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

Andrew Y. Chen

Federal Reserve Board

October 2021*

Abstract

Harvey, Liu, and Zhu (2016) "argue that most claimed research findings in financial economics are likely false." Surprisingly, their false discovery rate (FDR) estimates suggest most are true. I revisit their results by developing non- and semi-parametric FDR estimators that account for publication bias and empirical correlations. These estimators provide simple closed-form expressions and reliably produce an upper bound on the FDR in simulations that cluster-bootstrap from empirical predictor returns. Applying these estimators to the Chen-Zimmermann dataset of 205 predictors, I find that most claimed statistical findings in the cross-sectional predictability literature are likely true.

JEL Classification: G0, G1, C1

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

*First posted to SSRN: August 27, 2021. I thank Antonio Gil de Rubio Cruz and Rebecca John for excellent research assistance. I thank Jeffrey Pontiff and seminar participants at Boston College for helpful comments. The views expressed herein are those of the authors and do not necessarily reflect the position of the Board of Governors of the Federal Reserve or the Federal Reserve System.

1. Introduction

In an influential paper, Harvey, Liu and Zhu (2016) (HLZ) "argue that most claimed research findings in financial economics are likely false." This argument is based on multiple testing statistics—methods that account for the fact that dozens, or even hundreds of tests may underlie the data. The existence of many tests violates classical assumptions, suggesting that the false discovery rate (FDR) may be much higher than suggested by classical methods.

Surprisingly, HLZ's FDR estimates imply almost the opposite of their argument. HLZ find that a t-stat hurdle of around 3.0 implies an FDR ≤ 5% (pages 22, 24, and 30). Since most asset pricing t-stats exceed 3.0 (Chen 2021), this hurdle implies that most findings have a "true discovery rate" exceeding 95%. In other words, most claimed findings are likely true.

These surprising results warrant a thorough and robust re-examination. In this paper, I perform such a re-examination by developing non- and semi-parametric FDR estimators and applying them to the Chen and Zimmermann (Forthcoming) (CZ) dataset of published cross-sectional stock return predictors. These estimators build on the same framework used by HLZ's preferred Benjamini and Yekutieli (2001) (BY) algorithm. Unlike HLZ, I derive my estimators from a statistical model of publication bias, and provide simple expressions that explain why the FDR is so small. Additionally, I verify my estimators in simulations that closely match the empirical dependence in the CZ data. These simulations address concerns about correlations that are left unanswered by HLZ. Lastly, I make code available at https://github.com/chenandrewy/mostly-true.[1]

My re-examination consistently confirms HLZ's numerical result: a t-hurdle of 3.0 implies an FDR ≤ 5%. I extend this finding and show that for lower t-stat hurdles, the FDR remains relatively small. In my preferred estimate, the FDR among all CZ predictors is at most 30%. Overall, my findings thoroughly support the surprising conclusion that most claimed statistical findings in cross-sectional asset pricing are likely true.

In the remainder of the introduction, I summarize the intuition, the robustness to correlations, and the relation to the literature.

[1] This code automatically downloads the CZ data, which is available at https://www.openassetpricing.com/.

1.1. Summary of Intuition

The main result can be derived in just four lines of math.

Suppose we define "discoveries" as predictors with absolute t-stats $|t_i|$ exceeding 2.60. If $|t_i|$ is well-behaved (a Glivenko-Cantelli theorem holds),[2] the FDR among these discoveries satisfies

$$\text{FDR}(|t_i| > 2.60) \approx Pr\left(F_i||t_i| > 2.60\right),$$

where F_i is the event that predictor i is false. Intuitively, the FDR is the probability that a predictor is false, given that the predictor is declared a discovery. Applying Bayes rule and noting that $Pr\left(F_i\right) \leq 1$ yields

$$\begin{aligned} \text{FDR}(|t_i| > 2.60) &\approx \frac{Pr\left(|t_i| > 2.60|F_i\right)}{Pr\left(|t_i| > 2.60\right)} Pr\left(F_i\right) \\ &\leq \frac{Pr\left(|t_i| > 2.60|F_i\right)}{Pr\left(|t_i| > 2.60\right)}. \end{aligned} \tag{1}$$

The numerator is just the p-value corresponding to $|t_i| = 2.60$. Thus, if we have an estimate of $Pr\left(|t_i| > 2.60\right)$, we have an upper bound on the FDR.

Publication bias, however, means that the sample counterpart of $Pr\left(|t_i| > 2.60\right)$ is upward-biased. One way to handle this problem is to use external data. Yan and Zheng (2017) (YZ) show that, among trading strategies built off of random combinations of accounting variables, at least 15% of strategies have $|t_i| > 2.60$ (see YZ's Table 1). Thus, as long as the research process is more likely to generate $|t_i| > 2.60$ than random mining, we have the following bound on the FDR

$$\text{FDR}\left(|t_i| > 2.60\right) \leq \frac{1\%}{0.15} \approx 7\%, \tag{2}$$

where 1% is the p-value corresponding to $|t_i| = 2.60$. Since 70% of $|t_i|$ exceed 2.60, most claimed findings are at least 93% likely to be true.

Alternatively, one can address publication bias by fitting a parametric model to $|t_i|$ and then applying the non-parametric Equation (1). Fitting such a model amounts to extrapolating unobserved $|t_i|$ near zero. While this method sounds

[2] Glivenko-Cantelli theorems extend laws of large numbers (the convergence of sample moments) to cover the convergence of empirical distribution functions. Like the law of large numbers, Glivenko-Cantelli holds under weak dependence.

dangerous, it can be made conservative by assuming that $|t_i|$ has a mode at 0—that is, assuming the typical predictor has absolutely no predictive power, even in-sample.

