Most claimed statistical findings in cross-sectional return predictability are likely true

Andrew Y. Chen

Federal Reserve Board

May 2024[*]

Abstract

I develop simple and intuitive bounds for the false discovery rate (FDR) in cross-sectional return predictability publications. The bounds can be calculated by plugging in summary statistics from previous papers and reliably bound the FDR in simulations that closely mimic cross-predictor correlations. Most bounds find that at least 75% of findings are true. The tightest bound finds at least 91% of findings are true. Surprisingly, the estimates in Harvey, Liu, and Zhu (2016) imply a similar FDR. I explain how Harvey et al.'s conclusion that most findings are false stems from equating "false" and "insignificant."

JEL Classification: G0, G1, C1

Keywords: stock market predictability, stock market anomalies, p-hacking, multiple testing

[*] First posted to SSRN: August 27, 2021. More recent drafts: https://arxiv.org/abs/2206.15365. Replication code: https://github.com/chenandrewy/mostly-true. I thank Antonio Gil de Rubio Cruz and Rebecca John for excellent research assistance. For helpful comments, I thank an anonymous referee, Carter Davis (discussant), Wayne Ferson, Dalida Kadyrzhanova, Amit Goyal (discussant), Dino Palazzo, Jeffrey Pontiff, Russ Wermers, Dacheng Xiu (discussant), David Zeke, Min Zhu and seminar participants at the 2022 ASU Sonoran Winter Finance Conference, Boston College, the Future of Financial Information Webinar, 2024 Midwest Finance Association Meetings, the University of Nebraska, UC Riverside, and the University of Maryland. I am grateful to Sterling Yan and Lingling Zheng for sharing their data, which made the first draft possible. 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

Researchers have discovered hundreds of cross-sectional stock return predictors (Green, Hand, and Zhang (2013); Chen and Zimmermann (2022)). This abundance of discoveries has led to concerns about multiple testing, p-hacking, and a debate about the veracity of this literature (e.g. Harvey, Liu, and Zhu (2016); Chen and Zimmermann (2020)).

This paper provides simple arguments for why most statistical findings in this literature are true. The arguments build on Benjamini and Hochberg's (1995) insight that the false discovery rate (FDR) can be bounded using worst-case scenarios. Worst-case scenarios for publication bias can be formed from data-mining experiments (Yan and Zheng (2017)) or conservative extrapolations of published results. Plugging these scenarios into a simplified version of the Storey (2002) estimator leads to FDR bounds that are tractable yet valid under publication bias.

These bounds are so tractable that they can be evaluated by just plugging in summary statistics from data mining and predictor zoo studies. Using numbers reported in eight studies, performed by seven independent teams, I find that at least 75% of claimed statistical findings are true.[1]

I provide two derivations of these bounds. The first is a slight modification of Bayes rule and requires just a few lines of math. The second is a visual argument that does not require any equations at all—though it can be formally justified as a variant of Storey (2002). Applied to Chen, Lopez-Lira, and Zimmermann's (2024) (CLZ's) 29,000 data-mined long-short strategies, the visual argument leads to a much tighter bound on the FDR, and finds at least 91% of claimed findings are true.

The 75% and 91% bounds focus on equal-weighted predictability tests, matching the modal and primary tests in the cross-sectional literature (Green, Hand, and Zhang (2013); Chen and Zimmermann (2022)). However, value-weighted tests also imply that most findings are true. For example, applying the visual bound to the CLZ data implies that at least 82% of claimed findings regarding value-weighted CAPM alphas are true.

My FDR bounds are much simpler than the formulas that appear in the fi-

[1] I use the terms "discovery," "significant predictor," and "finding" interchangeably. I also use the terms "false predictor" and "null predictor" interchangeably. This terminology follows Sorić (1989) and Benjamini and Hochberg (1995). Predictors that are not false are "true."

nance literature (e.g. Harvey, Liu, and Zhu (2016)). Despite their simplicity, they are equivalent to the more complicated formulas under weak dependence (Storey and Tibshirani (2001)). To verify this equivalence, I run estimations on simulated data that cluster bootstrap from CLZ's data-mined returns. In these simulations closely mimic cross-predictor correlations, and show that my methods reliably bound the FDR.

My results appear to conflict with Harvey, Liu, and Zhu (2016) (HLZ), who "argue that most claimed research findings in financial economics are likely false." I show my FDR estimates are quantitatively consistent with HLZ's. HLZ's main figure implies an FDR of at most 35% and their SMM estimates imply an FDR of 9%. Thus, the conflict is entirely about interpretation. I explain how HLZ's interpretation equates "insignificant factors" with "false factors," deviating from the traditional separation of significance and truth (Wasserstein and Lazar (2016)). I illustrate how following HLZ's interpretation leads to absurd economic conclusions and extreme concentration of power among senior scholars.

This paper adds to the literature on multiple testing problems in cross-sectional predictability. Papers that find major problems include Harvey, Liu, and Zhu (2016), Harvey (2017), Linnainmaa and Roberts (2018), and Chordia, Goyal, and Saretto (2020). Papers that find the opposite include McLean and Pontiff (2016), Jacobs and Müller (2020), Chen and Zimmermann (2020), Jensen, Kelly, and Pedersen (2023), and Chen (Forthcoming). Among these papers, mine is unique in developing simple formulas that bound the FDR.[2] Simple formulas allow me to explore a more diverse range of data, demonstrating broader robustness compared to previous studies. They also provide a new level of transparency.

Transparency is important given the lack of consensus and complicated methods in the literature. Harvey, Liu, and Zhu's (2016) methods require so much exposition that their estimates that account for publication bias are not found until the 30th page of the article. Chordia, Goyal, and Saretto's (2020) analogous estimates are not found until page 21. Jensen, Kelly, and Pedersen's (2023) are not found until page 37. These complications are understandable. Publication bias is a slippery topic and accounting for it requires novel methods and thus convincing justifications. These complications, however, make the debate difficult to resolve. Debating any of these estimates requires, in principle, addressing

[2] Chen (2021) uses simple formulas but they cannot bound the FDR.

dozens of pages of motivation and methodology.

My formulas simplify the debate. My "easy bound" for the FDR is on page 4. FDR estimates that are valid under publication bias begin on page 5.

An important caveat is that I only examine predictability findings in the cross-sectional literature. I do not examine non-tradable factors, which are typically not shown to predict returns (Chen and Zimmermann (2022)). I also do not assess whether the claims made in the texts are true. Re-evaluating the claims in the texts typically requires structural modeling (Kozak, Nagel, and Santosh (2018)), though Chen, Lopez-Lira, and Zimmermann (2024) propose a test of textual claims using out-of-sample returns. I also do not examine whether predictability is stable over time. Chen and Velikov (2023) find that published strategies are not profitable in recent years after accounting for bid-ask spreads.

2. Why Most Claimed Statistical Findings Are True

Research generates predictors $i = 1,2,..,N$. Predictor i has t-stat t_d and may be false (F_d) (a.k.a. null). If i is false, its t-stat satisfies:

$$\Pr(|t_d| > 2|F_d) \approx 5\% \tag{1}$$

(i.e. standard errors are approximately correct). Define a "discovery" as a predictor with $|t_d| > 2$. Then the Benjamini and Hochberg (1995) FDR satisfies

$$\text{FDR}_{|t|>2} \approx \Pr\left(F_i||t_i| > 2\right), \tag{2}$$

if $|t_1|,|t_2|,...,|t_N|$ are well-behaved (Storey (2002)). I explain Storey's approximation in Section 4.1. Nevertheless, Equation (2) is so natural that one might as well define the FDR this way, as in Efron (2008).

Applying Bayes rule and noting $\Pr(F_i) \le 1$, leads to an "easy bound" on $\text{FDR}_{|t|>2}$:

$$\begin{aligned}
\text{FDR}_{|t|>2} &\approx \frac{\Pr\left(|t_i| > 2|F_i\right)\Pr\left(F_i\right)}{\Pr\left(|t_i| > 2\right)} && (3)\\
&\le \frac{\Pr\left(|t_i| > 2|F_i\right)}{\Pr\left(|t_i| > 2\right)} && (4)\\
&\approx \frac{5\%}{\Pr\left(|t_i| > 2\right)}, && (5)
\end{aligned}$$

where the third line plugs in Equation (1).

Equation (5) provides a neat interpretation of FDR methods. FDR methods just compare the tail area under the null (numerator) to the actual tail (denominator). If the tail under the null is much smaller, then the null distribution cannot explain the actual tail, and the FDR must be small.

The natural way to estimate the actual tail is to plug in the empirical tail. In other words, one might estimate $\Pr(|t_i| > 2)$ by just counting the share of $|t_i|$ that exceeds 2. However, since predictors with $|t_i| > 2$ are more likely to be published (publication bias), this estimate would overstate $\Pr(|t_i| > 2)$. This problem can be addressed, however, by considering worst-case scenarios.

2.1. Bounding the FDR with Atheoretical Data Mining.

One worst-case scenario comes from atheoretical data mining. Yan and Zheng (2017); Chordia, Goyal, and Saretto (2020); and Chen, Lopez-Lira, and Zimmermann (2024) systematically mine accounting data, with minimal use of theory, generating tens of thousands or even millions of long-short strategies. *All* of these papers find that at least 20% of data-mined strategies have absolute t-stats exceeding 2.

Assuming that research finds t-stats exceeding 2 at least as frequently as data mining, Equation (5) implies

$$\text{FDR}_{|t|>2} \leq \frac{5\%}{0.20} = 25\%. \tag{6}$$

Thus, at least 75% of published cross-sectional predictors are true.

The robustness of this bound is notable, given that each paper uses a different approach for data mining, as seen in Table 1. Furthermore, each study is conducted by an independent research team. Further details on these papers and their implied FDRs are in Appendix A.1.

The 25% bound can be considered both too aggressive and too conservative. Aggressiveness comes from focusing on equal-weighted returns, following the modal test in the cross-sectional literature (Green, Hand, and Zhang (2013)). Equal-weighted test focus on small, illiquid stocks, where predictability is known to be stronger, and thus the FDR is likely to be lower. Conservatism comes from plugging $\Pr(F_i) \leq 1.0$ into Equation (3). This amounts to assuming that all predic-

Table 1: FDR Bounds using Summary Stats from Data-Mining Studies

$\Pr(|t_i| > 2)$ is the share of predictors with $|t_i| > 2$, generated from the "Data Mining Method." Equal-weighted results are selected for comparability with the cross-sectional literature (Green, Hand, and Zhang (2013)). Chordia, Goyal, and Saretto's (2020) t_i largely use value-weighted returns, though they also examine FM regressions with many controls. Data mining studies consistently find $\Pr(|t_i| > 2) \geq 0.20$, implying $\text{FDR}_{|t|>2} \leq 5\%/0.20 = 25\%$.

Paper	Data Mining Method	Exhibit	$\Pr(\lvert t_i \rvert > 2)$
Yan and Zheng (2017)	Begin with 240 accounting variables apply 76 simple transformations, yielding 18,000 signals.	Table 1 Equal-Weighted	at least 20%
Chen, Lopez-Lira, Zimmermann (2023)	Begin with 242 accounting variables, apply all possible ratios and $\Delta x_1/\text{lag}(x_2)$, yielding 29,000 signals.	Table 4 Panel (a) Equal-Weighted	at least 20%
Chordia, Goyal, Saretto (2020)	Begin with 185 accounting variables, apply growth rates, ratios, and all combinations of $(x_1 - x_2)/x_3$, yielding 2.4 million signals.	Table A2	at least 20% in 12 out of 13 estimates

tors are false, while a more temperate approach would use data to bound $\Pr(F_i)$ (Benjamini and Hochberg (2000); Storey (2002)). Section 3 shows that these two effects largely cancel out.

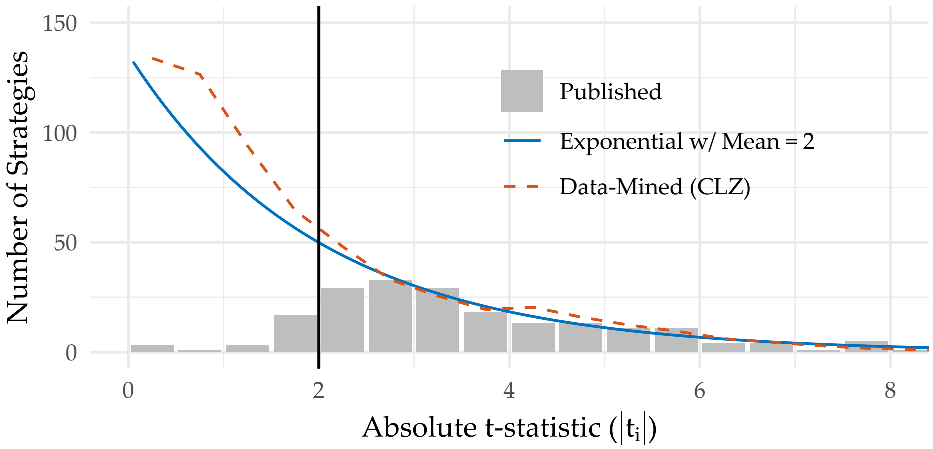
2.2. Bounding the FDR with Conservative Extrapolation

A second worst-case scenario can be constructed by conservative extrapolation. Figure 1 illustrates. It plots the distribution of t-stats from Chen and Zimmermann's (2022) open-source replications of 200 published predictors (bars) and overlays an exponential density with a mean of 2.0 (line). The plot shows that this exponential fits the observed region of the plot well.

