stats (triangle markers). The top panel uses the Chen and Zimmermann (2018) (CZ) replicated data, and the bottom panel uses moments from Harvey, Liu, and Zhu (2016) (HLZ).

[Figure 1 about here.]

As the model is exactly identified using the t-stat histogram, the model fit in Figure 1 is very good by construction. This fit illustrates the powerful logic of p-hacking: one can generate any pattern if the data is selectively published.

The implied preference for large t-stats, however, is very extreme. This preference is characterized by 8 parameters p_1, p_2, ..., p_8 corresponding to the probability of publishing a t-stat in each bin. The probabilities are so extreme that they need to be plotted on a log scale (triangles, right axis), and range from 10^{-14} for t-stats between 1 and 2 to 0.99977 for t-stats in excess of 8.0 for the CZ data. The HLZ data leads to similar results.

These highly skewed probabilities come from the very thin tail of a standard normal distribution. This can be seen in the bottom row of Table 1, which shows the histogram counts implied by a standard normal distribution that is truncated at 2.0. Roughly 0.00001% of t-stats exceed 6.0 in this distribution, compared to

the roughly 15% of t-stats that exceed 6.0 in the data. Thus, in order for the data to be generated by p-hacking, the publication probability for these large t-stats must be very high compared to those for smaller t-stats, leading to the extreme skewed probabilities seen in Figure 1.

3.2. A Thought Experiment

It's difficult to say if the estimated preference for large t-stats in Figure 1 is reasonable. The probability that a given t-stat is published depends on the choices of the both authors and journals. These choices interact, making interpretation difficult.

However, one can interpret the t-stat preferences easily in a thought experiment. This thought experiment tests the plausibility of the p-hacking story in the same way Mehra and Prescott's (1985) calibration exercise tests the plausibility of the power utility model of equity prices.

Suppose that N economists mine the data 8 hours per day, 365 days per year. Suppose further that the economists produce factors at a rate of x per economist-hour. How long would it take to produce 100 factors?

To answer this question, I need to calculate the probability that a random t-stat is published. This probability is found by integrating the probability of publication (2) over the distribution of t-stats. The staircase form of (2) implies a closed form expression:

$$\text{Probability of Publishing a Random t-stat} = \sum_{i=1}^{K} \hat{p}_i \left[\Phi(e_{i+1}) - \Phi(e_i)\right] \tag{6}$$

where, as a reminder, $\Phi(\cdot)$ is the standard normal CDF. Plugging in $\hat{p}_i$ from Figure 1 and the standard normal probabilities from Table 1 we have

$$\text{Probability of Publishing a Random t-stat} = 6.49e-15 \tag{7}$$

using the Chen and Zimmermann (2018) and

$$\text{Probability of Publishing a Random t-stat} = 1.23e-14. \tag{8}$$

using Harvey, Liu, and Zhu (2016).

These infinitesimal probabilities come from the fact that the estimated prob-

abilities of publication in Figure 1 and probabilities implied by the standard normal distribution (see Table 1) are largely disjoint. For t-stats below 6.0, $\hat{p}_i$ is extremely small, but for t-stats above 6.0, the standard normal density implies a tiny probability. Summing over the product of these probabilities, Equation (6) implies an extremely tiny probability of publication.

Using this probability of publication, I calculate the number of years it takes to publish 100 factors, assuming various numbers of economists and rates of factor production. As both probabilities are extremely small, I focus on the larger the probability implied by the Harvey, Liu, and Zhu (2016) dataset.

[Table 2 about here.]

Table 2 shows the result. The table begins by assuming that 10,000 economists mine the data. If these 10,000 economists produce factors at a rate of 1 per economist-hour, it takes 528 *million* years to publish 100 factors. To put this number in perspective, the number of economics professors in the United States was 12,770 in 2017, and the number of economists was 21,300 in 2016 according to the Bureau of Labor Statistics.

One might argue that factors can be mined at a much faster rate than 1 per economist-hour given modern computing power. However, factors need to come with supplementary results that satisfy journal review in order to be published. For example, portfolio sorts are often required to produce monotonic patterns in expected returns, alternative methods for factor construction are sometimes required, and the factors themselves are typically asked to be consistent with some kind of theory for journals to publish them. These additional restrictions are difficult to satisfy using computing power alone.

Regardless, I can pursue the idea of highly productive factor mining in this thought experiment. Table 2 shows that, even at a factor production rate of 10 per economist-second, it would take 15,000 years for 10,000 economists to publish 100 factors.

Table 2 also explores the possibility that more than 10,000 economists engage in p-hacking. Even if 1 million economists mine the data at 10 factors per economist-second, it would still take 145 years to publish 100 factors. To put these numbers in perspective, the Bureau of Labor Statistics estimates that there were 296,100 financial analysts in the United States in 2016.