Appendix A.1 of HLZ provides one such conservative estimate. They fit an exponential density to the right tail of their hand-collected t-stats, and consistently find a scale parameter of around 2.0, implying that

$$\text{FDR}\left(|t_i| > 2.60\right) \leq \frac{1\%}{Pr\left(|t_i| > 2.6\right)} = \frac{1\%}{\exp\left(-2.6/2.0\right)} \approx 4\%.$$

HLZ express doubts about this estimate due to the "underlying assumption" of "independence among t-statistics." However, I show that such an assumption is not necessary and that more conservative assumptions also lead to a small FDR.

In summary, the FDR is small because the research process readily generates very large t-stats compared to a standard normal distribution. This property can be seen in the frequency of large t-stats in random accounting-based strategies (Yan and Zheng (2017)) or in conservative extrapolations of the extremely large t-stats of published strategies. Either way, the multiple testing adjustment to the p-value is quite moderate.

1.2. Robustness to Correlations

Equation (2) makes minimal assumptions about correlations. I assume only that the share of $|t_i|$ that exceeds 2.60 is a good estimate of the probability that $|t_i|$ exceeds 2.60. Thus, the kind of weak dependence assumed in GMM is all that is required. That weak dependence is sufficient for the validity of simple FDR estimates was first demonstrated by Storey, Taylor and Siegmund (2004) (see also Ferreira and Zwinderman (2006); Genovese et al. (2006); Farcomeni (2007)).

This result contrasts with HLZ's statement that the Benjamini-Hochberg algorithm is "only valid when the test statistics are independent or positively dependent." This statement, however, is not true. Independence or positive regression dependence are *sufficient* conditions for Equation (2), but they are not necessary (see Theorems 1.2 and 1.3 of Benjamini and Yekutieli (2001)).

To verify that the data display sufficiently weak dependence, I simulate data by bootstrapping from residuals from the CZ data. The bootstraps are clustered so that residuals from the same month are always drawn together, thus ensuring that the simulation inherits the empirical correlation structure. Indeed, I show

that the distribution of pairwise correlations in the simulations closely matches the distribution in the data.

Estimations on these simulations show that the semi-parametric estimates reliably place an upper bound on the actual FDR. This reliability holds even in simulations in which 99% of predictors are false and true predictors return only 25 bps per month. It also holds under extreme forms of publication bias.

The simulations show that extrapolating from the observed $|t_i|$ distribution is safe under a realistic range of simulations. Thus, my semi-parametric estimates provide a simple alternative to the more complicated methods that impose more realistic structure (e.g. HLZ's model with correlations, Chen (2020)). Indeed, my preferred semi-parametric estimator can be calculated entirely by hand.

These results show that HLZ's preferred Benjamini and Yekutieli (2001) algorithm is excessively conservative. Though the finance literature frequently follows HLZ's use of Benjamini-Yekutieli (e.g. Jensen et al. (2021)), the statistics literature generally finds the implied FDR penalty is not necessary (Efron (2012)), consistent with my theoretical and simulation results.

1.3. Relation to the Literature

This paper addresses only the non- and semi-parametric methods in HLZ. These conservative methods *assume* that t-hurdles should be raised. Thus they cannot address the question of *whether* t-stat hurdles should be raised. I address this broader question with a re-examination of HLZ's parametric estimates in a companion paper (Chen (2020)).

HLZ's estimates imply that most findings are likely true, but this result is not easily discerned. From the abstract and introduction, the reader can see that t-hurdles should be raised to 3.0, but no other numerical results can be found. The next 18 pages describe the methodology. The main figure (on page 21) does not provide the necessary information either. The reader must dig into the text on page 22 and simulate HLZ SMM estimates to figure out that for 70% of findings, the FDR $\leq$ 5%. But even then, she would have doubts, as HLZ present a multitude of t-hurdle estimates, their primary estimates abstract from publication bias, and because they express unresolved concerns about correlations. The contribution of my paper is to cleanly tie all of these strands together, rigorously account for publication bias and correlations, and robustly demonstrate that most claimed

findings are likely true. I also show how external data can be used to bound publication bias effects and provide closed form expressions that illustrate the intuition.

Other followups to HLZ impose significant structure, resulting in relatively complicated estimations (Chen (2020); Chen and Zimmermann (2020); Chordia, Goyal and Saretto (2020); Giglio, Liao and Xiu (2021); Harvey and Liu (2021); Jensen, Kelly and Pedersen (2021); Zhu (2021)).[3] Nevertheless, these more structured estimates largely find that published cross-sectional predictors are largely true. For example, Chen and Zimmermann (2020) and Chen (2020) use maximum likelihood and SMM to estimate that in-sample mean returns are biased upward by around 15% and that the FDR is at most 20%. Harvey and Liu (2021) examine a multi-step mixture-bootstrap model and find FDRs of 10%, 21%, and 38% across three calibrations. My closed-form expressions provide the intuition behind these less transparent estimation procedures.

An important caveat is that I only examine the numerical findings of the cross-sectional literature. I do not assess whether these numerical findings address the written claims. Indeed, many of these numerical findings rely on the trading of illiquid stocks (Chen and Velikov (2021)), and many do not provide sufficient evidence for their economic conclusions (Kozak, Nagel and Santosh (2018)). More broadly, HLZ's claim that most research findings are likely false may still be correct. However, verifying this claim seems to be outside of the realm of multiple testing statistics.