In the unobserved region, the exponential extrapolates conservatively. A more moderate extrapolation might assume that the density peaks somewhere between 2 and zero, as it does with most distributions used in finance (e.g. log-normal). Instead, this extrapolation keeps increasing—it assumes there are exponentially more file drawer t-stats compared to published t-stats. Conservatism is also seen by comparing the exponential (solid line) with the data-mined distri-

Figure 1: Bounding the FDR using Conservative Extrapolation as a Worst Case

t_i tests the null that the mean long-short return is zero in a replication of a published predictor from Chen and Zimmermann (2022). Solid line extrapolates using an exponential distribution with a mean of 2.0. Dashed line overlays the distribution of t-stats from Chen, Lopez-Lira, and Zimmermann's (2024) equal-weighted data-mined strategies. The exponential is conservative: it assumes there are exponentially more file drawer t-stats than published t-stats and has a similar shape as the data-mined distribution. Plugging the exponential CDF into Equation (5) implies at least 86% of findings are true.



bution from Chen, Lopez-Lira, and Zimmermann (2024) (dashed line). To the left of 2.0, the two distributions are broadly similar, implying that the extrapolation is similar to the worst case of atheoretical data-mining.

Plugging an exponential CDF with $E(|t_i|) = 2.0$ into Equation (5) then bounds the FDR:

$$\text{FDR}_{|t|>2} \leq \frac{5\%}{\exp\left(-2/E\left(|t_i|\right)\right)} = 5\% \exp\left(1\right) = 13.6\%. \tag{7}$$

Thus, a very different estimate also finds that at least 85% of claimed findings are true.

A beautiful part of this FDR bound is that it can be calculated by hand, using summary statistics reported in published papers. The memoryless property of the exponential distribution implies

$$E\left(|t_i|\right) = E\left(|t_i| \,|\, |t_i| > 2\right) - 2, \tag{8}$$

that is, the mean t-stat can be estimated by just subtracting 2 from the mean

Table 2: FDR Bounds using Summary Stats from Predictor Zoo Studies

Table reports the mean published t-stat from various papers and the resulting $\text{FDR}_{|t|>2}$ bound (Equations (7)-(8)). t-stats use equal-weighting with the exception of Harvey, Liu, and Zhu (2016), which does not specify the weighting. These easy to use bounds consistently imply that most claimed findings are true.

Paper	Exhibit	Method	Mean Pub t-stat	$\text{FDR}_{|t|>2}$ bound (%)
Green, Hand, Zhang (2013)	Tables 3 and 4	Hand Collected	5.1	10
Chen, Zimmermann (2020)	Table 2	Hand Collected	4.6	11
Harvey, Liu, Zhu (2016)	Figure A.1 Panel B	Hand Collected	4.2	12
McLean, Pontiff (2016)	Page 16 , Par 1	Replicated In-Sample	3.6	18
Jacobs, Muller (2020)	Online App Table 2	Replicated Full-Sample	3.1	32

published t-stat, which is close to $E(|t_i|\,||t_i|>2)$. Thus, one can bound the FDR in the presence of publication bias, by just plugging in summary statistics.

Table 2 uses Equation (8) to bound $\text{FDR}_{|t|>2}$ in a variety of "predictor zoo" studies. For example, Chen and Zimmermann's (2020) Table 2 reports that the mean hand-collected t-stat across 77 published predictors is 4.6. Equation (8) then implies that $E(|t_i|) \approx 4.6 - 2 = 2.6$, and so Equation (7) implies $\text{FDR}_{|t|>2} \leq 5\% \exp(2/2.6) = 11\%$. Replications and post-publication data tend to result in lower t-statistics, leading to higher FDR bounds from McLean and Pontiff (2016) and Jacobs and Müller (2020).[3] Nevertheless, the results across five independent teams using various data collection methods are consistent. All imply that most claimed findings are true.

The entry for Harvey, Liu, and Zhu (2016) (HLZ) is notable because HLZ argue "that most claimed research findings in financial economics are likely false." Table 2 shows that the difference in conclusions does not arise from a difference in the underlying data. The mean published t-stat in HLZ is 4.2, not far from the mean of 4.0 in CZ. Based on Equation (7), a mean published t-stat of 4.2 implies $\text{FDR}_{|t|>2} \leq 12\%$.

[3] Chen and Zimmermann (2022) show that replicated t-statistics are as large as hand-collected ones, if the replications carefully follow the original methods.

In fact, even the methods are similar. HLZ's Figure A.1 also uses an exponential extrapolation. While this figure is buried in the appendix of HLZ, the same extrapolation is featured in Figure 1 of Harvey (2017). We will see that HLZ's own FDR estimates are consistent with the $\text{FDR}_{|t|>2} \leq 12\%$ (Section 5.1) and that the difference in conclusions comes from different interpretations of "false discovery" (Section 5.2).

For further details on the papers and exhibits in Table 2, see Appendix A.2.

In summary, the FDR is small because the research process readily generates large t-stats, much more readily than compared to a standard normal distribution. This property can be seen in the frequency of large t-stats in atheoretical data-mining studies or in conservative extrapolations of the large t-stats of published strategies. Similar results have been reported in previous studies, but these often involve structural estimation (e.g. Chen and Zimmermann (2020)). In contrast, my FDR bounds so simple they can be estimated by just plugging in numbers from summary statistic tables.

3. More Refined Estimates

One can argue that the easy FDR bounds are too conservative. Rather than make inference about $\Pr(F_i)$, they focus on the worst case of $\Pr(F_i) = 1.0$. But one can also argue that the previous bounds are too aggressive. They focus on raw equal-weighted returns, rather than value-weighted returns.

This section addresses both issues. Section 3.1 provides a visual approach to bounding $\Pr(F_i)$, following Storey (2002). Section 3.2 applies this method to value-weighted returns adjusted for various factor exposures.

Making inference on $\Pr(F_i)$ requires data on insignificant predictors (Chen (Forthcoming)). Thus, this section focuses on the CLZ data-mined predictors, and uses data-mining as a worst-case for the publication process, following Section 2.1.

3.1. A Tighter, Visual Bound

In essence, FDR methods decompose the empirical distribution into a false (null) component and a true (non-null) component. FDR methods are also conservative, so they generally consider cases in which the false component as large as possible. These facts motivate the following algorithm:

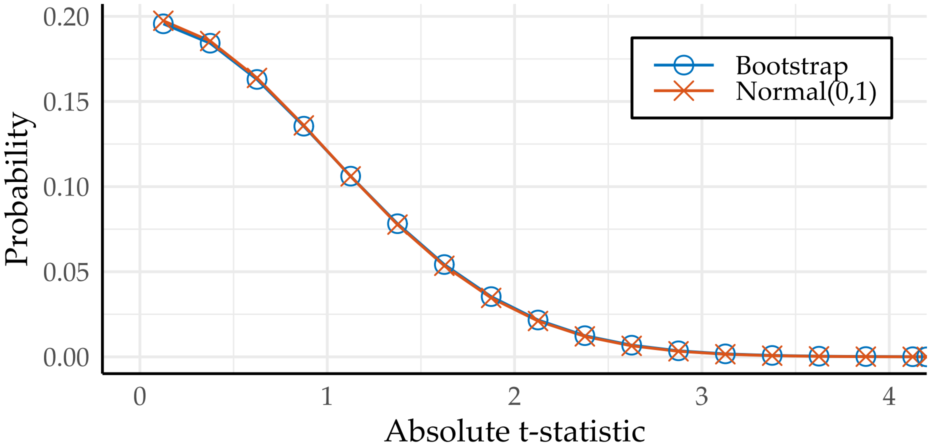
Algorithm 1 Visual FDR Bound

```
1. Plot the histogram of absolute t-stats for some data-mined strategies.
2. Draw the largest null distribution that still fits inside the data distribution.
3. Draw a vertical line at 2.0. To the right of this line, the ratio of the null area
   to total area is the "Visual Bound" on $\text{FDR}_{|t|>2}$.
```

The central limit theorem (CLT) implies that $\text{Normal}(0,1)$ is a good null (Chen (2021)). Figure 2 verifies this in my setting. It uses a cluster bootstrap (Fama and French (2010)) to simulate the null in a manner that captures both the fat tails in monthly returns and the potentially important cross-predictor correlations in the CLZ returns. The fat tails and correlations have almost no impact. The bootstrapped null (circles) is visually identical to $\text{Normal}(0,1)$ (crosses).

Figure 2: Cluster Bootstrap vs Standard Normal Null

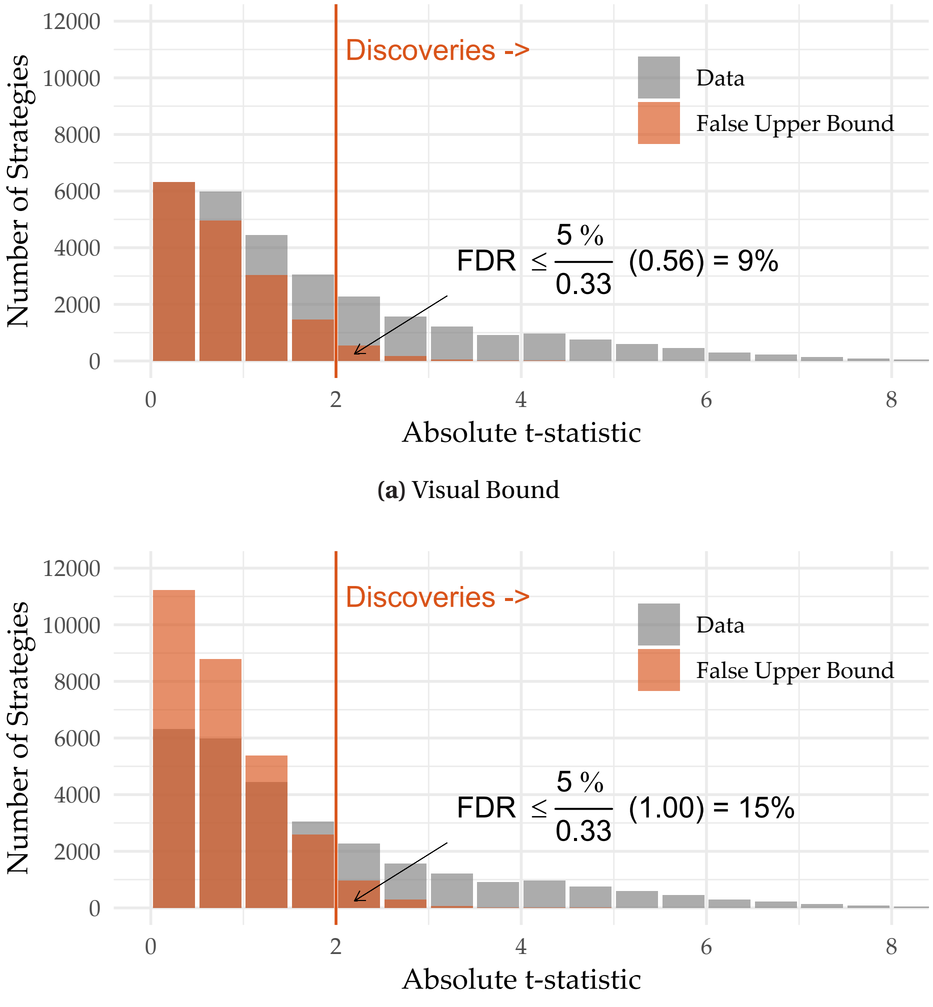
Bootstrap demeans CLZ equal-weighted returns, resamples months, calculates t-stats, and repeats 1,000 times (following Fama and French (2010)). The CLT holds and the standard normal is a good null.



Using $\text{Normal}(0,1)$ as the null, Panel (a) of Figure 3 applies Algorithm 1 to CLZ's equal-weighted raw returns (total bars). The false component (red) is chosen to be as large as possible, so it completely fills the first histogram bar. Nevertheless, the standard normal decays so much faster than the data, that to the right of 2.0, the red area is just 9% of the total area. Assuming that the research generates true predictability at at least the same rate as data mining, this visual bound implies *at least* 91% of claimed findings are true.

Figure 3: Bounding the FDR Visually, Following Storey (2002)

Panels decompose the CLZ equal-weighted raw t-stats (gray) into a false (red) and true (gray minus red) components. The false component is a scaled Normal$(0,1)$. Panel (a) uses Algorithm 1. Panel (b) uses the easy bound (Equation (5)). The easy bound is unreasonably conservative. A more reasonable decomposition implies $\text{FDR}_{|t|>2} \leq 9\%$.



(a) Visual Bound

(b) Easy Bound

3.1.1. Visual Justification for Algorithm 1

To describe Algorithm 1 more precisely, let b be a histogram bin. The law of total probability implies

$$\underbrace{\Pr(|t_d| \in b)}_{\text{Empirical Data (Total)}} = \underbrace{\Pr(|t_d| \in b|F_d)\Pr(F_d)}_{\text{False Component (Red)}} + \underbrace{\Pr(|t_d| \in b|T_d)\Pr(T_d)}_{\text{True Component (Remainder)}}, \tag{9}$$

where d is a data-mined predictor, t_d its t-stat, F_d is the event that d is false, and T_d is the event d is true (not false). This section uses the index d instead of i (which represents predictors from research) to emphasize that research should produce larger t-stats, and a smaller FDR, than data mining.

Equation (9) shows how FDR methods decompose the empirical histogram into false and true components. Moreover, Bayes rule provides the FDR for each bin:

$$\Pr(F_d||t_d| \in b) = \frac{\Pr(|t_d| \in b|F_d)\Pr(F_d)}{\Pr(|t_d| \in b)}. \tag{10}$$

In other words, the probability that a predictor in bin b is false is the ratio of the false component to the total. Then, averaging this ratio over bins to the right of $|t_d| = 2$ leads to $\text{FDR}_{|t|>2}$.