Finally, the bottom row of Table 2 shows that if 1 million economists produce

40 factors per economist-second, then 100 factors will be published in just 19 years. However, the idea that 1 million economists can, in every second, produce 40 factors that have the supplementary results required for publication, and do so consistently for 19 years, is ridiculous.

4. Conclusion

The idea that all asset pricing factors are due to p-hacking is very tempting. In one fell swoop, p-hacking can explain decades of puzzling financial research. A rigorous exploration of this explanation, however, shows that it is implausible. Though it may be difficult to understand, the stock return data does display cross-sectional variation in expected returns.

References

Barberis, Nicholas C. *Psychology-based Models of Asset Prices and Trading Volume*. Tech. rep. National Bureau of Economic Research, 2018.

Chen, Andrew Y. "Do t-stat Hurdles Need to be Raised? Direct Estimates of False Discoveries in the Cross-Section of Stock Returns". *Available at SSRN: https://papers.ssrn.com/abstract=3254995* (2018).

Chen, Andrew Y and Tom Zimmermann. "Publication Bias and the Cross-Section of Stock Returns". *Available at SSRN: https://ssrn.com/abstract=2802357* (2018).

Chordia, Tarun, Amit Goyal, and Alessio Saretto. "p-hacking: Evidence from two mmillion trading strategies" (2017).

Cochrane, John H. "Macro-finance". *Review of Finance* 21.3 (2017), pp. 945–985.

Fama, Eugene F. and Kenneth R. French. "A five-factor asset pricing model". *Journal of Financial Economics* 116.1 (2015), pp. 1–22. ISSN: 0304-405X. URL: http://www.sciencedirect.com/science/article/pii/S0304405X14002323.

Green, Jeremiah, John RM Hand, and Frank Zhang. "The remarkable multidimensionality in the cross-section of expected US stock returns". *Available at SSRN 2262374* (2014).

Harvey, Campbell R, Yan Liu, and Heqing Zhu. "... and the cross-section of expected returns". *The Review of Financial Studies* 29.1 (2016), pp. 5–68.

Harvey, Campbell and Yan Liu. "Multiple testing in economics" (2013).

Hou, Kewei, Chen Xue, and Lu Zhang. "Digesting anomalies: An investment approach". *The Review of Financial Studies* 28.3 (2015), pp. 650–705.

— *Replicating Anomalies*. Tech. rep. National Bureau of Economic Research, 2017.

Jacobs, Heiko and Sebastian Müller. "Anomalies across the globe: Once public, no longer existent?" (2017).

Kelly, Bryan T, Seth Pruitt, and Yinan Su. "Some characteristics are risk exposures, and the rest are irrelevant" (2017).

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh. "Shrinking the Cross Section" (2017).

Lettau, Martin and Markus Pelger. "Factors that Fit the Time Series and Cross-Section of Stock Returns" (2018).

Linnainmaa, Juhani T and Michael R Roberts. “The history of the cross-section of stock returns”. *The Review of Financial Studies* 31.7 (2018), pp. 2606–2649.

McLean, R David and Jeffrey Pontiff. “Does academic research destroy stock return predictability?” *The Journal of Finance* 71.1 (2016), pp. 5–32.

Mehra, Rajnish and Edward C Prescott. “The equity premium: A puzzle”. *Journal of monetary Economics* 15.2 (1985), pp. 145–161.

Stambaugh, Robert F and Yu Yuan. “Mispricing factors”. *The Review of Financial Studies* 30.4 (2016), pp. 1270–1315.

Exhibits

Figure 1: Model Fit and t-stat Preference. I estimate a model of pure p-hacking (Equations (1)-(2), circles) on large datasets of published t-stats (bars) by method of moments (Section 2.2). The top panel uses 156 replicated long-short portfolios from Chen and Zimmermann (2018). The bottom panel uses moments from Harvey, Liu, and Zhu (2016). The t-stat preference is modeled as publication probabilities (triangles). The estimated preference for large t-stats is extremely strong. t-stats < 6.0 have an absurdly low probability of publication.