2. Data and Statistical Framework

I describe the data (Section 2.1) and the statistical framework (Section 2.2). Along the way, I define key concepts like "statistical finding," "publication bias," and the "false discovery rate."

2.1. Data

I use the Chen and Zimmermann (Forthcoming) dataset of 205 reproduced cross-sectional stock return predictors. These data draw from a larger set of 319 characteristics examined in other meta-studies. Based on their reading of the

[3] An exception is Chen (2021), but that paper focuses on the null of no predictability anywhere, and cannot formally estimate the FDR.

original papers, CZ judge that these 205 predictors should produce t-stats exceed 1.96 in absolute value in long-short portfolio tests. These tests, in turn, attempt to follow the original papers' methods as closely as possible. This selection defines my notion of a "statistical finding."

The t-stats from CZ's long-short tests are shown in Table 1. Notably, 60% of predictors have $|t_i|$ that exceed HLZ's recommended t-stat hurdle of 3.0. The lowest decile $|t_i|$ is 1.92, indicating that almost all predictors are significant in the traditional sense. This is clearly the result of "publication bias"—that is, the idea that results with larger t-stats are more likely to be selected for sharing.[4] Indeed, as described above, the very definition of a "statistical finding" embeds publication bias.

[Table 1 about here]

The bottom panel of the Table 1 examines the distribution of pairwise correlations across monthly long-short returns. HLZ argue for using the highly conservative Benjamini and Yekutieli (2001) (BY) algorithm for FDR control to account for these correlations. Simu-theoretical evidence suggests that conservative dependence controls are mostly relevant for correlations close to 1.0 or -1.0 (Reiner-Benaim (2007)). However, this panel shows that these correlations cluster around zero. This clustering happens whether I use the longest overlapping in-sample data ("overlapping in-sample"), or restrict the data to be a balanced panel.

These average-zero correlations run counter to the common perception in finance that there is a strong factor structure in the cross-section of returns. While this result is certainly true for Fama and French's size and B/M sorted portfolios (Lewellen, Nagel and Shanken (2010)), many papers find that long-short returns like CZ's have typically zero correlation (McLean and Pontiff (2016)), and that many components are required to summarize anomaly data (Jensen, Kelly and Pedersen (2021)).

More broadly, these correlations suggest that the process of that generates $|t_i|$ is well-behaved. A large spike of correlations near 1.0 or -1.0 would bring about concerns that the process is in a sense non-stationary. In contrast, the moderate correlations in Table 1 suggest that fundamental theorems like the weak law of

[4] "p-hacking," "data-snooping," and "data-mining" all have extremely similar definitions. See Chen (2021) for a richer discussion.

large numbers should hold. This observation motivates the statistical framework that follows.

2.2. Statistical Framework

Suppose the cross-sectional predictability literature is generated in two steps. The first step generates N "unbiased" predictors with accompanying t-stats $t_1, t_2, ..., t_N$. N_F of these unbiased predictors are false, and these false predictors satisfy

$$t_i | F_i \sim \text{student's-t}(\nu) \tag{3}$$

where the event F_i indicates predictor i is false and ν is the degrees of freedom parameter. I do not assume independence in Equation (3).

In the second step, N_S predictors are selected for sharing based on the size of the t-stat. Let S_i denote the event that predictor i is selected, and assume

$$Pr(S_i | t_i) = s(|t_i|) \tag{4}$$

$$s(\cdot) \text{ weakly increasing}$$

$$s(|t|) = \bar{s}, \quad \text{for } |t| > t_{\text{good}} \tag{5}$$

where $s(\cdot)$ is a function that takes on values in $[0,1]$ and t_{good} is a real number. Equations (4)-(5) imply that a larger t-stat implies selection is more likely, but only up to a point (t_{good}). One can include additional variables that contribute to selection in Equation (4), but these can be averaged out leading to an equivalent form (e.g. Chen (2021)).

The idea that literature generation is separated this way is certainly an abstraction. In reality, it is likely that the two steps occur simultaneously. Nevertheless, this abstraction can be a useful framework for conceptualizing publication bias. One can think of the unbiased predictors as the ideal set of predictors the meta-econometrician would like to have—though this set may not exist in reality. Appendix A.1 shows how a model in which these two steps occur simultaneously can be "orthogonalized" into the framework above.

I assume that the joint process $\{t_i, F_i, S_i\}_{i=1,...,N}$ is well-behaved in the follow-

ing sense:

$$\frac{1}{N}\sum_{i=1}^{N} I\left(|t_i| \geq \bar{t}\right) \xrightarrow{p} Pr\left(|t_1| \geq \bar{t}\right) \tag{6}$$

$$\frac{1}{N_F}\sum_{i=1}^{N} I\left(t_i| \geq \bar{t} \cap F_i\right) \xrightarrow{p} Pr\left(|t_1| \geq \bar{t}|F_1\right) \tag{7}$$

$$\frac{1}{N_S}\sum_{i=1}^{N} I\left(t_i| \geq \bar{t} \cap S_i\right) \xrightarrow{p} Pr\left(|t_1| \geq \bar{t}|S_1\right). \tag{8}$$

where $I(\cdot)$ is an indicator function that takes a value of 1 if the argument is true and zero otherwise.

Equations (6)-(8) state that if we keep counting the share of t-stats that exceed the a hurdle (LHS), we'll eventually recover the probability that a given t-stat exceeds the hurdle (RHS). This kind of condition is important for interpreting order statistics like those shown in Table 1. Indeed, if Equations (6)-(8) do not hold, interpreting Table 1 is tricky in the way that interpreting the deciles of the level of U.S. GDP is tricky.