The key question is: how should one choose $\Pr(F_d)$? Equation (9) shows that $\Pr(F_d)$ is not only the proportion of false predictors overall, it is also the scaling of the false component. The visual bound chooses this scaling to be as conservative as is reasonable (Panel (a)). Any more conservative and the decomposition would imply that the number of false strategies could be greater than the number of total strategies in the first bin ($|t_d| \leq 0.5$).

This approach does not you what $\Pr(F_d)$ is—it only says $\Pr(F_d)$ is less than some number. As a result, the proper interpretation of the red bars is that they are an *upper bound* on the number of false predictors. Intuitively, some of the predictors with $|t_d| \leq 0.5$ could be true predictors that suffered from bad in-sample draws.

Instead of using the data to bound $\Pr(F_d)$, the easy bound simply chooses $\Pr(F_d) \leq 1.0$. The benefit of this choice is that it is easy to calculate and communicate. The downside is that it implies an unreasonably conservative decomposition, as seen in Panel (b). Plugging $\Pr(F_d) \leq 1.0$ into Equation (9) implies that the number of false strategies in the first bin ($|t_d| \leq 0.5$) could be as large as

11,000, far larger than the 6,000 or so strategies in the data. Equivalently, it implies that the number of true strategies could be as little as *negative* 5,000, which is nonsensical.

3.1.2. Algebraic Justification and Equivalence to Storey (2002)

Step 2 of Algorithm 1 can be formalized by solving Equation (9) for $\Pr(F_d)$ and noting that $\Pr(|t_d| \in b|T_d)\Pr(T_d) \geq 0$:

$$\Pr(F_d) = \frac{\Pr(|t_d| \in b) - \Pr(|t_d| \in b|T_d)\Pr(T_d)}{\Pr(|t_d| \in b|F_d)} \leq \frac{\Pr(|t_d| \in b)}{\Pr(|t_d| \in b|F_d)}. \tag{11}$$

Equation (11) provides many bounds on $\Pr(F_d)$, depending on the choice of bin b. Each bound chooses $\Pr(F_d)$ to be as conservative as is reasonable, given the data in that bin. Any more conservative, and we would be ignoring the constraint that $\Pr(|t_d| \in b|T_d)\Pr(T_d) \geq 0$. By considering the full distribution of the data, Step 2 minimizes across all bins, leading to the visual bound in Panel (a) of Figure 3.

Minimizing Equation (11) provides a "conservative estimate" of $\Pr(F_d)$, just as in the Storey (2002) method. In fact, Equation (11) is in a sense equivalent to Storey's Equation (6). Instead of examining t-stats and bins, Storey examines p-values p_d and a minimum p-value λ. Using p_d and λ, the numerator of the right-most expression in Equation (11) can be written as $\Pr(p_d > \lambda)$, which has the natural estimator $\#\{p_d > \lambda\}/M$, where M is the number of predictors in the data. The denominator can be written as $\Pr(p_d > \lambda|F_d) = (1-\lambda)$, since p-values are uniform under the null. Dividing $\#\{p_d > \lambda\}/M$ by $(1-\lambda)$ leads to Storey's Equation (6).

In Figure 3, the bin produces the smallest $\Pr(F_d)$ bound is $|t_d| \leq 0.5$. The chart shows that $\Pr(|t_d| \leq 0.5) \approx 6,300/29,300 = 21.5\%$. $\Pr(|t_d| \leq 0.5|F_d)$ is the classical "half sigma" probability: 38.3%. Combining, we have

$$\Pr(F_d) \leq \frac{\Pr(|t_d| \leq 0.5)}{\Pr(|t_d| \leq 0.5|F_d)} \approx \frac{0.215}{0.383} = 0.56, \tag{12}$$

showing that at most 56% of the 29,300 CLZ equal-weighted raw returns are false. Then plugging Equation (12) into Equation (3), we have the same FDR bound from the visual argument

$$\text{FDR}_{|t|>2} \leq \frac{\Pr(|t_d| > 2|F_d)}{\Pr(|t_d| > 2)} \frac{\Pr(|t_d| \leq 0.5)}{\Pr(|t_d| \leq 0.5|F_d)} \leq \frac{5\%}{0.33}(0.56) = 9\%, \qquad (13)$$

but derived algebraically.

The FDR bound of 9% is very close to the empirical Bayes shrinkage estimates of around 10% found in Chen and Zimmermann (2020) and Jensen, Kelly, and Pedersen (2023) (see also Chinco, Neuhierl, and Weber (2021)). It is also comparable to the point estimates found in Harvey, Liu, and Zhu (2016) (see Section 5.1). I provide simulation evidence supporting the validity of this bound in Section 4.2.

3.2. Adjustments for Liquidity and Factor Exposure

The previous estimates focus on equal-weighted portfolios, following the modal test in the cross-sectional literature. But about half of the papers also study value-weighted portfolios (Green, Hand, and Zhang (2013)), which can be though of as a simple liquidity adjustment. Similarly, the previous results focus on raw long-short returns, for simplicity. The literature also examines returns adjusted for various factor exposures.

Table 3 examines how these adjustments affect FDR bounds. The first four columns adjust equal-weighted returns for exposure to the CAPM, Fama-French 3, and Fama-French 3 + momentum factors. These adjustments have relatively little impact on the share of $|t_d| > 2$. Regardless of the factor adjustment, this share is roughly 1/3, as it is in the raw returns. Thus, the $\text{FDR}_{|t|>2}$ bounds are little changed, and lie around 9 percent (or a bit below).

Columns 4-6 examine value-weighted returns. This adjustment does make a difference. The share of $|t_d| > 2$ is much smaller, at around 16%, as one might expect from the many papers that document how predictability is concentrated in small, illiquid stocks (e.g Chen and Velikov (2023)). Thus, the easy FDR bound is larger, at around 0.05/0.16 = 0.30. Conversely, the share of $|t_d| < 0.5$ is much larger than in the equal-weighted case, leading to a larger bound on $\Pr(F_d)$ of around 0.80. But in the end, $\text{FDR}_{|t|>2}$ represents only a minority of the data. Indeed, using the visual bound and either the CAPM or FF3 model leads to $\text{FDR}_{|t|>2}$ of at most 21%.

Table 3: FDR Bounds with Liquidity and Factor Adjustments

Data consists of 29,000 long-short decile portfolios constructed by CLZ. $\Pr(|t| > 2)$ and $\Pr(|t| \leq 1)$ are the share of $|t| > 2$ and $|t| \leq 1$, respectively. "$\text{FDR}_{|t|>2}$ max Easy" is $5\% / \Pr(|t| > 2)$ (Equation (5)). "$\Pr(F)$ max" is $\Pr(|t| \leq 0.5)/0.38$ where 0.38 is the share implied by the null (Figure 2). "$\text{FDR}_{|t|>2}$ max Storey" multiplies "$\text{FDR}_{|t|>2}$ max Easy" by "$\Pr(F)$ max." After adjusting for liquidity and factor exposure, most statistical findings are still true.

	Equal-Weighted				Value-Weighted			
	Raw	CAPM	FF3	4-Fac	Raw	CAPM	FF3	4-Fac
$\Pr(\|t\| > 2)$	32.5	35.9	37.3	32.4	10.7	17.1	18.6	13.5
$\text{FDR}_{\|t\|>2}$ max Easy	15.4	13.9	13.4	15.4	46.7	29.2	26.9	37.2
$\Pr(\|t\| \leq 0.5)$	21.6	19.7	18.8	20.7	33.4	28.5	27.9	31.8
$\Pr(F)$ max	56.3	51.4	49.0	54.0	87.2	74.3	72.8	83.1
$\text{FDR}_{\|t\|>2}$ max Visual	8.7	7.2	6.6	8.3	40.8	21.7	19.6	30.9

4. More Rigorous Justifications

The FDR formulas in Sections 2 and 3 are much simpler than those found in the finance literature (e.g. Harvey, Liu, and Zhu (2016)). This section shows that the simple formulas are equivalent and provides simulation evidence of their validity.

4.1. Equivalence of FDR Definitions

HLZ's definition of the FDR looks more complicated than Equation (2). Following Benjamini and Hochberg (1995), HLZ define the FDR using:

$$\text{FDR} = E[\text{FDP}] \tag{14}$$

where FDP is the false discovery proportion, which in turn is defined as

$$\text{FDP} = \begin{cases} \frac{N_{0|r}}{R} & \text{if } R > 0 \\ 0 & \text{if } R = 0 \end{cases} \tag{15}$$

and where $N_{0|r}$ and R are the number of false discoveries and the number of discoveries, respectively. Implicit in Equations (14)-(15) is some definition of a discovery (e.g. predictor d is a discovery if $|t_d| > 2$).[4]

As shown by Storey and Tibshirani (2001) and Storey (2003), Equations (14) and (15) are equivalent to the simpler Equation (2) under weak dependence. To see the equivalence, assume that the following weak law of large numbers (WLLN) holds: as the number of predictors $\longrightarrow \infty$,

$$\frac{\text{number of false discoveries}}{\text{number of discoveries}} \xrightarrow{p} \Pr\left(F_d | d \text{ is a discovery}\right). \tag{16}$$

Then, assuming a positive number of discoveries, I can plug everything in and simplify:[5]

$$\begin{aligned} \text{FDR} &= E\left[\frac{\text{number of false discoveries}}{\text{number of discoveries}}\right] && (17) \\ &\xrightarrow{p} \Pr\left(F_d | d \text{ is a discovery}\right), && (18) \end{aligned}$$

as the number of predictors $\longrightarrow \infty$. Both of these expressions are intuitive. Equation (17) says the FDR is the expected share of discoveries that are false. Equation (18) says the FDR is the probability that a discovery is false. Finally, defining a discovery as predictors with $|t_d| > 2$ leads to Equation (2). This derivation is a simplified version of Storey and Tibshirani's (2001) Theorem 2 or Storey's (2003) Theorem 4.

The key assumption is the WLLN, Equation (16). It amounts to assuming that the method of moments works for probabilities. It is analogous to assuming that if you keep counting the share of economists who study finance, you will eventually recover the probability a randomly chosen economist studies finance.

This WLLN can formalized by placing weak cross-predictor dependence conditions on the underlying long-short returns (a la Wooldridge (1994)). The data on long-short returns suggest weak dependence: cross-predictor correlations cluster close to zero (Chen and Zimmermann (2022)). Section 4.2 examines this assumption using bootstrapped simulations.

The statistics literature features many other notions for the FDR. In addition to the Benjamini and Hochberg (1995) FDR, there is the "positive false discovery rate" (pFDR) and the "marginal false discovery rate," (mFDR) (Storey (2011)). There is also the "local fdr" (a.k.a. $\text{fdr}(z)$) and the tail probability counterpart

[4] Though I use the subscript d, the formulas in this section apply to both data-mined and research-produced predictors unless explicitly stated.

[5] The limit of the expectation in Equation (17) can be handled using the Portmanteau theorem.

(a.k.a. Fdr(z)) (Efron (2008)). All of these notions of the FDR are equivalent under the weak dependence in large sample settings (Benjamini (2008)).

4.2. Estimations on Data Mining Simulations

The FDR bounds in Section 2 and 3 can also be justified using simulations. This section examines simulations of data-mined returns. For simulations of published returns see Appendix B.2.

I simulate the long short return for predictor d in month τ using

$$r_{d,\tau} = \mu_d + \varepsilon_{d,k(\tau)} \tag{19}$$

where μ_d is the expected return, $k(\tau)$ is a random integer that depends on τ, and $\varepsilon_{d,k}$ is the de-meaned long-short return for predictor d in month k from the CLZ dataset. In other words, $\varepsilon_{d,k}$ is cluster-bootstrapped from the CLZ data in a way that ensures returns from the same month are drawn together. As a result, the simulated returns closely match the cross-predictor correlation structure (Appendix Figure A.4). For further details on the bootstrap, see Appendix B.

Expected returns follow

$$\mu_d = \begin{cases} 0 & \text{if } F_d \\ \gamma & \text{if } T_d \end{cases}, \tag{20}$$

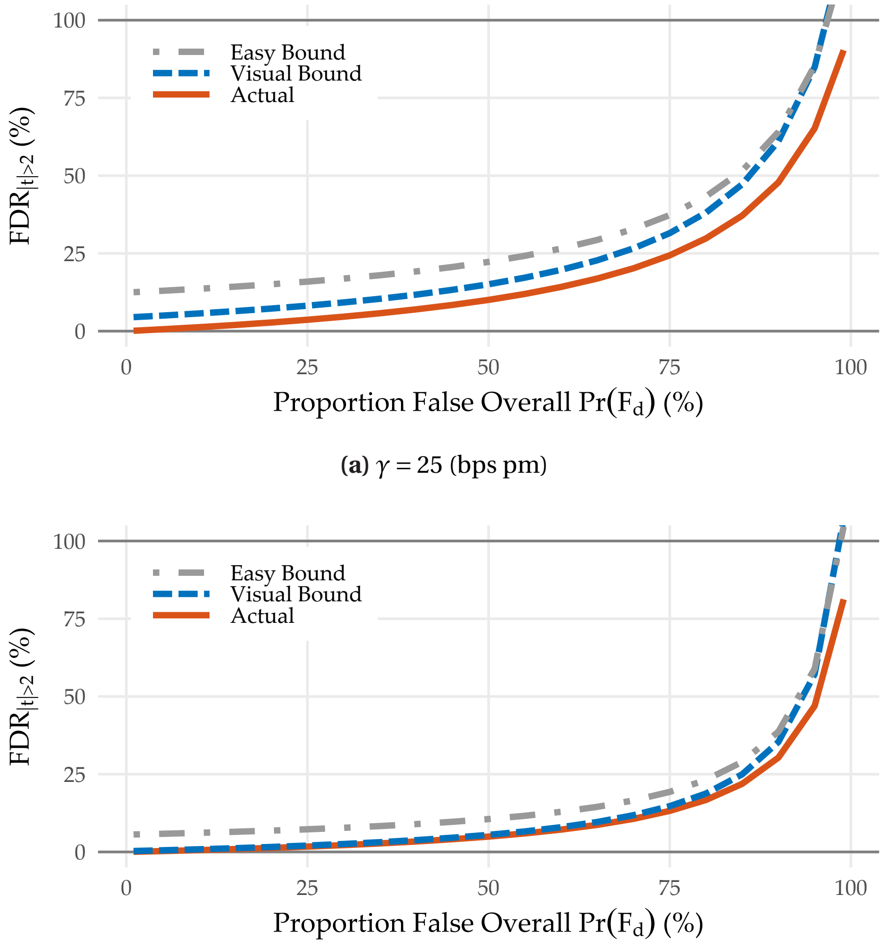
where γ is a constant. Equation (20) is the simplest structure that can underlie the FDR framework (Equations (14) and (15)). Each simulation has 29,000 predictors and 500 months. For each predictor I calculate t-stats the standard way (dividing the mean by the volatility and multiplying by $\sqrt{500}$). For comparison, the median number of monthly returns across the CLZ strategies is 594.

