

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

Andrew Y. Chen
Federal Reserve Board

October 2025*

Abstract

The false discovery rate (FDR) measures the share of false positives in a set of statistical tests. I develop simple and intuitive bounds on the FDR in cross-sectional predictability publications. The simplest bound requires just a few lines of math and finds $\text{FDR} \leq 25\%$ based on summary statistics in eight out of nine previous studies. A more refined bound finds $\text{FDR} \leq 9\%$. The FDR is small because randomly selecting accounting ratios produces statistically significant predictability far more often than would occur if there were no predictability. The bounds also reconcile the disparate FDR estimates in the literature.

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 Alec Erb, 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), Valentin Haddad, Cam Harvey (discussant), Dino Palazzo, Jeffrey Pontiff, A. Subra, 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, UCLA, UC Riverside, the University of Maryland, 2024 Western Finance Association Meetings, 2024 Chicago Booth Asset Pricing conference, 2024 BIGFI conference, and Northeastern University. I am grateful to Sterling Yan and Lingling Zheng for sharing their data. The views expressed herein 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 (Chen and Zimmermann 2022). This abundance has led to concerns about multiple testing. Intuitively, if researchers run many tests, some may be statistically significant by pure chance, even under the null hypothesis of no predictability.

The false discovery rate (FDR) quantifies this problem. The FDR originates from Sorić (1989), who defines a “discovery” as a statistically significant result, and a “false discovery” as a discovery for which the null (unfortunately) holds. Building on this language, Benjamini and Hochberg (1995) define the false discovery *rate* as the expected share of discoveries that are false. Discoveries that are not false are true.

This paper provides simple and intuitive bounds on the FDR in the cross-sectional predictability literature. My “Easy Bound” is derived in just a few lines of math and can be estimated by plugging in summary statistics from previous studies. This bound shows that the FDR is at most 25% based on eight out of the nine studies I examine. My “Visual Bound” is tighter, and can be calculated by drawing a picture. The Visual Bound implies that the FDR is at most 9%—in other words, at least 91% of discoveries are true.

The small FDR comes from a surprising fact: randomly selecting accounting ratios produces statistically significant predictability rather easily. In data mining studies, at least one out of five accounting ratios produces a t-stat exceeding 2.0 (Yan and Zheng 2017; Chordia, Goyal, and Saretto 2020; Chen, Lopez-Lira, and Zimmermann 2025). If there were no predictability, this rate would be one out of twenty—25% of what is found empirically. My Easy Bound shows that this 25% is precisely an upper bound on the FDR.

These bounds help reconcile the disparate estimates in the literature. Harvey, Liu, and Zhu (2016) and Chordia, Goyal, and Saretto (2020) find an FDR exceeding 45%, a result echoed by Harvey and Liu (2020). In contrast, Chen and Zimmermann (2018) and Jensen, Kelly, and Pedersen (2023) find an FDR close to zero, and Chen (2024) finds a similarly small FDR. My bounds show that these different conclusions are not due to differences in the underlying data. Data in all of these studies imply an FDR of at most 25%.

Instead, the debate is largely due to differences in interpretation. I explain how Harvey, Liu, and Zhu’s (2016) own methods imply an FDR of just 9%. But the paper interprets “insignificant factors” as “false discoveries,” leading to the claim that most findings are false. Similar interpretations occur in Harvey and Liu (2020). Chordia, Goyal, and Saretto (2020) uses the standard interpretation, but their high FDR estimates likely stem from the omission of key moments from their calibration.

Replication code is available at <https://github.com/chenandrewy/mostly-true>.

2. An Easy Bound on the FDR

The cross-sectional predictive power of signal i (a stock-level characteristic) is measured with $\bar{r}_i$. $\bar{r}_i$ is commonly constructed by forming a long-short portfolio based on i and computing a sample mean return, possibly with controls for factor exposure. The null hypothesis for i is $E(\bar{r}_i) = 0$. If this null hypothesis holds, I refer to i as “null” and denote this event by null_i . $\bar{r}_i$ has standard error SE_i and t-stat $t_i \equiv \bar{r}_i / \text{SE}_i$. The samples are long, so the central limit theorem holds:

$$t_i | \text{null}_i \sim \text{Normal}(0, 1). \quad (1)$$

Equation (1) can also be justified using a bootstrap (Chen 2021). These assumptions are standard in cross-sectional asset pricing (Cochrane 2009; Bali, Engle, and Murray 2016).

In a sense, FDR methods just add colorful terminology. Following Sorić (1989), a “discovery” is a significant result. For the bulk of the paper, I use the classical significance threshold of 5%, which I round to the following t-stat threshold:

$$i \text{ is a discovery if } |t_i| > 2. \quad (2)$$

A “false discovery” is a significant signal that is (unfortunately) null. Significant signals are also called “predictors” or “findings,” so I also refer to false discoveries as “false predictors” or “false findings.”

With this terminology, a natural definition of the false discovery *rate* is

$$\text{FDR}_{|t|>2} \equiv \Pr(\text{null}_i | |t_i| > 2), \quad (3)$$

where i is randomly selected from some set. In other words, $\text{FDR}_{|t|>2}$ is the probability a significant signal is null. Equation (3) is equivalent to the Benjamini and Hochberg (1995) definition under weak dependence (e.g. Chen 2024).

Applying Bayes rule and noting $\Pr(\text{null}_i) \leq 1.0$ leads to an “Easy Bound” on the FDR:

$$\text{FDR}_{|t|>2} = \frac{\Pr(|t_i| > 2 | \text{null}_i) \Pr(\text{null}_i)}{\Pr(|t_i| > 2)} \quad (4)$$

$$\leq \frac{\Pr(|t_i| > 2 | \text{null}_i)}{\Pr(|t_i| > 2)} \quad (5)$$

$$= \frac{5\%}{\Pr(|t_i| > 2)}, \quad (6)$$

where the last line applies Equation (1). The Easy Bound is just 5% divided by the probability a signal is significant.

It’s natural to estimate this probability using method of moments: replace $\Pr(|t_i| > 2)$

with the share of signals with $|t_i| > 2$. The resulting bound is intuitive: if the tail under the null (numerator) cannot explain the actual tail (denominator), then the FDR must be small.

However, estimating $\Pr(|t_i| > 2)$ runs into the problem of publication bias. Since signals with $|t_i| > 2$ are more likely to be published, the published share of $|t_i| > 2$ overstates $\Pr(|t_i| > 2)$. I address this problem by considering worst-case scenarios.

2.1. Bounding the FDR with Atheoretical Data Mining

I do not observe the sets of signals researchers consider. I do, however, observe sets of accounting ratios from data mining studies, namely Yan and Zheng (2017) (YZ); Chordia, Goyal, and Saretto (2020) (CGS); and Chen, Lopez-Lira, and Zimmermann (2025) (CLZ). If these accounting ratios are randomly selected, then Equation (6) produces a valid bound on $\text{FDR}_{|t|>2}$ from mining accounting ratios. While this is not the ideal bound, it is likely conservative. Given that researchers should use a more refined set of signals, the $\text{FDR}_{|t|>2}$ from selecting random accounting ratios should be a worst-case scenario for the cross-sectional literature.