More formally, Equations (6)-(8) hold if the joint process is strongly stationary (Tucker (1959)), or under a wide variety of general weak dependence conditions that are quite technical (Farcomeni (2007)).[5] Empirically, the correlations shown in Table 1 suggest that Equations (6)-(8) are valid, and I verify this assumption in bootstrap simulations in Section 4.2.

Following Benjamini and Hochberg (1995), define the FDR as

$$\text{FDR}_N\left(\bar{t}\right) \equiv \mathbb{E}\left\{\frac{\sum_{i=1}^{N} I\left(|t_i| > \bar{t} \cap F_i\right)}{\sum_{i=1}^{N} I\left(|t_i| > \bar{t}\right)}\right\} \tag{9}$$

where for simplicity I assume that the denominator is positive. Equation (9) has an intuitive definition: it is the expected fraction of "discoveries" (predictors that exceed $\bar{t}$) that are also false.

This intuitive definition means that the FDR is well-suited to address HLZ's argument that most claimed findings are likely false. HLZ's argument can formally defined as "for a hurdle $\bar{t}$ that covers more than 50% of claimed findings, $\text{FDR}_N\left(\bar{t}\right) > 0.50$." Due to this interpretability, I focus on the FDR. In contrast, the Bonferroni and Holm procedures studied by HLZ measure $\text{FWER}_N\left(\bar{t}\right) \equiv$

[5] Equations (6)-(8) are the result of Glivenko-Cantelli theorems, which are a fundamental result in empirical process theory that extend the laws of large numbers to empirical distributions, and are commonly used in machine learning.

$Pr\left(\sum_{i=1}^{N} I\left(|t_i| > \bar{t} \cap F_i\right) > 0\right)$. In words, the FWER is the probability of having just one false discovery. This quantity says very little about whether most claimed findings are likely false.

I denote $\text{FDR}_N\left(\bar{t}\right)$ with the subscript N to emphasize an unusual aspect of the Benjamini-Hochberg framework: this fundamental quantity $\text{FDR}_N\left(\bar{t}\right)$ is defined with respect to a small sample. This differs from statistical objects used in finance, which are typically independent of sample size (e.g. expected returns).

Indeed, many followups to Benjamini and Hochberg (1995) (e.g. Storey, Taylor and Siegmund (2004); Genovese, Roeder and Wasserman (2006)) focus on the large sample FDR:

$$\text{FDR}\left(\bar{t}\right) \equiv \operatorname*{plim}_{N \longrightarrow \infty} \text{FDR}_N\left(\bar{t}\right). \tag{10}$$

This large sample FDR has many appealing properties. Given the regularity conditions Equations (6)-(8), the large sample FDR boils down to a simple and intuitive expression:

$$\begin{aligned}
\text{FDR}\left(\bar{t}\right) &= \operatorname*{plim}_{N \longrightarrow \infty} \mathbb{E}\left\{ \frac{\left[N_F^{-1} \sum_{i=1}^{N} I\left(|t_i| > \bar{t} \cap F_i\right)\right]}{\left[N^{-1} \sum_{i=1}^{N} I\left(|t_i| > \bar{t}\right)\right]} \left[\frac{N_F}{N}\right] \right\} && (11)\\
&= \frac{Pr\left(|t_1| > \bar{t} | F_1\right)}{Pr\left(|t_1| > \bar{t}\right)} Pr\left(F_1\right) && (12)\\
&= Pr\left(F_1 || t_1| > \bar{t}\right), && (13)
\end{aligned}$$

where the second line applies the regularity conditions to each square bracket. In other words, the FDR is just the probability that a given predictor is false, given that the predictor's t-stat exceeds the hurdle $\bar{t}$. Moreover, the RHS of Equation (13) is equivalent to the empirical Bayes FDR (Efron et al. (2001)), and Storey's (2002) pFDR leads to the same expression in large samples (Storey, Taylor and Siegmund (2004)).

Thus, for the remainder of the paper, I focus on the large sample FDR defined in Equation (10). This approach reduces ambiguity, as Benjamini-Hochberg, Storey, and empirical-Bayes FDRs are all equivalent in this limit. Moreover, this large N limit is appropriate given that one of the main concerns in asset pricing is that N is "too large" in some sense. As noted by Benjamini (2008), the distinction between these approaches emerges if $Pr\left(F_1\right) = 1$ is actually possible, which is unlikely to be the case for cross-sectional predictability, as we will see shortly.

3. Non-Parametric FDR Estimates

This section presents the main result: simple and conservative FDR estimates that adjust for publication bias. Though the calculation is simple (Section 3.3), the theoretical justification takes some explanation (Section 3.1-3.2). I also compare my estimates with HLZ's (Section 3.4).

3.1. A Non-Parametric Estimator

Noting that $Pr(F_1) \leq 1$, Equation (12) implies the following upper bound on the FDR:

$$\text{FDR}(\bar{t}) \leq \frac{Pr(|t_1| > \bar{t}|F_1)}{Pr(|t_1| > \bar{t})}. \tag{14}$$

The numerator is just the p-value corresponding to a t-stat of $\bar{t}$ (Equation (3)). The denominator is the right tail of unbiased findings, suggesting a naive estimator for the RHS:

$$\widehat{\text{FDR}}_{\text{naive}}(\bar{t}) \equiv \frac{Pr(|t_1| > \bar{t}|F_1)}{\sum_{i=1}^{N} I(|t_i| > \bar{t} \cap S_i)/N_s}. \tag{15}$$

That is, the naive estimate just replaces $Pr(|t_1| > \bar{t})$ in Equation (14) with its observed sample counterpart, the observed share of findings with $|t_i| > \bar{t}$.