For each simulation, I compute the easy FDR bound (Equation (5)) and the visual bound (Algorithm 1). I compare these bounds with the actual $\text{FDR}_{|t|>2}$, which is computed by averaging the FDP (the ratio of false discoveries to discoveries) across simulations.

Figure 4 shows simulated estimation results for various choices for γ and $\Pr(F_d)$. Panel (a) shows γ = 25 bps per month, near the 10th percentile of in-sample returns in the CZ data. Panel (b) uses γ =75 bps, which is near the 70th percentile. These two choices for γ span what might be considered a reasonable range of values for the expected returns of true predictors. The figure examines values of $\Pr(F_d)$ ranging from 1% to 99%.

Figure 4: FDR Estimates in Cross-Sectional Predictability Simulations

I simulate long-short returns (Equation (19)-(20)) by cluster-bootstrapping from the CLZ strategies, computing FDR bounds, and comparing with the actual FDR. γ is the expected return of true predictors. "Easy Upper Bound" plugs in $\Pr(F_d) = 1$ into Bayes rule (Equation (5)). "Visual Bound" uses Algorithm 1. Both methods reliably bound the FDR.



(a) $\gamma = 25$ (bps pm)

(b) $\gamma = 75$ (bps pm)

Both methods reliably bound $\text{FDR}_{|t|>2}$. The visual bound comes closer to the actual FDR, and so provides a more accurate assessment of the veracity of the cross-sectional literature. Overall, these simulations show that the data mined FDR bounds in Sections 2 and 3 are valid for cross-sectional predictability.

4.3. Relationship with the Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001) algorithms

My estimates first define discoveries and then estimate an FDR bound. This method is known as the direct approach to FDRs (Storey (2002)). The finance literature focuses on FDR control instead, using the Benjamini and Hochberg (1995) (BH95) algorithm and Benjamini and Yekutieli's (2001) (BY's) Theorem 1.3.[6] These algorithms are the inverse of the direct approach: they first define an FDR bound and then estimate discoveries.

The two approaches are equivalent under weak dependence (Storey, Taylor, and Siegmund (2004); Farcomeni (2007)). The see the equivalence, note the BH95 algorithm can be written as

$$h^* = \min_{h>0}\left\{h : \widehat{\text{FDR}}_{\text{max,BH}}\left(|t_d| > h\right) \le q^*\right\}. \tag{21}$$

where h^* is the estimated t-stat hurdle, q^* is the FDR bound selected by the researcher, and

$$\widehat{\text{FDR}}_{\text{max,BH}}\left(|t_d| > h\right) \equiv \frac{\Pr\left(|t_d| > h|F_d\right)}{\text{Share of } |t_d| > h}. \tag{22}$$

Equation (22) generalizes the easy FDR bound (Equation (4)) to define discoveries as $|t_d| > h$ instead of $|t_d| > 2$. Thus, the BH95 algorithm is equivalent to choosing a t-stat hurdle that ensures the easy FDR bound is less than some prespecified value. For further details see Appendix C.

BY's Theorem 1.3 is equivalent to replacing $\widehat{\text{FDR}}_{\text{max,BH}}\left(|t_d| > h\right)$ in Equation (21) with

$$\widehat{\text{FDR}}_{\text{max,BY1.3}}\left(|t_d| > h\right) \equiv \frac{\Pr\left(|t_d| > h|F_d\right)}{\text{Share of } |t_d| > h} \times \sum_{j=1}^{M} \frac{1}{j},$$

where M is the total number of predictors. This expression amounts to plugging $\Pr\left(F_d\right) \le \sum_{j=1}^{M} \frac{1}{j} \approx \log M \gg 1.0$ into Bayes rule (Equation (3)). My methods omit this penalty, as Section 3.1 argues $\Pr\left(F_d\right) \le 1.0$ is already unreasonably conservative and Section 4.2 shows the tighter visual bound is reliable for cross-sectional predictability settings. Indeed, the original paper Benjamini and Yekutieli (2001)

[6] I specify the theorem number 1.3 from Benjamini and Yekutieli (2001) because the thrust of that paper regards Theorem 1.2.

describes Theorem 1.3 as "very often unneeded, and yields too conservative of a procedure."

5. Reconciliation and Interpretation

My results appear to conflict with HLZ, who "argue that most claimed research findings in financial economics are likely false." This section shows that, despite the apparent conflict, our FDR estimates are actually quite similar. It also argues in favor of my interpretation of these estimates.[7]

5.1. FDR Estimates in Harvey, Liu, and Zhu (2016)

HLZ shows empirical FDR estimates in two exhibits: Figure 3 (page 25) and Table 5 (page 34). These exhibits use different methods but arrive at similar numbers.

Figure 3 uses Benjamini and Yekuteli's (2001) Theorem 1.3 to estimate t-stat hurdles that imply FDR $\leq$ 1% using hand-collected data on published t-stats. It shows that a hurdle of 3.39 is required. This estimate implies that findings with $|t_i| > 3.39$ are almost certainly true, but says little about findings that fall below this cutoff. A more complete picture can be found in the text describing Figure 3, which states that the same method implies a hurdle of 2.81 if the FDR bound is raised to 5% (page 26, top paragraph).

The natural question, then, is how many t-stats in asset pricing papers meet the 2.81 hurdle? Perhaps surprisingly, this answer is not directly found in HLZ. But a closely related number is reported Chen (2021)'s Table 1, which simulates HLZ's SMM estimates to approximate their dataset. Chen finds that, of the $|t_i|$ that exceed 2.0, 69% also exceed 3.0. Thus, the FDR among $|t_i| > 2.0$ in HLZ's data is at *most* $0.69 \times 0.05 + (1 - 0.69) = 34\%$. This number is higher than the FDR bounds in Section 2, but it is qualitatively similar. In the end, both HLZ's Figure 3 and my analyses find that a minority of published findings are false, at most.

HLZ's Table 5 uses an SMM estimation based on a structural model that includes both published and unpublished t-stats (see Appendix D). This table shows a variety of estimates but the authors argue in favor of the "$\rho = 0.2$" esti-

[7] Some of the material in this section appears in the review article Chen and Zimmermann (2023), which was circulated after the first draft of this paper.

mate (page 35, last paragraph). This estimate says that choosing an FDR of 5% implies a hurdle of 2.27. Given that this threshold is close to 2.0, this estimate suggests that the FDR for published findings is also close to 5%. Indeed, of t-stats in the Chen and Zimmermann (2022) open-source dataset that exceed 2.0, 91% also exceed 2.27. This implies an FDR of at most $0.91(0.05) + 0.09 = 14\%$, exactly in line with Equation (7).

Figure 5 takes a closer look at HLZ's Table 5. Panel (a) shows a scatterplot of expected returns against absolute t-statistics from simulating HLZ's estimated model, with each marker representing one factor. HLZ's model features two kinds of factors: (1) factors with expected returns of zero (empty markers) and (2) factors with positive expected returns (filled markers). A complete description of HLZ's model can be found in Appendix D.

Following the traditional notion of a false discovery (Sorić (1989)), I label the markers with expected returns of zero as false. This labeling leads to a neat visualization of HLZ's Table 5. If we define discoveries as predictors with $|t_i| > 2.95$ (dot-dash line), we have about 210 discoveries. Only two of these factors are false, leading to an FDR of $2/210 \approx 1\%$, consistent with HLZ's famous 3.0 hurdle and the "FDR(1%)" column of their Table 5. If instead, we define discoveries using $|t_i| > 2.27$ (long-dash line), we have 305 discoveries, 13 false discoveries, and an FDR of $13/305 \approx 5\%$, consistent with the "FDR(5%)" column of their Table 5.

What is the FDR if we just define discoveries as $|t_i| > 2$? The figure shows that, in this case, the FDR is just 9%. In other words, about 91% of claimed findings in HLZ's dataset are true. This estimate is closely in line with the $\text{FDR}_{|t|>2}$ bound of 9% I derive following Storey (2002) (Section 3.1).

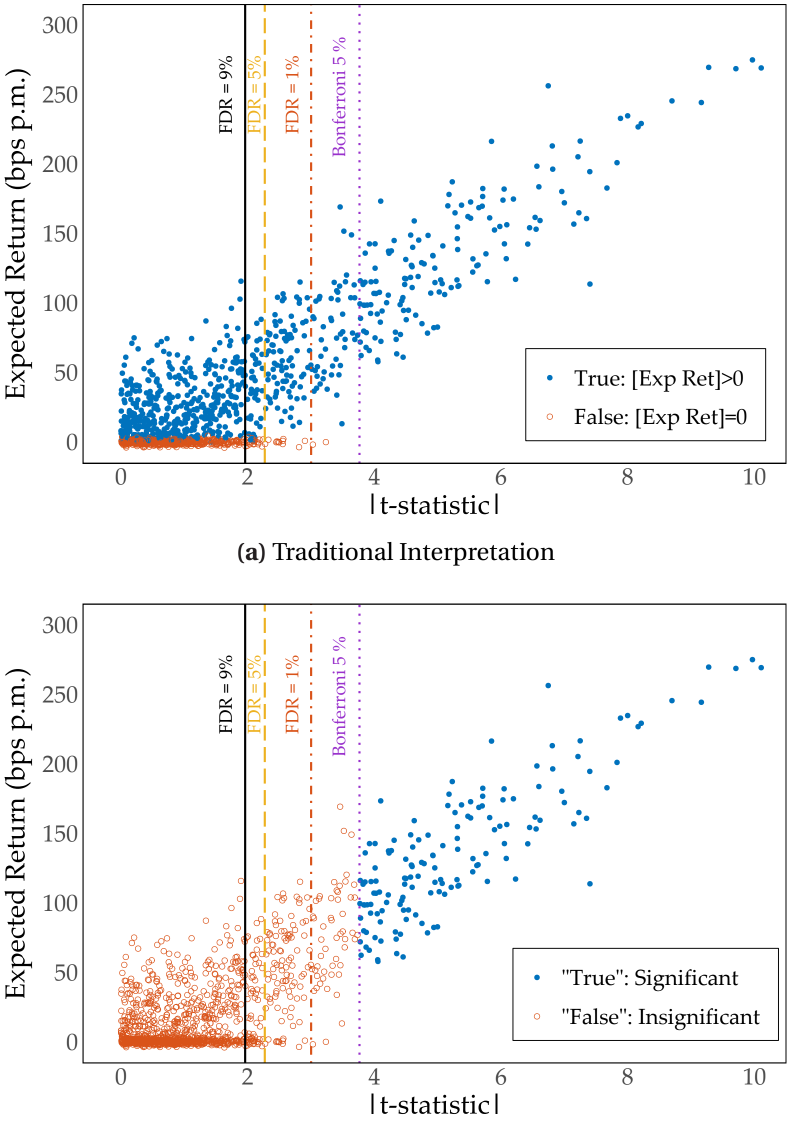
5.2. Two Interpretations of "False Factor"

If HLZ's estimates imply $\text{FDR}_{|t|>2} \approx 9\%$, why do they "argue that most claimed research findings in financial economics are likely false?" The answer can be found in their conclusion (page 37, paragraph 5):

> *In medical research, the recognition of the multiple testing problem has led to the disturbing conclusion that "most claimed research findings are false" (Ioannidis (2005)). Our analysis of factor discoveries leads to the same conclusion–many of the factors discovered in the field of finance are likely false discoveries: of the 296 published signif-*

Figure 5: Two Interpretations of Harvey, Liu, and Zhu's (2016) SMM Estimates

I simulate HLZ's SMM estimates (HLZ's Table 5, $\rho = 0.2$). Each marker represents one factor. "FDR = 5%" and "FDR = 1%" show t-hurdles from HLZ's Table 5. "Bonferroni 5%" uses the calculation implied by HLZ's conclusion (HLZ page 37 paragraph 5). Panel (a) defines false factors following the traditional interpretation (Sorić (1989)). Panel (b) uses an interpretation from HLZ's conclusion. HLZ's argument that "most claimed research findings in financial economics are likely false" comes from equating significance with truth. This interpretation leads to factors with expected returns of 75 bps or more being defined as false.