Table 1, Panel (a), describes sets of accounting ratios from the YZ, CGS, and CLZ data mining studies. As seen in the “Accounting Ratios Allowed” column, these sets are random: they employ little to no economic theory and instead are designed to systematically explore the data. Nevertheless, at least 20% of these ratios produce $|t_i| > 2$. Plugging 20% into Equation (6) implies

$$\text{FDR}_{|t|>2} \leq \frac{5\%}{0.20} = 25\% \quad (7)$$

for mining accounting ratios. Then, assuming that research does not produce a higher rate of false discoveries, $\text{FDR}_{|t|>2} \leq 25\%$ for the cross-sectional literature. In other words, at least 75% of claimed findings are true.

This robustness is notable, given that each paper is written by an independent research team. Indeed, each team chose a different set of accounting ratios, as seen in Table 1. Nevertheless, they all find that $|t_i| > 2$ for at least 20% of signals in their main specifications.

While the set of accounting ratios seems to have little effect, a close read of these papers shows that predictability measurement matters. Most importantly, measures that target large stocks (e.g. value-weighted portfolios) produce much smaller t-stats. Section 3.1 takes a closer look at this effect, using data on YZ’s accounting ratios. Nevertheless, Equation (7) provides a conservative estimate of what we’ll see there, and can be calculated from easily-available summary statistics.

Table 1: Easy Bounds on the FDR Based on Nine Studies

$FDR_{|t|>2}$ is defined in Equation (3). Papers in Panel (a) begin with ≈ 200 accounting variables and then apply the functional forms in “Accounting Ratios Allowed.” Papers in Panel (b) re-examine published predictors. ‘Pr($|t_i| > 2$) Bound’ and ‘Mean Pub t-stat’ are inferred from the ‘Reported Statistic’ or from the ‘Exhibit.’ ‘FDR $_{|t|>2}$ Bound’ uses Equation (7) (Panel (a)) or Equation (9) (Panel (b)). ‘BBH’ is Bessembinder, Burt, and Hrdlicka (2025). Using data from nine studies, by eight independent research teams, most claimed findings are true. For eight out of nine studies, $FDR_{|t|>2} \leq 25\%$.

Panel (a): Data-Mining as a Worst Case (Equation (7))				
Data Mining Study	Accounting Ratios Allowed	Reported Statistic	Pr($ t_i > 2$) Bound (%)	FDR $_{ t >2}$ Bound (%)
(1) Yan and Zheng (2017) Table 1 (EW)	$X/Y, \Delta(X/Y), \% \Delta(X/Y), \% \Delta X, \Delta X/\text{lag} Y, \% \Delta X - \% \Delta Y,$ Y restricted to 15 vars	90th and 10th ptile of t-stats are 2.41 and -3.48, respectively	20	25
(2) Chordia, Goyal, Saretto (2020) Table 3 (Alpha)	$\Delta X/\text{lag}(X), X/Y$ $(X - Y)/Z$	23% of $ t_i $ exceed 1.96	23	22
(3) Chen et al. (2025) Table 1 (Equal-Weighted)	$X/Y, \Delta X/\text{lag} Y$	Mean t-stat for 2nd quintile is -2.46	20	25
Panel (b): Exponential Extrapolation as a Worst Case (Equation (9))				
Meta Study	Exhibit	t-stat Construction	Mean Pub t-stat	FDR $_{ t >2}$ Bound (%)
(4) Green, Hand, Zhang (2013)	Tables 3 and 4	Hand Collected	5.1	10
(5) Chen, Zimmermann (2020)	Table 2	Hand Collected	4.6	11
(6) Harvey, Liu, Zhu (2016)	Figure A.1 Panel B	Hand Collected	4.2	12
(7) McLean, Pontiff (2016)	Page 16, Par 1	Replicated In-Sample	3.6	18
(8) Jensen, Kelly, Pedersen (2021)	Table 1 of BBH 2025	Replicated Full-Sample	3.6	18
(9) Jacobs, Muller (2020)	Online App Table 2	Replicated Full-Sample	3.1	32

Studies that focus on randomly-selected past return signals are less common. But the existing evidence also implies a small $\text{FDR}_{|t|>2}$. YZ and Chen and Dim (2024) examine past return signals as part of their studies, and both find that $|t_i| > 2$ is quite common. Applying Equation (7) to YZ’s Table 9 shows that the same $\text{FDR}_{|t|>2} \leq 25\%$ applies to their past return signals. One can also find the same bound using Chen and Dim’s publicly-available data.

2.2. Bounding the FDR with Conservative Extrapolation

A second worst-case scenario comes from conservatively extrapolating the t-stat density into the region affected by publication bias. While extrapolating point estimates is dangerous, we only need to extrapolate an upper bound. That is, we don’t need the true density for $|t_i| < 2$ —we only need a density that exceeds the true density. As long as this condition is satisfied, the $\text{FDR}_{|t|>2}$ implied by the extrapolation bounds the $\text{FDR}_{|t|>2}$ from the research process.

I extrapolate using an exponential distribution for three reasons:

1. It fits the observed data well. This is seen in Figure 1, which compares an exponential density with the distribution of t-stats from Chen and Zimmermann’s (2022) replications of 200 published predictors.
2. It is conservative. It assumes, in a sense, “exponentially more” file drawer t-stats compared to published t-stats. Conservatism is also seen in how the shape is similar to the data-mined t-stat distribution (dashed line, Figure 1).
3. It can be estimated by hand, using summary statistics reported in previous meta-studies.

An exponential extrapolation is also used in Appendix A.1 of Harvey, Liu, and Zhu (2016).

To understand the third point, note that if $|t_i|$ is exponentially distributed, the memoryless property implies

$$E(|t_i|) = E(|t_i| \mid |t_i| > 2) - 2, \quad (8)$$

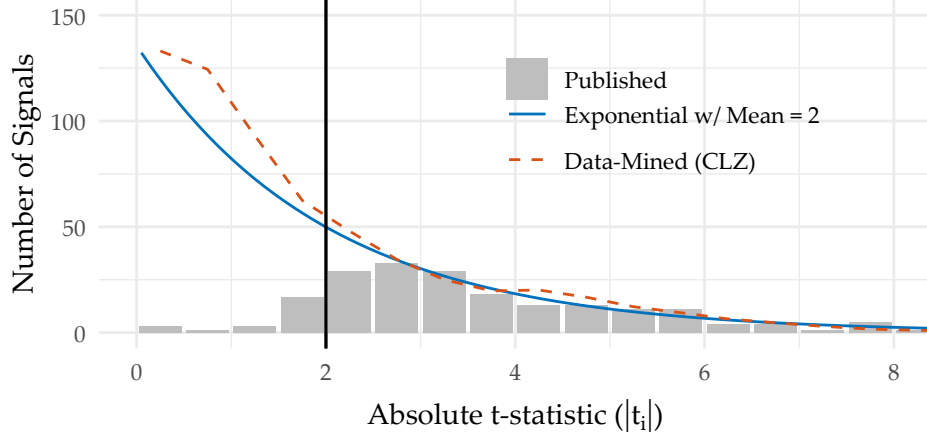
that is, $E(|t_i|)$ can be estimated by subtracting 2 from the mean $|t_i|$ conditional on $|t_i|$ exceeding 2. Since publication generally requires $|t_i| > 2$, the mean published t-stat is not far from this conditional mean.