Selection bias, however, means that the observed sample counterpart of $Pr(|t_1| > \bar{t})$ is upward biased:

$$\sum_{i=1}^{N} I(|t_i| > \bar{t} \cap S_i)/N_s \xrightarrow{p} Pr(|t_1| > \bar{t}) \underbrace{\left[\frac{Pr(|t_1| > \bar{t}|S_1)}{Pr(|t_1| > \bar{t})}\right]}_{>1}.$$

The term in the square brackets is larger than 1 because publication selects for large t-stats (Equation (4)). The trick is to offset the square bracket by multiplying $\widehat{\text{FDR}}_{\text{naive}}(\bar{t})$ with a penalty $\hat{c}$:

Proposition 1. *Let the publication bias adjusted estimator be*

$$\widehat{FDR}_{adj} \equiv \hat{c}\widehat{FDR}_{naive}(\bar{t}) \tag{16}$$

where $\hat{c}$ is chosen to satisfy

$$\underset{N\longrightarrow\infty}{\text{plim}}\ \hat{c} \geq \frac{1}{Pr\left(|t_1| > t_{good}\right)}. \tag{17}$$

Then for $\bar{t} \geq t_{good}$, $\widehat{FDR}_{adj}$ converges to an upper bound on the FDR

$$\underset{N\longrightarrow\infty}{\text{plim}}\ \widehat{FDR}_{adj} \geq FDR\left(\bar{t}\right). \tag{18}$$

The proof is in Appendix A.2. Intuitively, the bias $\frac{Pr(|t_1|>\bar{t}|S_1)}{Pr(|t_1|>\bar{t})}$ is close to the ratio of observed t-stats to total t-stats. This ratio can be bounded if we assume, conservatively, that only t-stats that exceed t_{good} are observed.

Notably, this FDR estimate does not depend on the total number of t-stats N. This feature stands in contrast to the common intuition that more tests imply more false discoveries, but is consistent with the FDR literature. In general, the FDR depends on the *distribution* of t-statistics, while the total number of t-statistics is in a way a nuisance parameter (for example, Genovese et al. (2006)). Similar results are seen in more recent papers that use a similar statistical framework to study publication bias (Andrews and Kasy (2019); Chen and Zimmermann (2020)). Perhaps there is an economic reason for assuming that larger N implies more false discoveries (decreasing returns to the production of predictor ideas), but this kind of modeling cannot be addressed with multiple testing statistics alone.

3.2. Implementation of Proposition 1

To implement Proposition 1, I need both $\hat{t}_{\text{good}}$ (an estimate of t_{good}) and a method for estimating a bound on $Pr\left(|t_1| > \hat{t}_{\text{good}}\right)$. I also need a choice for the degrees of freedom parameter $\hat{\nu}$.

I select $\hat{t}_{\text{good}} = 2.60$. This choice can be motivated several ways.

1. 2.60 is more than half a standard error beyond the ubiquitous 5% threshold for "statistical significance." In this sense, t-stats that exceed 2.60 are more than marginally significant.

2. 2.60 is close to the peak of the empirical t-stat distribution (see Figure 3). While, publication bias clearly plays a role in the left shoulder before 2.60, it seems less likely to play a role beyond that.

3. 2.60 is very close to the HLZ's choice of 2.57. In HLZ's model with correlations and the missing t-stat extrapolation in their Appendix (see also Harvey (2017)), they assume that all $|t_i| > 2.57$ are observed, implying that there is little incremental publication bias beyond this value. HLZ seem to choose 2.57 because this is the t-stat reported in Fama and MacBeth (1973)'s Table III.

One may be tempted to selected $\hat{t}_{\text{good}} > 2.60$ to be conservative, but Proposition 1 is already a conservative estimate of the FDR. This estimate effectively assumes that $Pr(F_1) = 1$ and $Pr\left(S_1 | |t_1| < t_{\text{good}}\right) = 0$. These assumptions overstate the rate of false discoveries and the effect of selection, respectively. As a result, choosing a larger $\hat{t}_{\text{good}}$ would likely decrease power more than it would improve robustness.

Given my choice of $\hat{t}_{\text{good}} = 2.60$, I estimate a bound on $Pr\left(|t_1| > \hat{t}_{\text{good}}\right)$ with the following identifying assumption: I assume that the unbiased research process (the first step in the model at the beginning of Section 2.2) is more likely to generate $|t_i| > \hat{t}_{\text{good}}$ than Yan and Zheng's algorithm for forming long-short portfolios from random combinations of accounting variables. A skeptical reader may doubt the ability of finance theory to produce good predictors, but even these readers would likely admit that adding common sense would not worsen performance relative to randomly combining accounting variables.

Moreover, this method is also almost entirely non-parametric, and is easily accessible. To implement this method, I simply look up values from YZ's Table 1. I focus on the equal-weighted 1-factor α specification, as it is closest to the simple long-short strategies studied in CZ. YZ's Table 1 shows that the 10th and 95th percentile t-stats for this class of strategies are -3.48 and 3.19, respectively. These values imply that $Pr\left(|t_1| > \hat{t}_{\text{good}} = 2.60\right) > 0.10 + 0.05 = 15\%$, and thus $\hat{B} = 1/0.15 = 6.7$ is a conservative way to satisfy the bound in Proposition 1.