(a) Traditional Interpretation

(b) Harvey, Liu, Zhu's (2016) Interpretation

> *icant factors, 158 would be considered false discoveries under Bonferonni [sic], 142 under Holm, 132 under [BY's Theorem 1.3] (1%), and 80 under [BY's Theorem 1.3] (5%).*

Traditionally, the Bonferroni correction provides a bound on the family wise error rate, which in turn is equal to the probability that the FDR is positive. It does not provide the number of false discoveries. The same is true of the Holm (1979) test, which is a variant of Bonferroni. In fact, BY's Theorem 1.3 does not provide the number of false discoveries either. It can provide an upper bound—but this bound may be far from the actual number (Section 4.3).

These issues make this paragraph difficult to understand. But the paragraph is straightforward if one interprets a "false factor" as an "insignificant factor under a chosen hypothesis test." With Bonferroni at the 5% level as the chosen test, the 158 insignificant factors are also false factors. 158/296 = 53%, which can be described as most, so most claimed research findings are false.

HLZ's conclusion paragraph also equates "most" and "many." Of the five FDRs described in this paragraph, only 158/296 could be described as most, using the traditional interpretation of most. Notably, 80/296 = 27% is rarely described as most.

The equating of "null" and "insignificant" is also found in HLZ. In describing their multiple testing framework, HLZ state "[i]n a factor testing exercise, the typical null hypothesis is that a factor is not significant. Therefore, a factor being insignificant means the null hypothesis is 'true'" (page 13, 2nd paragraph). A similar language is found further down on page 13: "under the null that all factors are insignificant" (3rd paragraph) and "the type I error in multiple testing is also related to false discoveries, by which we conclude a factor is 'significant' when it is not" (4th paragraph). The equating of false and insignificant is also found in Harvey and Liu (2020) (see Appendix E).

In contrast, the statistics literature defines a false discovery as a significant finding that is, in truth, null (Sorić (1989); Benjamini and Hochberg (1995)). This definition displays an important separation between significance and truth. No matter how well-designed a significance test is, it can still miss the truth, declaring nulls as significant, or non-nulls as insignificant. This separation is essentially "Principal 2" of the American Statistical Association's "Statement on Statistical Significance and P-Values" (Wasserstein and Lazar (2016)). This separation also has a long tradition, going back to Fisher (1935).

5.3. The Problems with Equating Significance and Truth

One might argue for HLZ's interpretation by pointing to multiple testing problems. These problems imply that the classical tests used throughout the literature are invalid. Thus, it is natural to describe published factors that fail multiple testing controls as "false discoveries."

The problems with this interpretation are illustrated in Panel (b) of Figure 5. This panel shows the same simulation as in Panel (a), but instead of using the traditional interpretation, it interprets false discoveries as factors that are insignificant under HLZ's chosen Bonferroni test (circles). HLZ choose a significance level of 5%, implying a hurdle of $|t_i| > 3.8$.

Factors with expected returns of 75 or even 150 bps per month fail to meet this hurdle. As a result, they are labeled false, despite their impressive expected returns. These are not sample mean returns, which may be due to luck, but the fundamental expected returns, cured of sampling error (see Equations (27) and (28) in Appendix D). Thus, HLZ's interpretation leads to absurd economic conclusions. The idea that a factor with a fundamental expected return of 150 bps per month (18% per year) is false is nonsensical.

Panel (b) illustrates an additional problem. The figure shows many significance hurdles. Which hurdle should be the one that defines a false discovery? The FDR = 5% hurdle (dashed line) would lead to many more factors being declared true compared to the Bonferroni 5% hurdle (dotted line). FDR 5% is also perhaps the natural hurdle, since FDR ≤ 5% is commonly used in fields like genetics and functional imaging (Benjamini (2020)). But this is not the hurdle that HLZ's conclusion emphasizes. Instead, it emphasizes the extremely stringent Bonferroni 5% hurdle, resulting in the headline statement that most claimed findings are false.

Thus, equating significance and truth could lead to a kind of "post-truth finance," where senior scholars decide which hurdles define the truth. Instead of seeking the truth, junior researchers would seek agreement with the truth defined by senior scholars. Since multiple testing controls can be arbitrarily strict (why not a Bonferroni 2% correction?), senior scholars can in principle declare *any* factor as false. While journal editors and chaired professors wield significant power in the current paradigm, equating significance with truth would bring this power to a new level. If significance is equated with truth, then senior scholars will not only decide which factors are interesting or economically meaningful.

They will literally decide which factors are true.

Panel (b) of Figure 5 also illustrates the problem with equating “most” and “many.” Of the many hurdles in Panel (b), only a single hurdle finds that most claimed findings are false, using HLZ’s interpretation. In fact, the more commonly used FDR = 5% control (dashed line) would lead a rather small share of false discoveries, as seen in the relatively small number of factors lying between the solid and dashed lines. Equating most with many, as is done in HLZ’s conclusion, implies ignoring this important evidence.

5.4. Economic Significance

Statistical significance does not imply economic significance. Similarly, estimates of a small FDR do not imply that most claimed findings are economically meaningful. Could it be that HLZ’s argument is really saying that most claimed findings are economically meaningless?

Panel (a) of Figure 5 shows that the answer is no. In HLZ’s own estimates, most factors that have t-stats greater than 2.0 have noteworthy expected returns. Among these factors, the mean expected return is 89 bps per month and 91% of factors have mean returns in excess of 25 bps.

Estimates throughout the literature support this economic significance. Using empirical Bayes, several papers find that the typical sample mean return is at least 80% due to expected returns, even after accounting for multiple testing (Chen and Zimmermann (2020); Jensen, Kelly, and Pedersen (2023); Chinco, Neuhierl, and Weber (2021); Chen (Forthcoming)). Out-of-sample tests find that 75% of mean returns persist in the first few years after the original samples end (McLean and Pontiff (2016); Jacobs and Müller (2020); Chen and Zimmermann (2020)). Notable out-of-sample returns are even found in data-mined strategies. By simply sorting on their past returns, Yan and Zheng (2017) and Chen, Lopez-Lira, and Zimmermann (2024) identify thousands of data-mined strategies with out-of-sample returns of 50 bps per month or more.

These results focus on equal-weighted portfolios, which arguably overstate economic significance due to trading costs (Chen and Velikov (2023)). However, these results also focus on simple-minded portfolio sorts, which ignore the economic significance that comes from combining predictors (DeMiguel et al. (2020)). Indeed, Jensen et al. (2022) and Simon, Weibels, and Zimmermann

(2022) find that combining predictors with machine learning leads to notable returns net of trading costs, even in recent years.

6. Conclusion

This paper provides simple arguments that bound the FDR under publication bias. They handle publication bias by using data mining experiments or conservative extrapolations as worst case scenarios. The arguments are so simple that they can be explained in just a few equations or a couple of diagrams. Yet simulations show that they are valid in settings that mimic the cross-predictor correlations found in empirical asset pricing.

These methods suggest a way forward for the literature on multiple testing problems. This literature features complicated methods that take dozens of pages to exposit, which likely contribute to the lack of consensus in the literature. My estimates provide a transparent language for understanding and debating these issues.

Applying these methods to numbers reported in many previous papers, I find that at least 75% of claimed findings in cross-sectional predictability are true. More refined estimates using Chen, Lopez-Lira, and Zimmermann's (2024) data-mined predictors find that at least 91% of claimed findings are true.

Since peer review should outperform data mining, some may wonder if the share of true research findings is much higher than 91%. Unfortunately, other results in Chen, Lopez-Lira, and Zimmermann (2024) suggest that this bound is tight. Chen et al. compare the returns of published and data-mined predictors, and find that their post-sample performance differs by just 1-2 basis points per month. Thus, even if most claimed *statistical* findings (e.g., estimates documenting predictability) are true, the veracity of most *textual* claims (e.g., interpretations and conclusions) may be questionable.

References

Andrews, Isaiah and Maximilian Kasy (2019). “Identification of and correction for publication bias”. *American Economic Review* 109.8, pp. 2766–94.

Benjamini, Yoav (2008). “Comment: Microarrays, empirical Bayes and the two-groups model”. *Statistical Science* 23.1.

— (2010). “Discovering the false discovery rate”. *Journal of the Royal Statistical Society: series B (statistical methodology)* 72.4, pp. 405–416.

— (2020). “Selective inference: The silent killer of replicability”.

Benjamini, Yoav and Yosef Hochberg (1995). “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300.

— (2000). “On the adaptive control of the false discovery rate in multiple testing with independent statistics”. *Journal of educational and Behavioral Statistics* 25.1, pp. 60–83.

Benjamini, Yoav and Daniel Yekutieli (2001). “The control of the false discovery rate in multiple testing under dependency”. *Annals of statistics*, pp. 1165–1188.

Chen, Andrew Y (2021). “The Limits of p-Hacking: Some Thought Experiments”. *The Journal of Finance.*

— (Forthcoming). “Do t-Statistic Hurdles Need to be Raised”. *Management Science.*

Chen, Andrew Y, Alejandro Lopez-Lira, and Tom Zimmermann (2024). “Does peer-reviewed research help predict stock returns?” *arXiv preprint arXiv:2212.10317.*

Chen, Andrew Y and Mihail Velikov (2023). “Zeroing in on the expected returns of anomalies”. *Journal of Financial and Quantitative Analysis* 58.3, pp. 968–1004.

Chen, Andrew Y and Tom Zimmermann (2020). “Publication bias and the cross-section of stock returns”. *The Review of Asset Pricing Studies* 10.2, pp. 249–289.

— (2023). “Publication Bias in Asset Pricing Research”. *Oxford Research Encyclopedia of Economics and Finance.*

— (2022). “Open Source Cross-Sectional Asset Pricing”. *Critical Finance Review* 27.2, pp. 207–264.

Chinco, Alex, Andreas Neuhierl, and Michael Weber (2021). “Estimating the anomaly base rate”. *Journal of financial economics* 140.1, pp. 101–126.

Chordia, Tarun, Amit Goyal, and Alessio Saretto (2020). “Anomalies and false rejections”. *The Review of Financial Studies* 33.5, pp. 2134–2179.

DeMiguel, Victor, Alberto Martin-Utrera, Francisco J Nogales, and Raman Uppal (2020). “A transaction-cost perspective on the multitude of firm characteristics”. *The Review of Financial Studies* 33.5, pp. 2180–2222.

Efron, Bradley (2008). “Microarrays, empirical Bayes and the two-groups model”. *Statistical science* 23.1, pp. 1–22.

Fama, Eugene F and Kenneth R French (2010). “Luck versus skill in the cross-section of mutual fund returns”. *The journal of finance* 65.5, pp. 1915–1947.

Farcomeni, Alessio (2007). “Some results on the control of the false discovery rate under dependence”. *Scandinavian Journal of Statistics* 34.2, pp. 275–297.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Green, Jeremiah, John RM Hand, and X Frank Zhang (2013). “The supraview of return predictive signals”. *Review of Accounting Studies* 18.3, pp. 692–730.

Harvey, Campbell R (2017). “Presidential address: The scientific outlook in financial economics”. *The Journal of Finance* 72.4, pp. 1399–1440.

Harvey, Campbell R and Yan Liu (2020). “False (and missed) discoveries in financial economics”. *The Journal of Finance* 75.5, pp. 2503–2553.

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

Holm, Sture (1979). “A simple sequentially rejective multiple test procedure”. *Scandinavian journal of statistics*, pp. 65–70.

Ioannidis, John PA (2005). “Why most published research findings are false”. *PLoS medicine* 2.8, e124.

Jacobs, Heiko and Sebastian Müller (2020). “Anomalies across the globe: Once public, no longer existent?” *Journal of Financial Economics* 135.1, pp. 213–230.

Jensen, Theis Ingerslev, Bryan Kelly, and Lasse Heje Pedersen (2023). “Is there a replication crisis in finance?” *The Journal of Finance* 78.5, pp. 2465–2518.

Jensen, Theis Ingerslev, Bryan T Kelly, Semyon Malamud, and Lasse Heje Pedersen (2022). “Machine learning and the implementable efficient frontier”. *Swiss Finance Institute Research Paper* 22-63.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh (2018). “Interpreting factor models”. *The Journal of Finance* 73.3, pp. 1183–1223.

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

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

Reiner-Benaim, Anat (2007). “FDR control by the BH procedure for two-sided correlated tests with implications to gene expression data analysis”. *Biometrical Journal* 49.1, pp. 107–126.

Simon, Frederik, Sebastian Weibels, and Tom Zimmermann (2022). “Deep parametric portfolio policies”. *Available at SSRN 4150292*.

Sorić, Branko (1989). “Statistical “discoveries” and effect-size estimation”. *Journal of the American Statistical Association* 84.406, pp. 608–610.

Storey, John D (2002). “A direct approach to false discovery rates”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.3, pp. 479–498.

— (2003). “The positive false discovery rate: a Bayesian interpretation and the q-value”. *The Annals of Statistics* 31.6, pp. 2013–2035.

— (2011). “False Discovery Rate.” *International encyclopedia of statistical science* 1, pp. 504–508.

Storey, John D, Jonathan E Taylor, and David Siegmund (2004). “Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66.1, pp. 187–205.

Storey, John D and Robert Tibshirani (2001). *Estimating false discovery rates under dependence, with applications to DNA microarrays*. Tech. rep. Technical Report 2001-28, Department of Statistics, Stanford University.

Wasserstein, Ronald L and Nicole A Lazar (2016). *The ASA statement on p-values: context, process, and purpose*.

Wooldridge, Jeffrey M (1994). “Estimation and inference for dependent processes”. *Handbook of econometrics* 4, pp. 2639–2738.

Yan, Xuemin Sterling and Lingling Zheng (2017). “Fundamental Analysis and the Cross-Section of Stock Returns: A Data-Mining Approach”. *The Review of Financial Studies* 30.4, pp. 1382–1423.

A. Bounding the FDR by plugging in numbers from previous papers

A.1. FDR bounds from plugging in data mining numbers

This section describes the data mining procedures and t-statistics for the papers listed in Table 1.