This reasoning leads to the following bound:

$$\text{FDR}_{|t|>2} \leq \frac{5\%}{\exp(-2/(\hat{\mu}_{\text{pub}} - 2))}, \quad (9)$$

Figure 1: Conservative Extrapolation from Published t-stats

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. Dashed line overlays the distribution of (equal-weighted) absolute t-stats from CLZ’s data-mined signals. The extrapolation is conservative, has a similar shape as the data-mined distribution, and fits the data well.



where $\hat{\mu}_{\text{pub}}$ is the mean published t-stat and the denominator comes from plugging an exponential distribution with mean $\hat{\mu}_{\text{pub}} - 2$ into Equation (6).

Panel (b) of Table 1 applies Equation (9) to a variety of 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 (9) then implies $\text{FDR}_{|t|>2} \leq 5\% / \exp(-2/(4.6 - 2)) = 11\%$. Replications and post-publication data tend to result in lower t-stats, leading to higher FDR bounds using McLean and Pontiff’s (2016) and Jacobs and Müller’s (2020) replications. Nevertheless, the results across six independent teams are consistent. All imply that most findings are true.

The entry for Harvey, Liu, and Zhu (2016) (HLZ) is notable. It shows that the Easy Bound, applied to HLZ’s dataset, implies an FDR of at most 12%. This result contrasts with HLZ’s conclusion that most findings are false, and suggests that the disparate FDR estimates in the literature are not due to differences in data. We will see that different methods are not the issue either, and that the debate is driven by different interpretations of multiple testing methods (Section 4).

In summary, the FDR is small because the research process readily generates signals with large t-stats, much more readily than compared to a standard normal distribution. This property can be seen in the relatively high frequency of large t-stats in data mining experiments or in conservative extrapolations of published t-stats.

3. More Refined Estimates

The bounds in Section 2 are coarse. Rather than bound $\Pr(\text{null}_i)$ based on data, they use the theoretical bound $\Pr(\text{null}_i) \leq 1.0$. They gloss over important details of predictability measurement like equal- vs value-weighting. They use only summary statistics from published tables, rather than actual datasets.

This section addresses these issues. Section 3.1 bounds $\Pr(\text{null}_i)$ with data. Section 3.3 conditions the bound on the details of predictability measurement. These methods are applied to the Chen, Lopez-Lira, and Zimmermann (2025) and Yan and Zheng (2017) datasets. Along the way, Section 3.2 discusses correlations and standard errors.

3.1. A Tighter, Visual Bound

To obtain a tighter bound, we can use data to make inferences on $\Pr(\text{null}_i)$. A natural data point to use is the fraction of t-stats that are small. Intuitively, if $\Pr(\text{null}_i)$ is large, then many t-stats should be small.

To formalize this intuition, apply the law of total probability to the event that $|t_i|$ is small, say less than 0.5:

$$\Pr(|t_i| < 0.5) = \Pr(|t_i| < 0.5 | \text{null}_i) \Pr(\text{null}_i) + \Pr(|t_i| < 0.5 | \text{alt}_i) \Pr(\text{alt}_i). \quad (10)$$

where alt_i is the event that i is not null. Noting that $\Pr(|t_i| < 0.5 | \text{alt}_i) \Pr(\text{alt}_i) \geq 0$ and applying Equation (1) yields

$$\Pr(|t_i| < 0.5) \geq \Pr(|t_i| < 0.5 | \text{null}_i) \Pr(\text{null}_i) \quad (11)$$

$$= (0.38) \Pr(\text{null}_i) \quad (12)$$

where 0.38 is the probability that a standard normal r.v. falls between -0.5 and +0.5. Equation (12) provides a bound on $\Pr(\text{null}_i)$ based on the observed frequency of small t-stats.

Plugging Equation (12) into Equation (4) implies a “Visual Bound” on $\text{FDR}_{|t|>2}$:

$$\text{FDR}_{|t|>2} \leq \left[\frac{5\%}{\Pr(|t_i| > 2)} \right] \left[\frac{\Pr(|t_i| < 0.5)}{0.38} \right]. \quad (13)$$

This expression tightens the Easy Bound (first bracket) with a data-based bound on $\Pr(\text{null}_i)$ (second bracket). As in the Easy Bound, Equation (13) can be estimated with method of moments: replace probabilities with observed frequencies. This estimation is a variation of Storey’s (2002) Algorithm 1, though my argument does not require independence. I return to this point in Section 3.2.

To obtain numbers for Equation (13), I use data from Chen, Lopez-Lira, and Zim-

mermann (2025) (CLZ), which is publicly available.¹ It's important to use data-mined rather than published signals, since Equation (13) requires reliable data on signals with small t-stats. I use equal-weighted raw returns to match the modal specification in the literature (Green, Hand, and Zhang 2013) but Section 3.3 examines other performance measures. 9,700 of 29,000 signals have $|t_i| > 2$, and 6,300 have $|t_i| < 0.5$. Plugging these numbers into Equation (13) implies

$$\text{FDR}_{|t|>2} \leq \left[\frac{5\%}{9,700/29,000} \right] \left[\frac{6,300/29,000}{0.38} \right] = 8.5\% \quad (14)$$

for CLZ's data mining process. Then assuming that research does not produce a higher rate of false discoveries, $\text{FDR}_{|t|>2} \leq 8.5\%$ for the cross-sectional literature. In other words, at least 91.5% of equal-weighted raw return findings in the literature are true.

I call Equation (13) the Visual Bound because it can be estimated using the following algorithm:

Algorithm 1 Visual Bound Estimation

1. Plot the data histogram for $|t_i|$ for some data-mined signals.
 2. Draw the largest null histogram that still fits inside the data.
 3. Draw a vertical line at 2.0. To the right of this line, the ratio of the null area to the data area estimates the bound on $\text{FDR}_{|t|>2}$ in Equation (13).
-

This algorithm is equivalent to estimating Equation (13) with method of moments, under the assumption that the density of $|t_i|$ is decreasing and that the histogram bins include $[0, 0.5)$. To see this, note that estimating Equation (13) is equivalent to taking the ratio of the right tail areas of two histograms: (1) a null histogram scaled by an estimate of $\Pr(|t_i| < 0.5) / 0.38$, and (2) a data histogram. Step 2 imposes a scaling factor on the null histogram. This scaling factor is equal to $\Pr(|t_i| < 0.5) / 0.38$ under the two assumptions. Step 3 then compares the scaled histogram tails.