Last, I calculate $Pr\left(|t_1| > \bar{t}|F_1\right)$ by assuming that $t_1|F_1$ (Equation (3)) follows a student's t-distribution with $\hat{\nu} = 100$ degrees of freedom. This can be considered a conservative assumption, as the 10th percentile of the number of in-sample months is 164 (Table 1), implying $\nu > 164$ for 90% of predictors if monthly returns are normally distributed.

3.3. Results: Non-Parameteric FDRs

Table 2 shows the main result: FDRs upper bounds from applying Proposition 1 to the CZ data. The table examines several t-stat hurdles $\bar{t}$, ranging from 2.60 to 3.80. For a hurdle of 2.60, the FDR upper bound is only 10.2%. Since 70.7% of CZ's reproductions meet this hurdle, most claimed findings are highly likely to be true.

[Table 2 about here]

The intuition can be found by breaking down the calculation in steps. First, the naive FDR bound (Equation (15)) just divides the p-value for $\bar{t}$ by the share of observed findings that meet the hurdle $\bar{t}$. This results in a very small FDR bound, as the p-value for a t-stat of 2.60 is only about 1%, and 70% of observed t-stats exceed 2.60. The naive estimate ignores publication bias, which implies that we should multiply the naive bound by a factor of $1/0.15 \approx 6.7$, corresponding to the share of randomly generated strategies that have large t-stats. Still, the resulting FDR bound is only 10.2%, since the p-value is so small to begin with.

Table 2 also shows that the FDR becomes negligible as the hurdle is raised above 3.0. $\bar{t} = 3.4$ implies an FDR upper bound of 3.9%, and $\bar{t} = 3.8$ implies an FDR of at most 0.4%. These results imply that predictors with $|t_i| > 3.8$ are almost certainly true predictors, consistent with Chen (2021)'s thought experiments. This commonality likely steps from the fact that both analyses revolve around comparing the thin tails of the null distribution to the very fat tails of the empirical distribution.

3.4. Reconciliation with Harvey, Liu, and Zhu (2016)

Readers who know HLZ well may feel the numbers in Table 2 are eerily familiar. The table shows that a hurdle of $\bar{t} = 3.4$ implies an FDR upper bound of 1.4%, very close to HLZ's Figure 3, which shows an FDR upper bound of 1% implies a t-hurdle of about 3.4 percent. Indeed, in the text HLZ state an FDR upper bound of 5% implies a t-hurdle of 2.78 based on the same method, also similar to the numbers in Table 2. Strikingly, these numerical results are quite similar, despite the fact that the methods appear to be quite different.

To understand the reconciliation, it helps to know how FDR estimation is equivalent to t-hurdle estimation. This equivalence is known in the statistics lit-

erature going back to Efron and Tibshirani (2002). But since FDR methods are quite foreign to finance researchers, I restate the equivalence here, adapted to nest BY.

I begin by defining the BY algorithm:

1. Sort the observed p-values from smallest to largest: $p_1 \leq p_2 \leq ... \leq p_{N_s}$.

2. Find the cutoff p-value p^* using

$$p^* \equiv \max_j p_j \quad \text{s.t.} \quad p_j \leq \frac{j}{c_{BY} N_s} q \tag{19}$$

$$c_{BY} \equiv \left(\sum_{k=1}^{N_s} \frac{1}{k} \right) \tag{20}$$

   where q is the desired FDR bound.

3. Reject the null hypotheses corresponding to p_j if and only if $p_j \leq p^*$.

In a setting where there is no selection bias, BY prove this procedure implies an FDR $\leq q$.

Stated in this form, the BY algorithm is rather mysterious. It's not at all clear how this procedure leads to FDR control. The logic becomes clear, however, if the BY algorithm is restated in terms of FDR estimation:

Lemma 1. *Define $|\tilde{t}_1| \geq |\tilde{t}_2| \geq ... \geq |\tilde{t}_{N_s}|$ as the observed absolute t-stats ordered from largest to smallest. The BY algorithm is equivalent to choosing a t-hurdle t^* that solves*

$$t^* \equiv \min_{j \in \{1,...,N_s\}} |\tilde{t}_j| \quad s.t. \quad \widehat{FDR}_{BY}\left(\tilde{t}_j\right) \leq q$$

where

$$\widehat{FDR}_{BY}\left(\tilde{t}_j\right) \equiv c_{BY} \widehat{FDR}_{naive}\left(\tilde{t}_j\right) \tag{21}$$

and

$$c_{BY} \equiv \left(\sum_{k=1}^{N_s} \frac{1}{k} \right) \tag{22}$$

is a "correlation penalty."

Proof. The constraint in the optimization (19) equivalent to

$$Pr\left(|t_1| > \tilde{t}_j | F_1\right) \leq \frac{\sum_{i=1}^{N} I\left(|t_i| \geq \tilde{t}_j \cap S_i\right)}{N_s} \frac{1}{c_{BY}} q,$$

since $Pr\left(|t_1| > \tilde{t}_j | F_1\right)$ is the p-value corresponding to $\tilde{t}_j$, and j is equal to the number of observed t-stats that exceed $\tilde{t}_j$ (see step 1 of the algorithm). Solving for q on the RHS and plugging in Equation (15) leads to Equation (19). □

The lemma shows that BY just uses $c_{BY}\widehat{\text{FDR}}_{\text{naive}}\left(\tilde{t}_j\right)$ as an estimator for the upper bound on the FDR and the empirical distribution of t-stats as an estimator of the true distribution. If these are valid estimators (and regularity conditions hold), then the plug-in principle implies that solving the minimization problem finds the most generous t-hurdle that implies the FDR $\leq q$. Under more restrictive dependence assumptions, BY prove that using $c_{BY} = 1$ controls the FDR. Thus $c_{BY} \equiv \left(\sum_{k=1}^{N_s} \frac{1}{k}\right) \gg 1$ can be thought of as a "correlation penalty."