A.1.1. FDR Bounds using Yan and Zheng (2017)'s (YZ's) Table 1

YZ choose 240 Compustat accounting variables based on data availability requirements (e.g. non-missing values in at least 20 years). They also choose 15 intuitive base variables (e.g. total assets, sales). They then form all combinations of X/Y, $\Delta(X/Y)$, $\%\Delta(X/Y)$, $\%\Delta X$, $\%\Delta X - \%\Delta Y$, and $\Delta X/\text{lag}(Y)$, where X is one of the 240 variables and Y is one of 15 base variables, leading to 18,000 signals. These functional forms are selected based on a survey of accounting textbooks and academic papers. For each variable, YZ form long-short decile portfolios.

Figure A.1: Yan and Zheng's Table 1

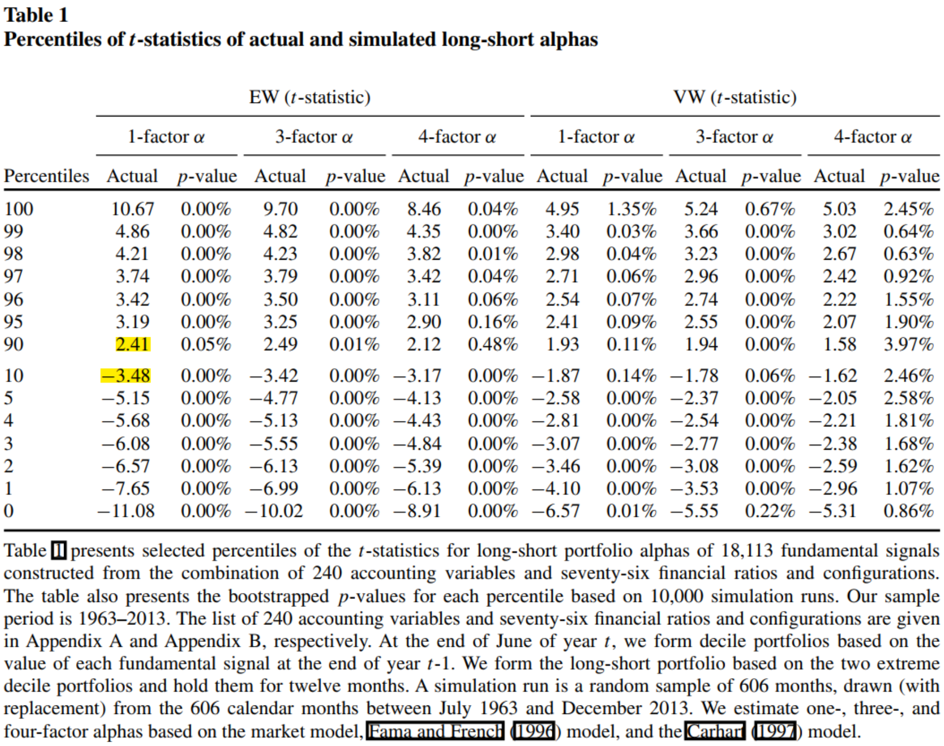


Table 1
Percentiles of t-statistics of actual and simulated long-short alphas

	EW (t-statistic)						VW (t-statistic)					
	1-factor α		3-factor α		4-factor α		1-factor α		3-factor α		4-factor α	
Percentiles	Actual	p-value	Actual	p-value	Actual	p-value	Actual	p-value	Actual	p-value	Actual	p-value
100	10.67	0.00%	9.70	0.00%	8.46	0.04%	4.95	1.35%	5.24	0.67%	5.03	2.45%
99	4.86	0.00%	4.82	0.00%	4.35	0.00%	3.40	0.03%	3.66	0.00%	3.02	0.64%
98	4.21	0.00%	4.23	0.00%	3.82	0.01%	2.98	0.04%	3.23	0.00%	2.67	0.63%
97	3.74	0.00%	3.79	0.00%	3.42	0.04%	2.71	0.06%	2.96	0.00%	2.42	0.92%
96	3.42	0.00%	3.50	0.00%	3.11	0.06%	2.54	0.07%	2.74	0.00%	2.22	1.55%
95	3.19	0.00%	3.25	0.00%	2.90	0.16%	2.41	0.09%	2.55	0.00%	2.07	1.90%
90	2.41	0.05%	2.49	0.01%	2.12	0.48%	1.93	0.11%	1.94	0.00%	1.58	3.97%
10	−3.48	0.00%	−3.42	0.00%	−3.17	0.00%	−1.87	0.14%	−1.78	0.06%	−1.62	2.46%
5	−5.15	0.00%	−4.77	0.00%	−4.13	0.00%	−2.58	0.00%	−2.37	0.00%	−2.05	2.58%
4	−5.68	0.00%	−5.13	0.00%	−4.43	0.00%	−2.81	0.00%	−2.54	0.00%	−2.21	1.81%
3	−6.08	0.00%	−5.55	0.00%	−4.84	0.00%	−3.07	0.00%	−2.77	0.00%	−2.38	1.68%
2	−6.57	0.00%	−6.13	0.00%	−5.39	0.00%	−3.46	0.00%	−3.08	0.00%	−2.59	1.62%
1	−7.65	0.00%	−6.99	0.00%	−6.13	0.00%	−4.10	0.00%	−3.53	0.00%	−2.96	1.07%
0	−11.08	0.00%	−10.02	0.00%	−8.91	0.00%	−6.57	0.01%	−5.55	0.22%	−5.31	0.86%

Table 1 presents selected percentiles of the t-statistics for long-short portfolio alphas of 18,113 fundamental signals constructed from the combination of 240 accounting variables and seventy-six financial ratios and configurations. The table also presents the bootstrapped p-values for each percentile based on 10,000 simulation runs. Our sample period is 1963–2013. The list of 240 accounting variables and seventy-six financial ratios and configurations are given in Appendix A and Appendix B, respectively. At the end of June of year t, we form decile portfolios based on the value of each fundamental signal at the end of year t-1. We form the long-short portfolio based on the two extreme decile portfolios and hold them for twelve months. A simulation run is a random sample of 606 months, drawn (with replacement) from the 606 calendar months between July 1963 and December 2013. We estimate one-, three-, and four-factor alphas based on the market model, Fama and French (1996) model, and the Carhart (1997) model.

YZ's Table 1 (shown in Figure A.1) shows order statistics for t-statistics measured using various methods. For EW 1-factor α's, the 90th and 10th percentile t-statistics are +2.41 and -3.48, respectively (highlighted). This means that at least 20 percent of absolute t-stats exceed 2. Applying to Equation (3):

$$\text{FDR}_{|t|>2} \leq \frac{5\%}{\Pr(|t_i|>2)} \leq \frac{5\%}{0.20} = 25\%.$$

EW 3- and 4-factor α's lead to the same result.

Of course, this bound may be much tighter if one had more detailed order statistics. -3.48 is quite far from -2.0, implying that one can form a much tighter bound. Section 2.1 fills this gap using the CLZ data, and finds $\text{FDR}_{|t|>2} \leq 15\%$.

For VW α's, one needs to use a slightly more general form of this bound. The 1-factor α's 90th and 10th percentile t-statistics are 1.93 and -1.87, respectively, so one needs to use the more general form

$$\text{FDR}(|t_i|>1.87) \leq \frac{\Pr(|t_i|>1.87|F_i)}{\Pr(|t_i|>1.87)} = \frac{6\%}{0.20} = 30\%.$$

Since the FDR decreases in the t-stat hurdle, $\text{FDR}_{|t|>2} \leq \text{FDR}(|t_i| > 1.87) \leq 30\%$. So even using VW portfolios, one still finds that most data-mined findings with $|t_i| > 2$ are true.

The picture changes slightly with VW 4-factor α's. Here the table shows that $\Pr(|t_i| > 2) \approx 10\%$, implying $\text{FDR}_{|t|>2} \leq 50\%$. Nevertheless, for most of the VW results, the Equation (3) implies that most claimed statistical findings are true. Moreover, these bounds come from an atheoretical data mining exercise, and researchers should be able to obtain a smaller FDR. Indeed, Section 3.2 shows that the Chen and Zimmermann (2023) and Chen, Lopez-Lira, and Zimmermann (2024) data mining procedure leads to an FDR of at most 34% when using 4-factor value-weighted mean returns.

A.1.2. FDR Bounds using Chen, Lopez-Lira, and Zimmermann (2024)'s Table 4

CLZ study 29,000 accounting signals. Table 4 forms long-short portfolios on these signals, calculates t-statistics for each portfolio using the past 30 years of returns, sorts portfolios into 5 bins based on the past 30 year return, and then averages t-statistics within each bin and across bin-formation years.

The equal-weighted panel shows that the mean t-stat for in-sample bin 2 is -2.46. Assuming volatility is relatively stable across portfolios, this implies that the maximum t-stat for bin 1 is at most -2.46, so at least 20% of signals produce $|t_i| > 2$. Plugging this result into Equation (5) implies $\text{FDR}_{|t|>2} \leq 5\%/0.20 = 25\%$.

The table also shows value-weighted results, but the statistics shown are not refined enough to estimate the share of $|t_i| > 2$. Section 3.2 examines this issue using the full CLZ dataset, and finds that the $\text{FDR}_{|t|>2}$ for value-weighted returns is typically less than 30%.

A.1.3. FDR Bounds using Chordia, Goyal, and Saretto (2020)'s Tables

CGS begin with 185 accounting variables and then apply the following transformations: (1) annual growth rates of a single variable, (2) ratios of two variables, and (3) difference ratios involving three variables as in $(x_1 - x_2)/x_3$. They examine a wide variety of predictability t-stats, leading to numerous reports of the $|t_i| > 1.96$ in Tables 3, 7, 8, and A2, among other locations.

Figure A.2: Chordia, Goyal, and Saretto's (2020) Table A2

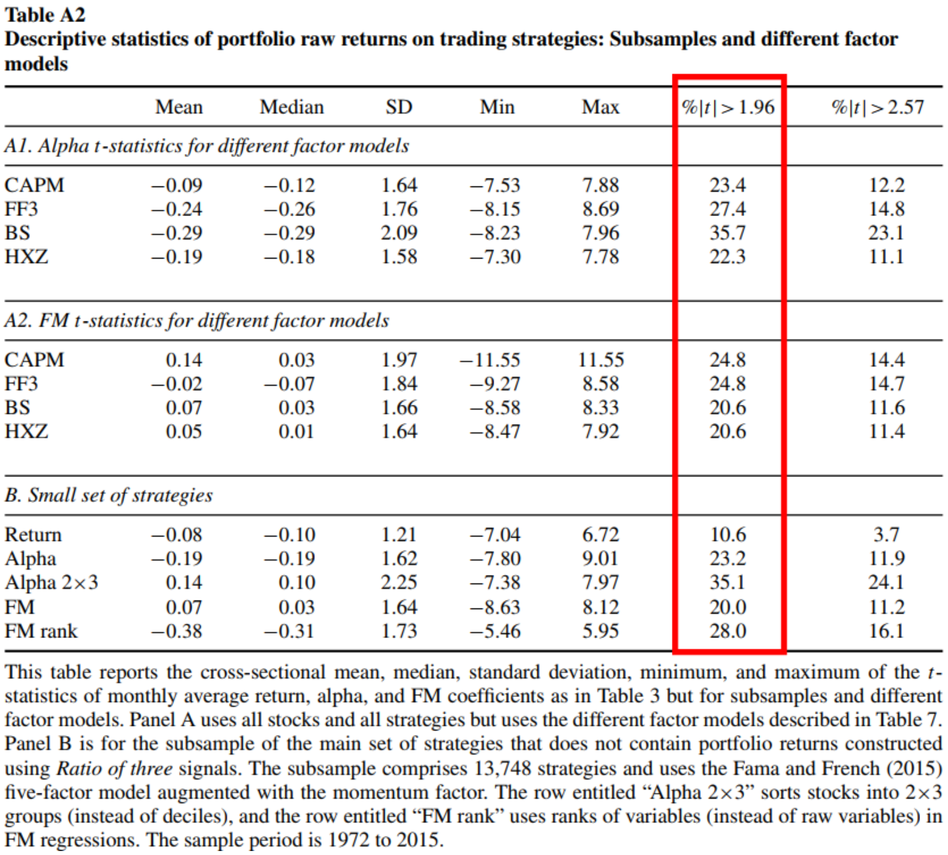


Table A2
Descriptive statistics of portfolio raw returns on trading strategies: Subsamples and different factor models

	Mean	Median	SD	Min	Max	$\%\lvert t\rvert > 1.96$	$\%\lvert t\rvert > 2.57$
A1. Alpha t-statistics for different factor models							
CAPM	−0.09	−0.12	1.64	−7.53	7.88	23.4	12.2
FF3	−0.24	−0.26	1.76	−8.15	8.69	27.4	14.8
BS	−0.29	−0.29	2.09	−8.23	7.96	35.7	23.1
HXZ	−0.19	−0.18	1.58	−7.30	7.78	22.3	11.1
A2. FM t-statistics for different factor models							
CAPM	0.14	0.03	1.97	−11.55	11.55	24.8	14.4
FF3	−0.02	−0.07	1.84	−9.27	8.58	24.8	14.7
BS	0.07	0.03	1.66	−8.58	8.33	20.6	11.6
HXZ	0.05	0.01	1.64	−8.47	7.92	20.6	11.4
B. Small set of strategies							
Return	−0.08	−0.10	1.21	−7.04	6.72	10.6	3.7
Alpha	−0.19	−0.19	1.62	−7.80	9.01	23.2	11.9
Alpha 2×3	0.14	0.10	2.25	−7.38	7.97	35.1	24.1
FM	0.07	0.03	1.64	−8.63	8.12	20.0	11.2
FM rank	−0.38	−0.31	1.73	−5.46	5.95	28.0	16.1

This table reports the cross-sectional mean, median, standard deviation, minimum, and maximum of the t-statistics of monthly average return, alpha, and FM coefficients as in Table 3 but for subsamples and different factor models. Panel A uses all stocks and all strategies but uses the different factor models described in Table 7. Panel B is for the subsample of the main set of strategies that does not contain portfolio returns constructed using *Ratio of three* signals. The subsample comprises 13,748 strategies and uses the Fama and French (2015) five-factor model augmented with the momentum factor. The row entitled "Alpha 2×3" sorts stocks into 2×3 groups (instead of deciles), and the row entitled "FM rank" uses ranks of variables (instead of raw variables) in FM regressions. The sample period is 1972 to 2015.