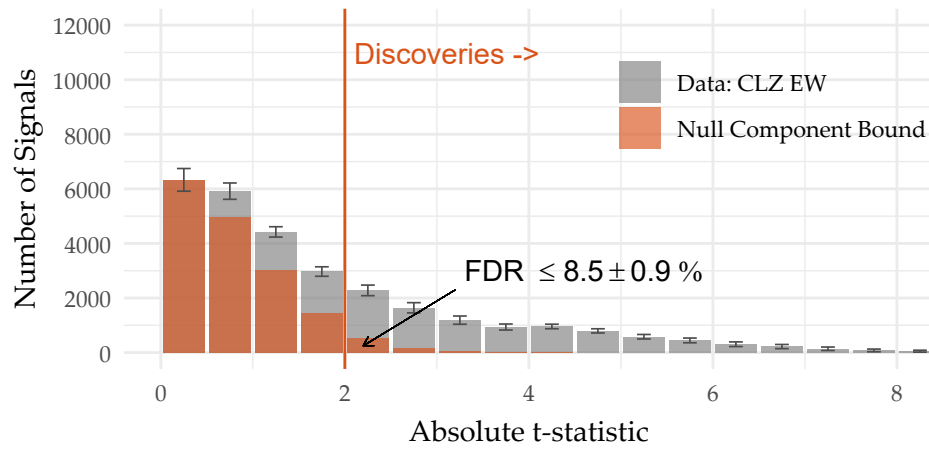
Figure 2, Panel (a), applies Algorithm 1 to CLZ's equal-weighted raw returns. The gray bars show the data histogram of $|t_i|$. The red bars show the largest null distribution that still fits inside the data. The result is an intuitive visualization of Equation (14). To the right of the vertical line at 2.0, the ratio of the null area (red) to the gray area is only 8.5%.

More broadly, Algorithm 1 shows how to interpret FDR methods as a histogram decomposition. This interpretation is seen in Figure 2, but it can be derived from first prin-

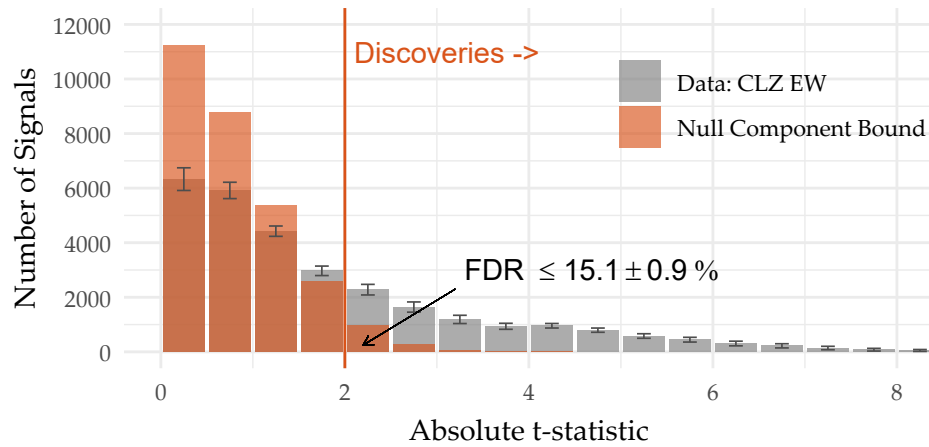
¹The CLZ data can be found at <https://sites.google.com/site/chenandrewy/>

Figure 2: A Visual Bound on the FDR

Bars show the distribution of CLZ equal-weighted raw return t-stats. The red component is Normal(0,1) scaled by a constant. Panel (a) uses Algorithm 1 to determine the constant. Panel (b) uses the Easy Bound. Error bars show one standard error calculated with a cluster bootstrap. Algorithm 1 provides an visual interpretation of Equation (13) and shows that at least 91% of equal-weighted raw return findings are true.



(a) Visual Bound



(b) Easy Bound

ciples.

To formalize this interpretation, let b be a histogram bin, and apply the law of total probability to the event that $|t_i|$ falls in bin b :

$$\underbrace{\Pr(|t_i| \in b)}_{[\text{Data Component}]_b} = \underbrace{\Pr(|t_i| \in b | \text{null}_i) \Pr(\text{null}_i)}_{\equiv [\text{Null Component}]_b} + \underbrace{\Pr(|t_i| \in b | \text{alt}_i) \Pr(\text{alt}_i)}_{\equiv [\text{Alternative Component}]_b}. \quad (15)$$

This equation says that each bar of a t-stat histogram (LHS) is the sum of a null component and an alternative component (RHS). We do not know the null component. But Algorithm 1 can bound it, using an intuitive constraint: the null component cannot be larger than the data component. If it were larger, then the law of total probability would be violated.

The null component, in turn, provides the key quantities in FDR methods. It provides $\text{FDR}_{|t|>2}$:

$$\text{FDR}_{|t|>2} = \frac{\sum_{b \subseteq [2, \infty)} [\text{Null Component}]_b}{\sum_{b \subseteq [2, \infty)} [\text{Data}]_b}, \quad (16)$$

where I've assumed that the bins have an edge at 2.0. To see why Equation (16) holds, plug in the definition of the null component, and compare to Equation (4). The null component also provides $\Pr(\text{null}_i)$:

$$\Pr(\text{null}_i) = \sum_{b \in B} [\text{Null Component}]_b \quad (17)$$

where I've assumed that B covers the support of $|t_i|$. Algorithm 1 provides upper bounds on these null components, and thus upper bounds on $\text{FDR}_{|t|>2}$ and $\Pr(\text{null}_i)$.

Equation (17) is visualized in Panel (a) of Figure 2: the ratio of the total red area to the total gray area is an upper bound on $\Pr(\text{null}_i)$. This ratio implies $\Pr(\text{null}_i) \leq 56\%$, which may be surprising. $\Pr(\text{null}_i) \leq 56\%$ means that at least 44% of randomly-selected accounting ratios have real, non-zero expected returns.

Though this result may be surprising, a larger $\Pr(\text{null}_i)$ is implausible, as illustrated in Panel (b). This panel shows an alternative decomposition, in which $\Pr(\text{null}_i) > 56\%$ is possible. The result is an unappealing decomposition, in which red bars exceed the gray bars. This feature is not only visually unappealing, it also violates principles of probability. Yet it is precisely the decomposition implicit in the Easy Bound (Equation (6)). Thus, while the FDR bounds in Section 2 are easy to calculate, they are likely too conservative. They violate the “budget constraint” imposed by the law of total probability.

3.2. Correlations, Standard Errors, and Factor Structure

The Visual and Easy Bounds do *not* require independence. As seen in Algorithm 1, FDR methods boil down to histogram estimation, which is valid under the weak dependence assumptions described in Wooldridge (1994).