Comparing Equations (16) and (21), we see that my estimates and HLZ's Figure 3 differ only in the choice of the penalty term c_{BY} vs $\hat{c}$. By sheer coincidence, these two penalties are quantitatively similar. HLZ use a dataset of roughly 300 "factors," implying $N_s \approx 300$ and $c_{BY} \equiv \sum_{k=1}^{N_s} \frac{1}{k} \approx 6.3$. In contrast, I use the numbers available in Yan and Zheng (2017)'s Table 1, implying $\hat{c} \equiv \frac{1}{Pr\left(|t_1| > t_{\text{good}}\right)} \leq \frac{1}{0.15} \approx 6.7$. There is no logic for why these two formulas should lead to similar penalties—they just happen to lead to numbers close to 6.5.

Clearly one should include a publication bias penalty. But perhaps one should also include a penalty for correlations? Using both could have profound effects, as multiplying the numbers in Table 2 by a factor of 6 leads to FDR upper bounds as high as 60%, consistent with the idea that most findings are likely false.

Arguing for the correlation penalty, HLZ state that the original Benjamini and Hochberg (1995) algorithm is "only valid when the test statistics are independent or positively dependent." This statement, however, is not true. The correct statement omits "only": independence or positive regression dependence are *sufficient* conditions for the validity of Equation (19), but they are not necessary (see Theorems 1.2 and 1.3 of BY). Indeed, Storey, Taylor and Siegmund (2004) prove that the weak dependence conditions like those in Equations (6)-(8) are sufficient, though they do so in a setting without publication bias. Intuitively, $\widehat{\text{FDR}}_{\text{naive}}\left(\tilde{t}\right)$ is just a method of moments estimator for the upper bound (14). For

fixed $\bar{t}$, all that is needed is a weak law of large numbers, which does not require independence.[6]

The fact that the correlation penalty is often not needed is noted in the original paper (Benjamini and Yekutieli (2001)), which states "[o]bviously, 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 of a procedure." Similarly, Efron (2012) writes that Equation (22) "represents a severe penalty... ...and is not really necessary." Farcomeni (2007) provides a long list of technical conditions under which this penalty is not necessary and Reiner-Benaim (2007) provides simutheoretical results that demonstrate the same.[7]

Deep in their paper, HLZ acknowledge that the correlation penalty may be "overly stringent" and offer an estimate that adjusts for publication bias while omitting the correlation penalty. This estimate finds a hurdle of 3.05 implies an FDR $\leq$ 5%, very similar to my estimates in Table 2 (page 24). HLZ express little confidence in this estimate, however. They list it at the very end of their analyses on Benjamini and Hochberg (1995) style controls, after they list estimates that include only the correlation penalty and estimates that include both penalties. Indeed, these estimates are not found in any exhibits in the paper, and the estimation of the publication bias penalty is relegated to the appendix.

My analysis shows that, counterintuitively, the very last of HLZ's many FDR upper bound estimates are the ones that should be used. Moreover, one does not need to rely on HLZ's appendix for the publication bias adjustment. Proposition 1, along with Table 1 of Yan and Zheng (2017), are all that are needed. I also show that the resulting t-hurdles imply that the FDR for most of the literature is quite small.

Of course, one additional difference with HLZ is the interpretation. Based on numbers very close to those in Table 2, HLZ "argue that most claimed research findings in financial economics are likely false." Instead, I claim that "most statistical findings in cross-sectional asset pricing are likely true."

[6] As seen in Lemma 1, Benjamini and Hochberg (1995) requires a stochastic input to $\widehat{\text{FDR}}_{\text{naive}}(\cdot)$, which then requires extending the law of large numbers and continuous mapping theorem to empirical distribution functions. The details are quite technical (see Van der Vaart 2000), but these extensions are implicitly used whenever a sample median is used to estimate a non-parameterized population median, or when one uses any plug-in estimate without a parametric model.

[7] For additional proofs, see Genovese, Roeder and Wasserman (2006), Ferreira and Zwinderman (2006). In Chris Genovese's lecture slides on "A Tutorial on False Discovery Control," he writes "Practically speaking, BH is quite hard to break even beyond what [h]as been proven."

4. Semi-Parametric FDR Estimates

This section takes a closer look at the FDR estimates by adding a parametric model of the unbiased t-stats. Though I model t-stats parametrically, I estimate FDR bounds using non-parametric methods following Benjamini and Hochberg (1995) and Storey (2002), so I call these estimates "semi-parametric."

These semi-parametric estimates extend the findings in Section 2. Simulations verify the theoretical assumptions and show that the BY correlation adjustment is unnecessarily conservative (Section 4.2). Empirical estimates provide the intuition behind the Yan-Zheng bound, and also show that the FDR that findings in the CZ data are likely to be true overall (Section 4.3).

4.1. A Semi-Parametric FDR Estimator

The semi-parametric estimates build on the statistical framework in Section 2.2. On top of this framework, add the assumption that $|t_1|$ is drawn from a one parameter distribution:

$$|t_1| \sim f(|t_1||\lambda), \tag{23}$$

where $f(\cdot|\lambda)$ is a density function to be chosen, and λ is a parameter to be estimated.