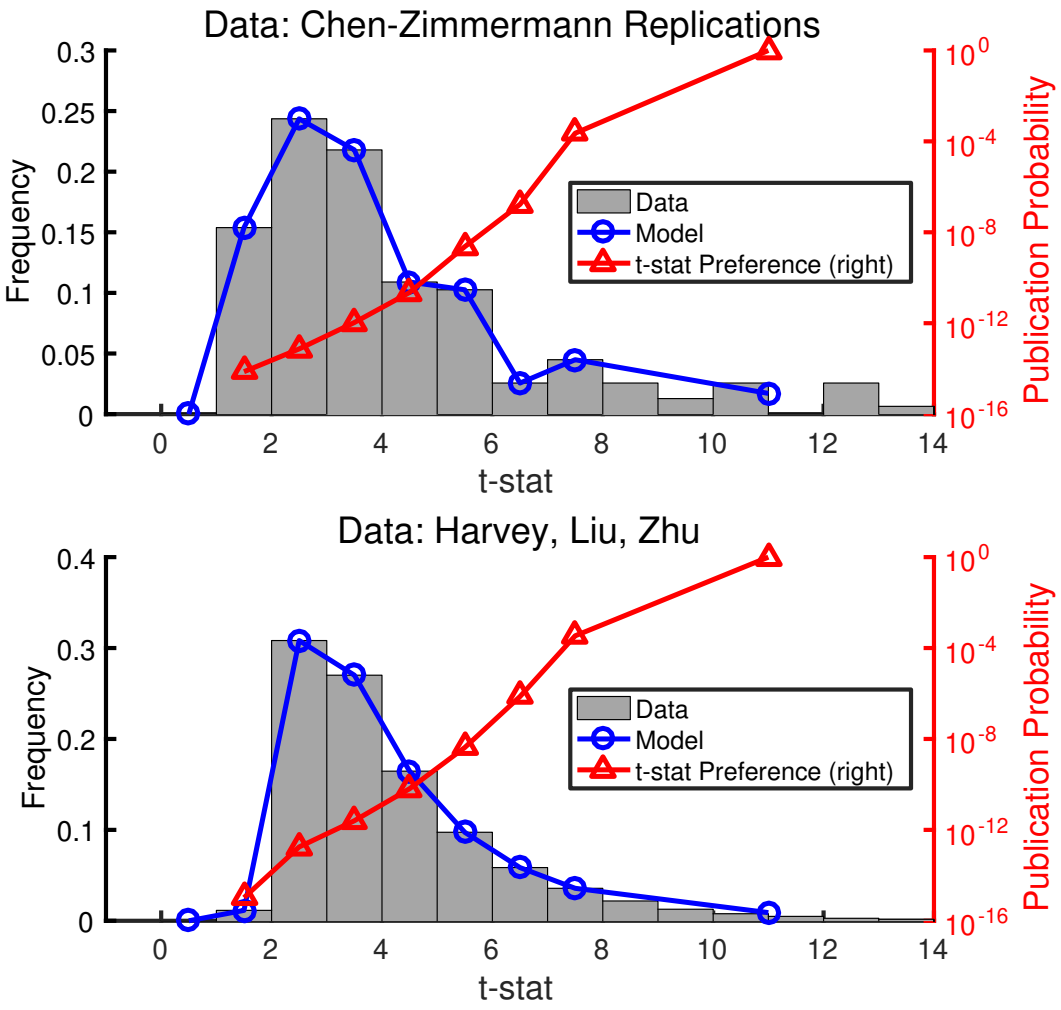

Table 1: Distribution of Published t-stats

This table summarizes the data and provides moments used in the estimation. CZ replications is the 156 replications of equal-weighted long-short quintile portfolios in Chen and Zimmermann (2018). HLZ estimated model simulates the model from Harvey, Liu, and Zhu (2016) Table 5, first row. For comparison, I show the 77 hand collected statistics from Chen and Zimmermann (2018) (CZ hand collection) and a standard normal truncated at 2.0.

	percent of t-stats between							
	1, 2	2, 3	3, 4	4, 5	5, 6	6, 7	7, 8	> 8
Used in Estimation								
CZ replications	15.4	24.4	21.8	10.9	10.3	2.6	4.5	10.3
HLZ estimated model	1.2	30.9	27.1	16.2	9.7	5.9	3.5	5.4
For Comparison								
CZ hand collection	6.4	29.5	20.5	9.0	14.1	5.1	1.3	14.1
standard normal truncated at 2.0	-	94.1	5.8	0.1	1E-03	4E-06	6E-09	3E-12

Table 2: A Thought Experiment

I calculate the probability that a random t-stat is published (Equation (6)). Using the Harvey, Liu, and Zhu (2016) data, this probability is 1.23e-14. Applying this probability to the assumed number of economists and factors per economist-hour in the table leads to the number publications per year and years to publish 100 factors. For comparison, there were 12,770 economics professors in the United States in 2017 and 21,300 professional economists in 2016 according to the Bureau of Labor Statistics.

Number of Economists	Factors per Economist-Hour	Factors per Year (Millions)	Publications per Year	Years to Publish 100 Factors
10,000	1	29	3.60E-07	277,524,922
10,000	60	1,752	2.16E-05	4,625,415
10,000	3,600	105,120	1.30E-03	77,090
10,000	36,000	1,051,200	1.30E-02	7,709
100,000	36,000	10,512,000	0.13	771
500,000	36,000	52,560,000	0.65	154
1,000,000	36,000	105,120,000	1.30	77
1,000,000	144,000	420,480,000	5.19	19