Correlations do increase estimation uncertainty. Figure 2 captures this effect with the Fama and French (2010) cluster bootstrap, which preserves cross-sectional correlations. The error bars are non-negligible but the histogram is well-estimated overall. The standard error on the FDR bound is only 0.9 percentage points. In fact, this standard error is likely overstated, as the Fama and French (2010) method suffers from “bootstrap dilation” in this large scale setting (Efron 2003; 2010a; 2010b).²

A small standard error is consistent with the large number of signals and moderate correlations in the data. Consider Panel (b), which builds on the estimate of $\Pr(|t_i| > 2) = 0.33$. Under independence, the standard error would be only $\sqrt{0.33(1 - 0.33)/29,000} = 0.3$ percentage points, due to the 29,000 signals.³ Correlations imply a larger standard error, but the correlations in cross-sectional predictors are not particularly high. CLZ find that among data-mined predictor returns, more than 80% of pairwise correlations are less than 0.25 in absolute value, and that roughly 50 principal components are required to explain 75% of total variance (see also McLean and Pontiff 2016; Chen and Zimmermann 2022).

Correlations also impact the interpretation. Figure 2 shows roughly 10,000 true discoveries, but they cannot all be independent. If they were, one could combine them to form a portfolio with a Sharpe ratio $\sqrt{10,000} = 100$ times larger than the individual Sharpe ratios. In reality, combining data-mined predictors leads to annualized out-of-sample Sharpe ratios of about 1.5 (YZ, CLZ, Chen and Dim 2024), which, while high, are far short of the 100 $\times$ boost. The observed Sharpe ratios are broadly consistent with CLZ’s principal components analysis, which implies roughly 50 independent discoveries rather than 10,000.

3.3. Controlling for Value Weighting and Factor Exposure

There are multiple ways to measure the predictive power of a signal i . A more refined estimate of $\text{FDR}_{|t|>2}$ should control for details of this measurement. Table 2 examines how stock weighting and factor model adjustments affect Equation (13), applied to accounting ratios from the YZ dataset. Similar results are found using the CLZ data (see the Github replication code).

The first three columns construct $|t_i|$ using equal-weighted long-short decile returns,

²Correlations also imply that FDR point estimates are too conservative (Schwartzman and Lin 2011)

³A small standard error is also seen applying the delta method to the Easy Bound, which implies a standard error of $5(1/0.33)^2 \sqrt{0.33(1 - 0.33)/29,000} \approx 0.1$ percentage points.

Table 2: FDR Estimates Controlling for Value Weighting and Factor Adjustments

Data consists of long-short decile strategies formed on 18,000 accounting ratios constructed by Yan and Zheng (2017). t_i tests the null that $\alpha = 0$ using the stock weighting and the factor adjustment in the column header. ‘4-Fac’ is FF3+Momentum. Panel (a) uses Equation (13) to bound $\text{FDR}_{|t_i|>2}$. ‘Pr(null _{i}) Upper Bound’ uses Equation (12). Panel (b) uses Equation (22) to control the FDR following Storey (2002), as implemented in Harvey and Liu (2020). After adjusting for value weighting and factor exposure, at least 77% of findings are true under most specifications, though the bound falls to 59% controlling for FF3 + Momentum exposure. Harvey and Liu’s finding of near-zero “true strategies” can be reconciled by interpreting significant factors as “true factors” and focusing on the most extreme estimate based on the six specifications in YZ.

	Equal-Weighted			Value-Weighted		
	CAPM	FF3	4-fac	CAPM	FF3	4-fac
Panel (a): Visual Bound on $\text{FDR}_{ t_i >2}$						
Share of $ t_i > 2.0$	32.7	34.9	29.3	17.9	17.1	10.5
Share of $ t_i < 0.5$	21.4	19.3	21.5	25.8	29.6	33.4
Pr(null _{i}) Upper Bound	56.4	50.8	56.5	68.0	77.9	87.8
$\text{FDR}_{ t_i >2}$ Upper Bound	8.6	7.3	9.6	19.0	22.8	41.7
Panel (b): Significance using Storey (2002) with $\text{FDR} \leq 10\%$						
t -statistic Hurdle h	1.8	1.7	1.9	2.4	2.5	3.3
Percent Significant	36.1	41.3	31.3	11.3	9.2	1.0

adjusted for exposure to various factor models. The factor model adjustments have little impact on the share of $|t_i| > 2$ or the share of $|t_i| < 0.5$. Thus, the $\text{FDR}_{|t_i|>2}$ bounds are close to the estimates based on raw returns (Figure 2), and range from 7.3% to 9.6%.

The other columns examine value-weighted returns, which do make a difference. While equal-weighted returns consistently imply a share of $|t_i| > 2$ of around 30%, the share under value weighting is much smaller, ranging from 11% to 18%. This leads to larger $\text{FDR}_{|t_i|>2}$ bounds ranging from 21% to 42%, depending on the choice of factor model. Nevertheless, the Visual Bound still implies that the majority of findings are true, even under value weighting. Notably, findings based on value-weighted CAPM alphas are at least 81% likely to be true discoveries.

The 42% FDR bound for the value-weighted 4-factor (FF3 + momentum) alpha is notable. This high FDR is consistent with weaker predictability found in large and liquid stocks (Chen and Velikov 2023), and the fact that momentum shows up in all size groups (Fama and French 2008). One interpretation is that liquidity and momentum are important for understanding cross-sectional predictability, and potentially account for 42% of findings in the literature. Nevertheless, liquidity and momentum still leave the majority of predictability unexplained.

4. Reconciliation with the Literature

The bounds in Sections 2 provide a partial reconciliation. Table 1 shows that $\text{FDR}_{|t|>2} \leq 25\%$, based on many previous studies. This finding is consistent with YZ; Chen and Zimmermann (2018); Jensen, Kelly, and Pedersen (2023); and Chen (2024); who find either a small FDR or many valid predictors.

This section aims to complete the reconciliation. I re-examine papers that find $\text{FDR}_{|t|>2} \gg 25\%$, specifically Harvey, Liu, and Zhu (2016), Harvey and Liu (2020), and Chordia, Goyal, and Saretto (2020). I close by discussing the literature on out-of-sample tests (McLean and Pontiff 2016; YZ; Linnainmaa and Roberts 2018; CLZ).

Unlike previous sections where significance is defined as $|t_i| > 2.0$, this section considers the alternative hurdles. Accordingly, “FDR” in this section refers to notions of significance beyond $|t_i| > 2.0$. “ $\text{FDR}_{|t|>2}$ ” still refers to the classical $|t_i| > 2.0$.

4.1. Reconciliation with Harvey, Liu, and Zhu (2016) (HLZ)

HLZ’s abstract states: “A new factor needs to clear a much higher hurdle, with a t-statistic greater than 3.0. We argue that most claimed research findings in financial economics are likely false.” Reconciling this claim with my estimates centers around the interpretation of terms like “insignificant” and “false discovery.”