In most exhibits, the share of $|t_i| > 1.96$ ranges from 20 to 30 percent. Table A2 (shown in Figure A.2) is representative and is helpful since it shows 13 estimates in a single table (red box). Here, the shares range from 20% to 35%, with one exception. The one exception is the value-weighted long short mean return test, which leads to a share of 10.6%, which is consistent with the finding in Section 3.2 that value-weighted long-short raw return tests have unusually small t-stats in the CLZ data. All of CGS's alpha estimates use value-weighted returns.

If one can exclude this outlier, then 20% is a lower bound on $\Pr(|t_i| > 2)$, leading to $\text{FDR}_{|t|>2} \leq \frac{5\%}{0.20} = 25\%$.

A.2. FDR Bounds from plugging in predictor zoo numbers

This section describes the t-statistics collection for the predictor zoo papers listed in Table 2.

A.2.1. FDR Bounds using hand-collected Green, Hand, and Zhang (2013)

GHZ hand-collect numbers for 330 predictors from cross-sectional predictability studies. Table 3 reports the mean number of months used is 291. Table 4 reports the mean equal-weighted annual Sharpe ratio is 1.04. Thus, the mean published t-statistic is approximately $1.04 \times \sqrt{291/12} = 5.1$. Plugging into Equation (7) and (8) implies $\text{FDR}_{|t|>2} \leq 5\% \exp(2/(5.1-2)) = 10\%$.

Table 5 reports the mean value-weighted Sharpe ratio is 0.70, implying a mean published t-statistic is approximately $0.70 \times \sqrt{291/12} = 3.44$. Plugging into Equation (7) and (8) implies $\text{FDR}_{|t|>2} \leq 5\% \exp(2/(3.44-2)) = 20\%$.

A.2.2. FDR bounds using hand-collected t-stats in Chen and Zimmermann (2020)

The focus of CZ 2020 is on replications of published cross-sectional predictors. But as a robustness check on regarding their replicated t-stats, Chen and Zimmermann (2020) report results using 77 hand-collected t-stats from cross-sectional predictability studies. Given that CZ's replications use equal-weighting, it is sensible to assume that most all of these hand-collected t-stats also use equal-weighting.

Table 2 reports that the mean hand-collected t-statistic is 4.57. Plugging into Equation (7) and (8) implies $\text{FDR}_{|t|>2} \leq 5\% \exp(2/(4.57-2)) = 11\%$.

Of course, one can also use the replicated t-stats in CZ 2020. Since CZ 2020 find the replicated and hand-collected t-stats are very similar, the $\text{FDR}_{|t|>2}$ bound is also similar.

A.2.3. FDR bounds using hand-collected t-stats in Harvey, Liu, and Zhu (2016)

HLZ describe their data collection as follows: "Our goal is not to catalog every asset pricing paper ever published. We narrow the focus to papers that propose and test new factors." They arrive at 313 articles and "catalogue 316 different factors."

HLZ do not report the mean published t-statistic. However, one can back out the mean published t-statistic using Appendix A.1, which uses exponential distributions to extrapolate unpublished t-stats. Table A.1 illustrates the extrapolations (see Figure A.3).

Figure A.3: Harvey, Liu, and Zhu's (2016) Table A.1

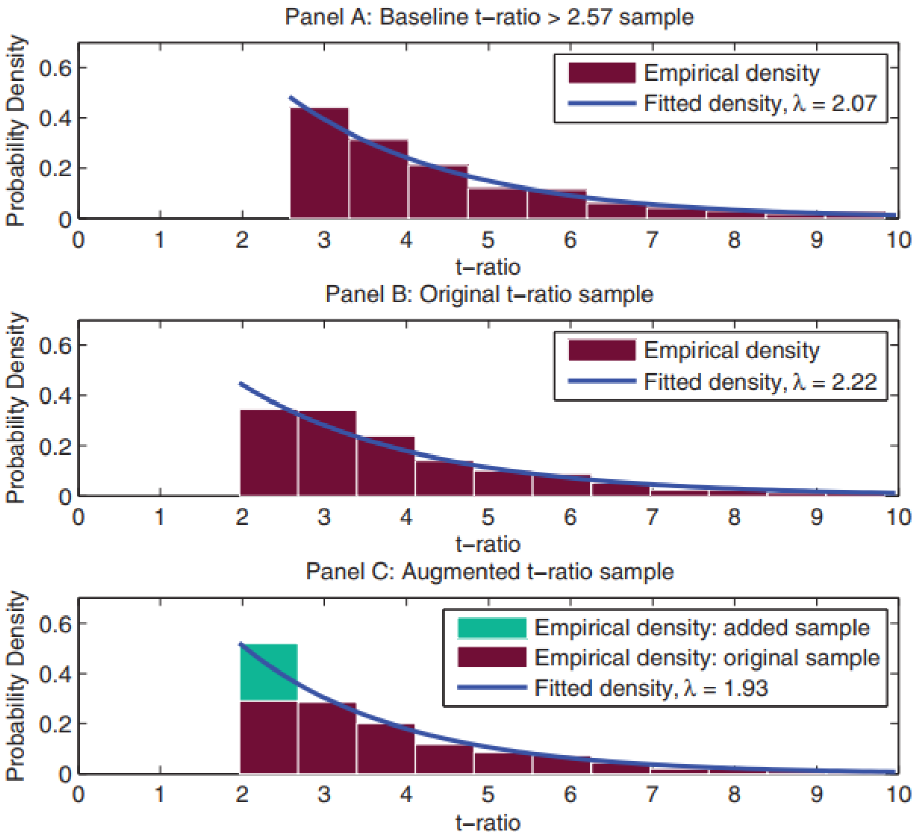


Figure A.1
Density plots for *t*-statistic
Empirical density and fitted exponential density curves based on three different samples. Panel A is based on the baseline sample that includes all *t*-statistics above 2.57. Panel B is based on the original sample with all *t*-statistics above 1.96. Panel C is based on the augmented sample that adds the subsample of observations that fall in between 1.96 and 2.57 to the original *t*-statistic sample. It doubles the number of observations within the range of 1.96 and 2.57 in the original sample. λ is the single parameter for the exponential curve. It gives the population mean for the unrestricted (i.e., nontruncated) distribution.

Using a variety of specifications, HLZ find that the mean t-stat overall is around 2.0. Applying Equation (8), then, implies a mean published t-stat of about 4.0.

In Table 2, I focus on the estimate in Panel B, because it is more directly comparable to the other estimates. Indeed, HLZ's footnote 50 implies that one can back out the mean published t-stat by applying Equation (8) to the estimate in Panel B. A caveat here, however, is that there seems to be a typo in footnote 50 ($\hat{\lambda} = 1/\left(\bar{t} - c\right)$ should say $\hat{\lambda} = \bar{t} - c$ to match the numbers in HLZ's Table A.1).

Regardless, all of the estimates in HLZ's Table A.1 imply that $\text{FDR}_{|t|>2}$ is quite small. Using the estimate of the mean overall t-stat implies $\text{FDR}_{|t|>2} = 5\% \exp(2/1.96) = 13.9\%$.

A.2.4. FDR bounds using replicated t-stats in McLean and Pontiff (2016)

MP replicate 97 long-short quintile strategies based on from published studies. Page 16 (paragraph 1) reports that the average t-statistic in their replications is 3.55. Plugging into Equation (7) and (8) implies $\text{FDR}_{|t|>2} \leq 5\% \exp(2/(3.55-2)) = 18\%$.

A.2.5. FDR bounds using replicated t-stats in Jacobs and Müller (2020)

This paper replicates 241 long-short quintile strategies based on published studies. The online appendix Table 2 reports a mean return of 55 bps per month and a t-statistic of 3.08 for US stocks over the full sample. Plugging into Equation (7) and (8) implies $\text{FDR}_{|t|>2} \leq 5\% \exp(2/(3.08-2)) = 32\%$.

The FDR over the full sample is likely higher than the FDR in the original samples, however. Table 2 reports pooled long-short returns in US and in-sample data have a mean return of 74.2 bps per month, much higher than the 55 bps per month over the full sample. Scaling the full sample t-stat by the mean return differential leads to rough estimate of the mean in-sample t-stat: $3.08 \times 74.2/55 = 4.2$, and $\text{FDR}_{|t|>2} \leq 5\% \exp(2/(4.2-2)) = 8\%$.

B. Simulation Appendix

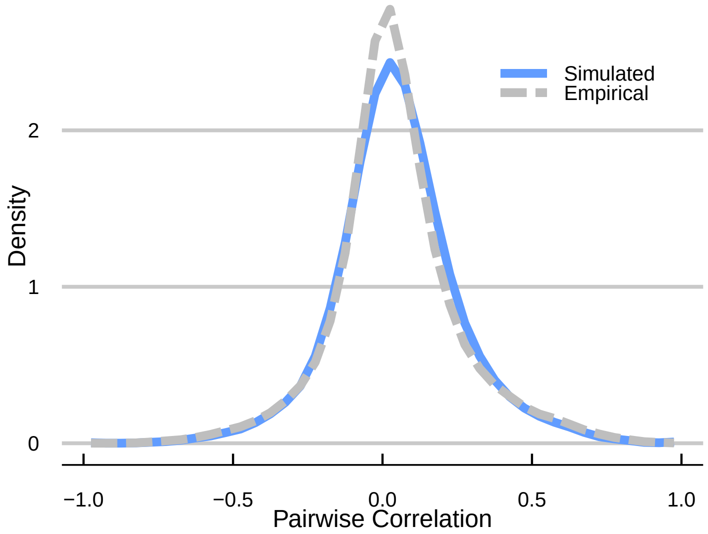
B.1. Details on simulations of data-mined strategies

To deal with missing values, I keep only strategies with at least 200 months of return observations and keep only months with at least 100 observations of return signals. I then construct residuals by subtracting the sample mean at the signal level from returns.

From this empirical panel, the cluster bootstrap selects 500 months with replacement ($\kappa(\tau)$, for $\tau = 1, 2, .., 500$), and then constructs a bootstrapped panel of residuals $\varepsilon_{d,\kappa(\tau)}$. The cluster bootstrap ensures that the simulated correlation

Figure A.4: Data-Mining Simulation vs Empirical Correlations

I simulate monthly long-short return residuals by cluster-bootstrapping from CLZ's data-mined strategies and compare with the correlations in the original CLZ data. Return residuals subtract the sample mean at the signal level from returns. Both distributions use a random sample of 1,000 signals for tractability.



structure closely matches the empirical structure, as seen in Figure A.4.

B.2. Estimations of simulations of published data

The simulation model is a slight modification of the data mining simulations (Section 4.2).

The key modification is that N returns are bootstrapped from the Chen and Zimmermann (2022) (CZ) replications as follows:

$$r_{i,\tau} = \mu_i + \varepsilon_{i,\tau}$$

where $\varepsilon_{i,\tau}$ is defined as

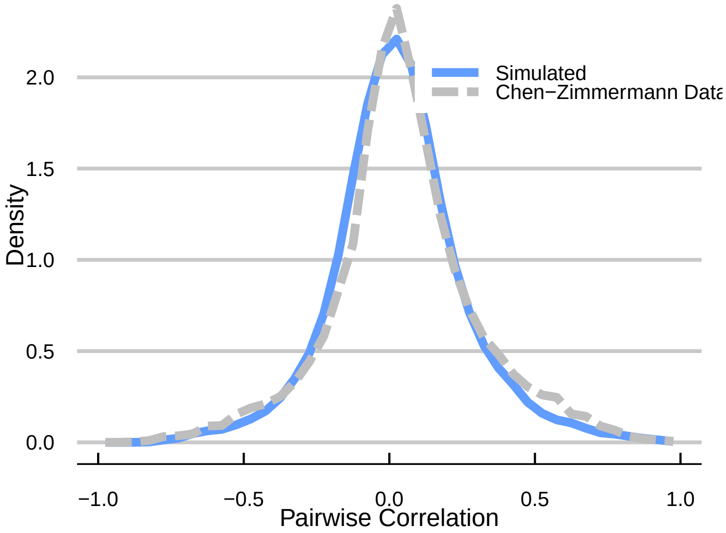
$$\varepsilon_{i,\tau} = 0.65\hat{\varepsilon}_{k(i),v(\tau)} + 0.35\delta_{i,\tau} \tag{23}$$

where $\hat{\varepsilon}_{k,v}$ is the de-meaned long-short return for signal k in month v in the CZ data, $k(i)$ and $v(\tau)$ are random integers, and $\delta_{i,t} \sim \text{Normal}(0, 3.32)$ i.i.d. In other words, I cluster-bootstrap residuals from empirical data, where the clustering preserves cross-sectional correlations, but I mix in 35% random noise, where the noise has a volatility similar to the empirical data. 3.32% is the mean volatility of returns in the CZ dataset.