To understand the interpretations, consider the model used in HLZ’s simulated method of moments (SMM) estimates (Section 4.1, page 30). HLZ’s model adds parametric assumptions to the framework from Sections 2 and 3. It models t_i using

$$t_i | \mu_i \sim \text{Normal}\left(\frac{\mu_i}{\sigma/\sqrt{N}}, 1\right) \quad (18)$$

$$\text{Corr}(t_i, t_j) = \rho \quad \text{for } i \neq j, \quad (19)$$

where $\mu_i = E(\bar{r}_i)$ is the expected return of factor i , ρ is the correlation between factor returns, σ is the volatility of factor returns, and N is the length of the sample. μ_i is modeled using

$$\mu_i = 0 \quad \text{with prob } p_0 \quad (20)$$

$$\mu_i \sim \text{Exponential}(\lambda) \quad \text{with prob } (1 - p_0), \quad (21)$$

where p_0 is a constant and λ is the scale parameter of the exponential distribution.

In the standard interpretation, a factor is null if $\mu_i = 0$, and it is significant if $|t_i| > h$, where h is a significance hurdle. “Null” and “significant” are related through Equation (18), which implies that μ_i and $|t_i|$ are correlated. Nevertheless, they are distinct concepts. A null factor may very well have $|t_i| > h$. An alternative factor (drawn from

Equation (21)) may have $|t_i| < h$. Multiple testing controls add colorful language (a “discovery” is a significant factor), but the distinction between null and significant remains (e.g. Benjamini and Hochberg 1995). This distinction is embedded in Principle 2 of the American Statistical Association’s “Statement on Statistical Significance and P-Values” (Wasserstein and Lazar 2016).

In HLZ, this distinction is removed. Section 3.2 “A multiple testing framework” states: “[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). The same page defines “false discoveries” as when “we conclude a factor is ‘significant’ when it is not,” and this interpretation is also used in the conclusion (page 37). These definitions imply that HLZ consider a “false discovery” a factor that meets the classical 2.0 hurdle, but fails to meet an alternative hurdle selected by HLZ. Using the 3.0 hurdle from HLZ’s abstract, factors that satisfy $2.0 < |t_i| < 3.0$ are “false discoveries”—regardless of the value of μ_i .

To visualize this distinction, consider Figure 3, which simulates Equations (18)-(21) using HLZ’s SMM estimates (Table 5, $\rho = 0.2$). These estimates come from selecting $\rho = 0.2$, $\sigma = 0.15/\sqrt{12}$ per month, and $N = 240$ to match basic asset pricing facts, and then choosing $p_0 = 0.444$ and $\lambda = 0.555$ to fit the cross-factor distribution of t-stats. I run 1,000 simulations of 10,000 factors. The figure shows 1,378 factors (HLZ’s preferred number, Table 5) drawn from across these simulations, each represented by one marker.⁴

Panel (a) follows the standard interpretation, labelling factors as “Null” (red circles) or “Alt” (blue dots), depending on their position on the y-axis (“Expected Return”). Significance is determined by the position on the x-axis (“|t-statistic|”). 355 factors fall to the right of the classical 2.0 hurdle. 30 of these are null, implying an FDR of only 30/355 $\approx 9\%$ using the classical hurdle for significance. This number is consistent with the Easy Bound applied to HLZ’s summary stats (Table 1) as well as the Visual Bound based on CLZ (Figure 2). HLZ suggest that one should aim for an FDR of 1% (page 24), which one can obtain with a significance hurdle of $|t_i| > 3.0$ (dashed line). Nevertheless, “null” and “significant” are distinct concepts, and require separate axes for plotting.

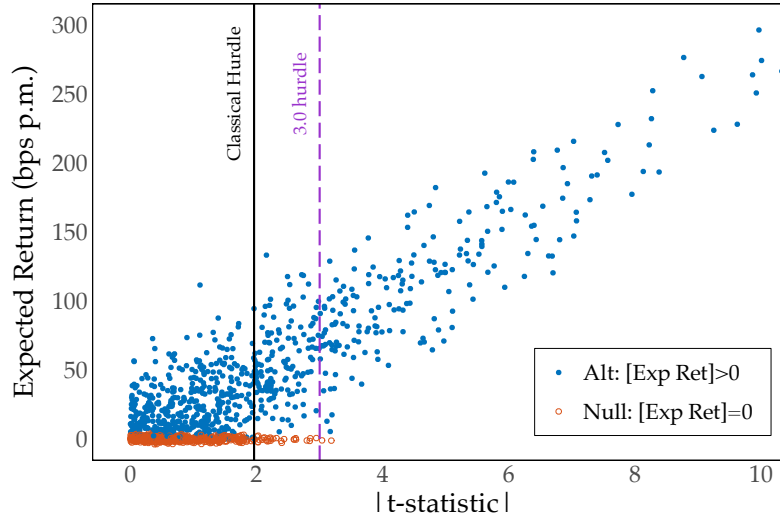
Panel (b) follows HLZ’s interpretation. Using HLZ’s preferred 3.0 hurdle, factors that fall to the left of the dashed line are “null,” regardless of their position on the y-axis. As seen in Equations (18)-(21), the y-axis is not the sample mean return, which may be due to luck, but the fundamental expected return, cured of sampling error. Thus, HLZ’s interpretation leads to questionable economic conclusions, like the idea that a factor with an expected return of 100 bps per month (12% per year) is a “false discovery.”

Panel (b) shows that, even using HLZ’s interpretation of “false discovery,” it is difficult

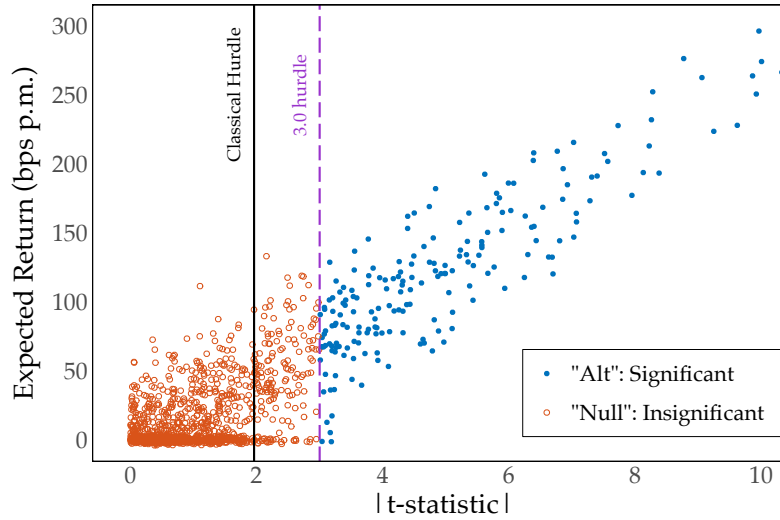
⁴I simulate more factors than HLZ’s preferred number to reduce variation in the Monte Carlo calculations. The FDR is invariant to the number of factors.

Figure 3: Two Interpretations of Harvey, Liu, and Zhu's (2016) SMM Estimates

I simulate HLZ's SMM estimates (Equations (18)-(21)). "Expected Return" is μ_i , "t-statistic" is t_i , and each marker represents one factor. "FDR = 5%" and "FDR = 1%" show t-hurdles from HLZ's Table 5. Panel (a) uses the standard interpretation (Benjamini and Hochberg 1995). Panel (b) follows HLZ, which interprets insignificant factors as "false discoveries." The standard interpretation finds $\text{FDR}_{|t|>2} = 9\%$, in-line with Figure 2. HLZ's interpretation leads to "false discoveries" with expected returns of 12% per year.



(a) Standard Interpretation



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

to conclude that most findings are false. Of the factors that meet the classical 2.0 hurdle, 44% fall to the left of the 3.0 hurdle. Under HLZ's interpretation, this number implies that "many" findings are "false," or that about half are "false." But it does not quite permit the conclusion that *most* findings are false.

To complete the reconciliation, it helps to read HLZ’s conclusion (page 37), which refers to Benjamini and Yekutieli (2001) (BY) Theorem 1.3:

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 significant factors, 158 would be considered false discoveries under Bonferroni [sic], 142 under Holm, 132 under [BY Theorem 1.3] (1%), and 80 under [BY Theorem 1.3] (5%).

This quote presents four false discovery rate estimates: 53% (158/296), 48% (142/296), 45% (132/296), and 27% (80/296), and summarizes them by saying “many” discoveries are likely false. Yet it also arrives at the “same” conclusion as Ioannidis (2005), which argues that *most* claimed research findings are false.

The final step of this argument either selects only the largest out of the four estimates (53%) or makes an interpretive leap from “many” to “most.” Either way, HLZ’s conclusion that “most claimed research findings are likely false” requires an additional nonstandard interpretation beyond equating “insignificant” and “null.”

4.2. Reconciliation with Harvey and Liu (2020) (HL)

HL also examine the YZ data using Storey (2002). But they come to a very different conclusion (page 2531):

Our results suggest that only about 0.1% (or 18) of the 18,000 anomaly strategies in Yan and Zheng (2017) are classified as “true” to control the FDR at 10%.

This 0.1% is two orders of magnitude smaller than the 12.2% to 49.2% share of “true” (more precisely, non-null) strategies in Table 2, which also uses the YZ data and Storey (2002).

The reconciliation once again centers on interpretations of significance. To understand the interpretations, generalize Equation (13) as follows:

$$\text{FDR}(h) \leq \left[\frac{\Pr(|t_i| > h | \text{null}_i)}{\Pr(|t_i| > h)} \right] \left[\frac{\Pr(|t_i| < c)}{\Pr(|t_i| < c | \text{null}_i)} \right], \quad (22)$$

where $|t_i| > h$ generalizes $|t_i| > 2$, $|t_i| < c$ generalizes $|t_i| < 0.5$, and $\text{FDR}(h)$ is the false discovery rate from defining significance with $|t_i| > h$. Storey (2002) uses Equation (22) to construct a hypothesis test: choose a value for c and a desired worst-case FDR q . Then solve for the smallest h that ensures the RHS is less than q .

In the standard interpretation, strategies that satisfy $|t_i| > h$ are “significant” when “controlling” the FDR at q (e.g. Storey and Tibshirani 2003). But HL follow HLZ’s interpretation, and describe strategies that satisfy $|t_i| > h$ as “true.”

Table 2 Panel (b) takes a closer look, by fixing $c = 1.28$ (equivalent to HLZ’s $\theta = 0.8$ on page 2531) and solving for the smallest h that ensures the RHS is less than 10%. It examines all performance measures from YZ’s Table 1, which includes both equal- and value-weighted specifications. The resulting hurdles vary widely, as do the shares of strategies that meet the hurdles.

The table uses the standard interpretation, and reports that 30% to 40% of strategies are “significant” using equal-weighting, but only about 10% are significant under most value-weighted specifications. These significance rates are smaller than the share of non-null strategies, which is roughly 50% under equal-weighting, and ranges from 12% to 32% under value-weighting.

In HLZ’s interpretation, 30% to 40% of equal-weighted strategies are “true,” and about 10% are true under most value-weighted specifications. This helps with the reconciliation, but these numbers are still two orders of magnitude larger than HL’s 0.1%.

To reconcile further, note there is an outlier in Table 2: if one measures predictability using value-weighting *and* adjusts for FF3 + Momentum exposure, only 1.0% of strategies are “true” using HLZ’s interpretation. This 1.0% is almost the same magnitude as HL’s 0.1% and is consistent with HL’s footnote 18, which states “[w]e focus only on the Fama-French-Carhart four-factor model to save space.”

Thus, one can largely reconcile HL’s claims by interpreting “significant” as “true” and selecting the smallest estimate based on the six specifications in YZ.

4.3. Reconciliation with Chordia, Goyal, and Saretto (2020)

CGS report the headline FDR number in their abstract:

We estimate the expected proportion of false rejections that researchers would produce if they failed to account for multiple hypothesis testing to be about 45%.

This conflicts with Equation (7), which shows that this proportion is at most 25%.

CGS use standard statistical language. Throughout their paper, a “rejection” is a discovery, and a false rejection is a discovery that is null (e.g. pages 2135 and 2153). So the reconciliation is not in the interpretation.

Instead, the reconciliation lies in CGS’s calibration procedure. The headline number comes from a model of return signals, in which π fraction of signals are not null (“informative,” page 2142). They calibrate π to their data-mined returns, arriving at $\pi = 0.06\%$ (page 2154), implying $\Pr(\text{null}_i) = 99.94\%$.

$\Pr(\text{null}_i) = 99.94\%$ is difficult to square with the law of total probability and figures reported in CGS. As shown in Equation (12), total probability implies bounds on $\Pr(\text{null}_i)$ of the form:

$$\Pr(\text{null}_i) \leq \frac{\Pr(|t_i| < c)}{\Pr(|t_i| < c | \text{null}_i)} \quad (23)$$

where c is a constant. CGS show in Figure 1 that at least 100,000 of 2.39 million t-stats fall in the bin containing zero. Inspecting the figure closely shows that at most 11 bins fit between zero and 2.0, and thus the width of the bin that contains zero is at least $2/12 = 0.166$. Plugging these values into Equation (23) leads to

$$\Pr(\text{null}_i) \leq \frac{(1 \times 10^5) / (2.39 \times 10^6)}{\Pr(|t_i| < (0.166/2) | \text{null}_i)} = 63\%, \quad (24)$$

not far from the bounds in Table 2, and far lower than the 99.94% calibrated by CGS. The intuition can be seen by applying Algorithm 1 to Figure 1 of CGS. The largest standard normal distribution that still fits inside this histogram will fail to cover the tails, which extend well beyond -2.0 and +2.0.