Purely bootstrapping from the CZ data would lead to many redundant strategies and excessively high correlations for $N \gg 200$, since there are only 200 strategies in the CZ data. 35% random noise is selected to fit the empirical distribution of correlations, as seen in Figure A.5.

Figure A.5: Publication Simulation vs Empirical Correlations

I simulate monthly long-short return residuals by mixing a cluster-bootstrap of de-meaned returns from the Chen and Zimmermann (2022) data with noise ($\varepsilon_{i,t}$ in Equation (23)). Empirical return residuals subtract the sample mean at the predictor level (using in-sample data only) from returns. I compare simulated correlations with correlations in the original Chen and Zimmermann (2022) data. The simulated distribution uses a random sample of 1,000 predictors for tractability.



As in Section 4.2, expected returns follow

$$\mu_i = \begin{cases} 0 & \text{if } F_i \\ \gamma & \text{if } T_i \end{cases}.$$

I fix $N = 10,000$ and $T = 200$.

The publication process is the same as in HLZ:

$$\Pr(S_i|t_i) = \begin{cases} 0 & |t_i| < 1.96 \\ 0.5\bar{s} & |t_i| \in (1.96, 2.57] \\ \bar{s} & |t_i| > 2.57 \end{cases}, \tag{24}$$

where S_i indicates i is selected for publication and $\bar{s}$ is the maximum probability of selection. WLG, I set $\bar{s} = 1.0$. A lower $\bar{s}$ is equivalent to a smaller N.

In each simulation, I estimate the FDR bound using the easy exponential extrapolation (Equation (7) and (8)). I also calculate the FDP for published strategies with $|t_i| > 2.0$. I then average across simulations to obtain the actual FDR and the average FDR bound. The results are shown in Figure A.6.

For most parameter values, the easy extrapolation formula bounds the actual FDR.The exceptions occur when $\Pr(F_i)$ is around 90% or $\gamma = 75$ bps per month.

These settings imply that the distribution of t-stats displays an extremely sharp curvature below $|t_i| < 2.0$. Since the data mined returns do not display such a sharp curvature, it is unlikely that this setting is relevant, as it would mean that researchers uncover $|t_i| < 2.0$ at a much higher rate than pure data mining.

C. FDR Control Equivalence Details

This section provides further details on the expressions in Section 4.3.

To derive Equation (21), start with HLZ's description of BH95 control:

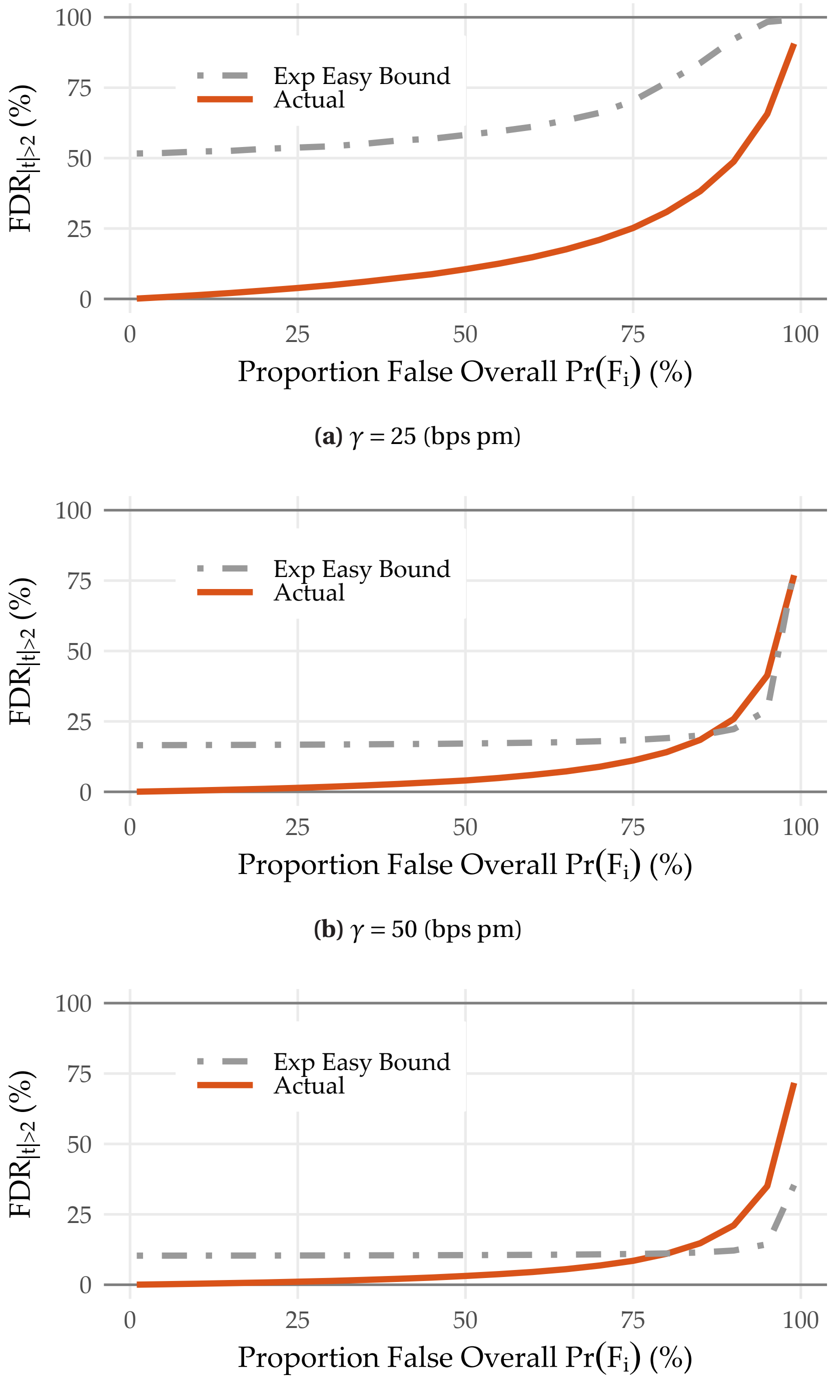
- Order the p-values such that $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(M)}$ where M is the total number of predictors.
- Find k^* that solves

  $$k^* = \max_{k \in \{1,2,..,M\}} \left\{ k : p_{(k)} \leq \frac{k}{M} q^* \right\} \tag{25}$$

  where q^* is the FDR bound selected by the researcher.
- Reject null hypotheses corresponding to $p_{(1)}, p_{(2)}, ..., p_{(k^*)}$.

Figure A.6: FDR Estimates in Simulations of Published Data

I simulate published returns, estimate the easy FDR bound using exponential extrapolation, and compare to the actual FDR (see Section B.2)



(a) $\gamma = 25$ (bps pm)

(b) $\gamma = 50$ (bps pm)

(c) $\gamma = 75$ (bps pm)

To see the equivalence, first note that Equation (25) can be written as

$$h^* = \min_{h \in \{|t_1|, |t_2|, \ldots, |t_M|\}} \left\{ h : \Pr\left(|t_i| > h | F_i\right) \leq \frac{[\text{Number of } |t_i| > h]}{M} q^* \right\}. \qquad (26)$$

That is, finding the largest p-value is equivalent to finding the smallest t-stat hurdle that satisfies the constraint. The constraint can be written in terms of a t-stat hurdle, $p_{(k)} = \Pr\left(|t_i| > h|F_i\right)$ for $h = |t_{(k)}|$, where $t_{(k)}$ ranks the absolute t-statistics in descending order. Finally, since k is the ranking of the p-value $p_{(k)}$, it is also the number of p-values smaller than $p_{(k)}$, or equivalently the number of $|t_i| > h$.

BY Theorem 1.3 simply modifies Equation (25) with a correlation penalty $\sum_{j=1}^{M} 1/j$. On page 1183, BY describe this penalty as follows: "Obviously, as the main thrust of this paper shows, the adjustment by $\sum_{i=1}^{m} \frac{1}{i} \approx \log(m) + \frac{1}{2}$ is very often unneeded, and yields too conservative a procedure." The BY paper is more often cited for Theorem 1.2, which omits this penalty.

FDR controls like BH95 are significantly more difficult to prove compared to direct FDR estimates because controls imply that the hurdle is stochastic. In contrast, direct FDR estimates fix the hurdle, leading to simple manipulations like Equations (16).

Nevertheless, FDR control under dependence has been proven in a variety of settings. Many such settings are listed in Farcomeni (2007). Simutheoretical evidence suggests that FDR control generally obtains for t-tests like the one in this paper (Reiner-Benaim (2007)) but a complete proof of this result has remained elusive (Benjamini (2010)). The robustness of BH95 seems to be even broader than in Reiner-Benaim, 2007. In his lecture notes on "A Tutorial on False Discovery Control," Christopher Genovese describes the BH95 algorithm as "In practice, it is quite hard to 'break'."

D. Harvey, Liu, and Zhu's (2016) Parametric Model

HLZ's model starts with a description of all factors, $i = 1, 2, ..., N$

$$\mu_i \sim \begin{cases} \text{Dirac}\,(0) & \text{with prob } p_0 \\ \text{Exponential}\,(\lambda) & \text{with prob } (1 - p_0) \end{cases} \tag{27}$$

$$t_i | \mu_i \sim \text{Normal}\left(\frac{\mu_i}{SE}, 1\right) \tag{28}$$

$$\text{Corr}\left(t_i, t_j\right) = \rho \text{ for } i \neq j$$

where μ_i is the expected return on factor i, Dirac (0) is the distribution with a point mass at zero, Exponential (λ) is an exponential distribution with shape parameter λ, p_0 if the proportion of nulls, SE is the (constant) standard error of the sample mean return, and ρ is the constant pairwise correlation between monthly returns.

HLZ choose $SE = (1500/\sqrt{12})/\sqrt{240}$ (page 29) and $\rho = 0.20$ (page 35).

Factors are selected for publication according to a staircase function

$$\Pr\left(S_i | t_i\right) = \begin{cases} 0 & |t_i| < 1.96 \\ 0.5\bar{s} & |t_i| \in (1.96, 2.57] \ , \\ \bar{s} & |t_i| > 2.57 \end{cases} \tag{29}$$

where S_i indicates i is selected and $\bar{s}$ is the maximum probability of selection. HLZ do explicitly describe this function but it is implied by their data adjustment for publication bias (page 31). HLZ implicitly assume $\bar{s} = 1.0$, which allows identification of N. It is unclear if $\bar{s} = 1.0$ is a reasonable assumption and indeed Andrews and Kasy (2019) and Chen (Forthcoming) do not attempt to identify N or $\bar{s}$.

This model contains parameters (p_0, λ, N) which HLZ estimate by SMM. They select as target moments the number of published and significant factors, as well as three order statistics t-statistics: the 20th, 50th, and 90th percentiles. HLZ's Table 5, Panel A, $\rho = 0.2$ arrives at $p_0 = 0.444$, $\lambda = 55.5$ bps per month, and $N = 1{,}378$.

E. Equating of false, null, and insignificant in Harvey and Liu (2020)

Harvey and Liu (2020) (HL) “show exactly what went wrong with the inference in Yan and Zheng (2017)” (page 2506, paragraph 4). In particular, HL state that Yan and Zheng “claim that a large fraction of the 18,000 anomalies in their data are true.” (page 2506, paragraph 2). This statement about Yan and Zheng is repeated several times in HL (e.g. page 2519, paragraph 3) and HL devote an entire section (II.A.5) to refuting this claim.

However, Yan and Zheng do not make this claim. Instead, they “find that many fundamental signals are significant predictors of cross-sectional stock returns even after accounting for data mining” (abstract). Similar statements are found throughput Yan and Zheng (e.g. page 1416, paragraph 3). HL’s statement and Section II.A.5 are thus hard to understand—unless one equates significance (under HL’s chosen test) with truth.

HL explain their argument in more detail in section II.A.5, “Revisiting Yan and Zheng (2017).” This section begins (page 2541) with:

> *Applying the preferred methods based on Table III to the 18,000 anomalies, the fraction of true strategies is found to be 0.000% ([BY Theorem 1.3]) and 0.015% (Storey, $\theta = 0.8$) under a 5% significance level, and 0.006% ([BY Theorem 1.3]) and 0.091% (Storey, $\theta = 0.8$) under a 10% significance level.*

There are multiple issues with this paragraph. First, the fraction of true strategies does not traditionally depend on the significance level.

Second, BY Theorem 1.3 traditionally provides an *upper bound* on the FDR (Section 4.3). It can say the fraction of true strategies is *at least 0.000%.* But this theorem cannot say that the fraction of true strategies is 0.000%.

Last, it is not possible that Storey (2002) finds an FDR of less than 1% in the Yan and Zheng (2017) strategies. This can be seen applying Equation (11) to the numbers reported in Table 1 of Yan and Zheng. This table shows that at most 80 percent of $|t_i|$ fall between -1.62 and +1.58 (using VW 4-factor αs, which is the only part of the Yan and Zheng data examined in HL). Thus, the law of total probability implies that

$$\begin{aligned}\Pr(F_i) &= \frac{\Pr(t_i \in [-1.62, +1.58]) - \Pr(t_i \in [-1.62, +1.58] | T_i) \Pr(T_i)}{\Pr(t_i \in [-1.62, +1.58] | F_i)} \\ &\leq \frac{\Pr(t_i \in [-1.62, +1.58])}{\Pr(t_i \in [-1.62, +1.58] | F_i)} \\ &= \frac{0.80}{0.89} = 89\%,\end{aligned}$$

where $\Pr(t_i \in [-1.62, +1.58] | F_i) = 0.89$ is calculated using the standard normal distribution. In other words, the fraction of true strategies is at least 11%.