So how does CGS's calibration arrive at $\pi = 0.06\%$? They calibrate π to match the share of signals that exceed the $|t_i| > 2$ threshold but fail a more stringent threshold based on Romano and Wolf (2007) (page 2153). This target moment contains no information about small t-stats, which, as seen in Figure 2, are highly informative about $\Pr(\text{null}_i)$. As shown in Equation (23), if CGS had instead targeted the share of small t-stats, they would have arrived at a much higher π . Since CGS choose the dispersion of informative alphas outside of their calibration (page 2152), this much higher π would then imply a much lower headline FDR.

4.4. Reconciliation with Out-of-Sample (OOS) Tests

Assuming no structural changes, the mean OOS return of a false discovery is zero. Thus, if the FDR is more than $X\%$, we would expect at least $X\%$ return decay in OOS tests.

At a cursory glance, the literature on OOS tests finds mixed evidence, with decay ranging from 0% (Jacobs and Müller 2020) to 60% (Linnainmaa and Roberts 2018). But a closer look reveals that the decay is consistently less than 36%, if one focuses on OOS tests that are temporally and methodologically close to the original tests. Using OOS periods that end at the publication dates, and using portfolio tests that generally follow the original papers, Chen and Zimmermann (2020), McLean and Pontiff (2016), and Jacobs and Müller (2020) find OOS decay of 20%, 26%, and 36%, respectively. These numbers suggest that at least 36% of discoveries are true.

Conceptually, OOS tests that are close to the original tests are the most appropriate for validating FDR estimates. OOS samples that are far from the original samples introduce economic effects that are distinct from multiple testing. For example, improvements in information technology can reduce predictability (Chordia, Subrahmanyam, and Tong 2014), while the introduction of retail traders can increase it (Goetzmann and Huang 2018). Though these effects are important, they are economic rather than statistical, and should not be measured using FDR methods.

With the large datasets available from data mining, one can examine OOS periods that are even closer to the in-sample periods. CLZ find OOS decay as small as 17% among data-mined predictors by zooming in on the 12-months after the in-sample periods. YZ and Chen and Dim (2024) use data mining to produce OOS returns of roughly 50 bps per month. This mild decay and strong OOS performance are quite consistent with the $\text{FDR}_{|t|>2} \leq 25\%$ derived in this paper.

5. Conclusion and Limitations

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 can be explained in just a few equations and provide clarity to the FDR literature. Using data from many papers, they show that the FDR cannot exceed 25%. This small FDR comes from a surprising fact: one out of five random accounting ratios produces statistically-significant predictability. If most discoveries were false, finding predictability would not be so easy.

Claims that the FDR exceeds 25% are driven by interpreting “insignificant factors” as “false discoveries” and selecting the extremes among multiple estimates. Neglecting to fit moments that are highly informative about the FDR also results in overestimates.

As highlighted by the title of this paper, my analysis is limited to statistics. A “true discovery” is a statistically significant predictor that has a non-zero alpha. Many economic issues are buried in this statement, like trading costs (Novy-Marx and Velikov (2016)), information costs (Chordia, Subrahmanyam, and Tong (2014)), and arbitrage effects (McLean and Pontiff (2016)). Chen and Velikov (2023) find that, net of these effects, most findings in cross-sectional predictability produce measly profits—even if they are true, statistically speaking.

References

- Bali, Turan G, Robert F Engle, and Scott Murray (2016). *Empirical asset pricing: The cross section of stock returns*. John Wiley & Sons.
- 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.
- Benjamini, Yoav and Daniel Yekutieli (2001). “The control of the false discovery rate in multiple testing under dependency”. *Annals of statistics*, pp. 1165–1188.
- Bessembinder, Hendrik, Aaron Burt, and Christopher M Hrdlicka (2025). “Can we tame the factor zoo? The roles of alternative databases and extraction methods”. *Can we tame the factor zoo*.
- Chen, Andrew Y (2021). “The Limits of p-Hacking: Some Thought Experiments”. *The Journal of Finance*.
- (2024). “Do t-Statistic Hurdles Need to be Raised”. *Management Science*.
- Chen, Andrew Y and Chukwuma Dim (2024). “High-Throughput Asset Pricing”. *arXiv preprint arXiv:2311.10685*.
- 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 (2018). *Publication bias and the cross-section of stock returns*. <https://www.federalreserve.gov/econres/feds/files/2018033pap.pdf>.
- (2020). “Publication bias and the cross-section of stock returns”. *The Review of Asset Pricing Studies* 10.2, pp. 249–289.
- Chen, Andrew Y., Alejandro Lopez-Lira, and Tom Zimmermann (2025). *Does Peer-Reviewed Research Help Predict Stock Returns?* arXiv: 2212.10317 [q-fin.GN].
- Chen, Andrew Y. and Tom Zimmermann (2022). “Open Source Cross-Sectional Asset Pricing”. *Critical Finance Review* 27.2, pp. 207–264.
- Chordia, Tarun, Amit Goyal, and Alessio Saretto (2020). “Anomalies and false rejections”. *The Review of Financial Studies* 33.5, pp. 2134–2179.
- Chordia, Tarun, Avanidhar Subrahmanyam, and Qing Tong (2014). “Have capital market anomalies attenuated in the recent era of high liquidity and trading activity?” *Journal of Accounting and Economics* 58.1, pp. 41–58.
- Cochrane, John H (2009). *Asset pricing: Revised edition*. Princeton university press.
- Efron, Bradley (2003). “Second thoughts on the bootstrap”. *Statistical science*, pp. 135–140.
- (2010a). “Correlated z-values and the accuracy of large-scale statistical estimates”. *Journal of the American Statistical Association* 105.491, pp. 1042–1055.
- (2010b). “Rejoinder-Correlated z-Values and the Accuracy of Large-Scale Statistical Estimates”. *Journal of the American Statistical Association* 105.491, p. 1067.

- Fama, Eugene F and Kenneth R French (2008). "Dissecting anomalies". *The Journal of Finance* 63.4, pp. 1653–1678.
- (2010). "Luck versus skill in the cross-section of mutual fund returns". *The journal of finance* 65.5, pp. 1915–1947.
- Goetzmann, William N and Simon Huang (2018). "Momentum in Imperial Russia". *Journal of Financial Economics*.
- 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 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.
- 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.
- 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.
- Novy-Marx, Robert and Mihail Velikov (2016). "A taxonomy of anomalies and their trading costs". *Review of Financial Studies* 29.1, pp. 104–147.
- Romano, Joseph P and Michael Wolf (2007). "Control of Generalized Error Rates in Multiple Testing". *The Annals of Statistics*, pp. 1378–1408.
- Schwartzman, Armin and Xihong Lin (2011). "The effect of correlation in false discovery rate estimation". *Biometrika* 98.1, pp. 199–214.
- 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.
- Storey, John D and Robert Tibshirani (2003). "Statistical significance for genomewide studies". *Proceedings of the National Academy of Sciences* 100.16, pp. 9440–9445.
- 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.