Importantly, Equation (23) regards only the marginal distribution of a single t-stat. It thus makes minimal assumptions about correlations, the structure of true returns, or any of the other issues brought up in more structured models (e.g. Chen and Zimmermann (2020); Chen (2020)). Instead, I rely on the convergence assumptions (Equations (6)-(8)), which I verify in simulations.

To estimate λ, note that the model implies

$$\mathbb{E}\left(|t_1||S_1, |t_1| > t_{\text{good}}; \lambda\right) = \mathbb{E}\left(|t_1||\, |t_1| > t_{\text{good}}; \lambda\right), \tag{24}$$

since the model implies that the selection probability $s(|t_1|)$ is constant for $|t_1| > t_{\text{good}}$.[8] As the LHS is observed, this suggests a simple method of moments

[8] To see this, note that the distribution of $|t_1|$ conditional on S_1 and $|t_1| > t_{\text{good}}$ is

$$\frac{f(|t_1||\lambda)\, s(|t_1|)|_{|t_1|>t_{\text{good}}}}{Pr\left(S_1 \cap |t_1| > t_{\text{good}}|\lambda\right)} = \frac{f(|t_1||\lambda)\, \bar{s}}{\bar{s} Pr\left(|t_1| > t_{\text{good}}|\lambda\right)} = f\left(|t_1||\, |t_1| > t_{\text{good}}, \lambda\right).$$

estimation in which the LHS is replaced with the sample counterpart. Though I use only one moment, I find that estimations using the first two moments or several quantiles lead to similar results. As in Section 3.2, I use $\hat{t}_{\text{good}} = 2.60$ when applying Equation (24).

Given $\hat{\lambda}$, I estimate a bound on the FDR using

$$\widehat{\text{FDR}}_{\text{semi}}\left(\bar{t};\hat{\lambda}\right) \equiv \frac{Pr\left(|t_1| > \bar{t}|F\right)}{Pr\left(|t_1| > \bar{t}|\hat{\lambda}\right)}. \tag{25}$$

where the subscript "semi" stands for semi-parametric. This estimator converges to an upper bound on the FDR:

$$\begin{aligned}\frac{Pr\left(|t_1| > \bar{t}|F_1\right)}{Pr\left(|t_1| > \bar{t}|\hat{\lambda}\right)} &\xrightarrow{p} \frac{Pr\left(|t_1| > \bar{t}|F_1\right)}{Pr\left(|t_1| > \bar{t}|\lambda\right)} \\ &= \frac{1}{Pr\left(F_1\right)}\text{FDR}\left(\bar{t}\right) \\ &\geq \text{FDR}\left(\bar{t}\right)\end{aligned} \tag{26}$$

due to the law of large numbers, the continuous mapping theorem, and Equation (12).

Unlike Proposition 1, Equation (26) is valid for $\bar{t} < t_{\text{good}}$, and thus can estimate a conservative FDR for the entire cross-sectional predictability literature. The price for this expanded inference is a reliance on assumptions, in particular one needs to assume a form for $f(\cdot|\lambda)$.

I examine two choices for $f(\cdot|\lambda)$, an exponential, and a conservative gamma.

4.1.1. The exponential estimator

The first choice assumes $f(\cdot|\lambda)$ is an exponential distribution with scale λ. This distribution fits the right tail of the data well, offers closed form estimation, is conservative, and is the same distribution examined in HLZ's Appendix A.1 (see also Harvey (2017)).

To motivate this assumption, Table 3 shows the mean $|t_i|$ in the CZ data conditional on $|t_i|$ exceeding some cutoffs $\bar{t}$. The table shows that the conditional mean increases by about 1.0 for every 1.0 increase in $\bar{t}$—consistent with memoryless property of an exponential distribution.[9] Thus, the exponential assump-

So multiplying by $|t_1|$ and integrating leads to Equation (24).

[9]The memorylessness of the exponential distribution implies $\mathbb{E}\left(|t_1||t_1| > \bar{t}\right) = \bar{t} + \lambda$ if $|t_1|$ is

tion should fit the data well, and offers a closed form estimator

$$\hat{\lambda} = \hat{\mathbb{E}}\left(|t_1||S_1, |t_1| > \hat{t}_{\text{good}}\right) - \hat{t}_{\text{good}}, \tag{27}$$

where $\hat{\mathbb{E}}\left(|t_1||S_1, |t_1| > \hat{t}_{\text{good}}\right)$ is the sample counterpart to the conditional expectation.

[Table 3 about here]

The exponential distribution also offers a closed form expression for the FDR bound. The bound is simply

$$\widehat{\text{FDR}}_{\text{semi}}\left(\bar{t};\hat{\lambda}\right) = \frac{Pr\left(|t_1| > \bar{t}|F\right)}{\exp\left(-\bar{t}/\hat{\lambda}\right)} \propto \exp\left(-\frac{1}{2}\bar{t}^2\right)\exp\left(\frac{1}{\hat{\lambda}}\bar{t}\right)$$

providing a quantitative intuition behind the FDR bound estimates. The $\exp\left(-\frac{1}{2}\bar{t}^2\right)$ term decays much faster than the $\exp\left(\frac{1}{\hat{\lambda}}\bar{t}\right)$ term, leading to a sharp drop in the FDR bound as $\bar{t}$ increases. How fast this decrease occurs is determined by the scale parameter $\hat{\lambda}$.

Additionally, the exponential distribution is a conservative estimate in the sense that it assumes the modal t-stat is zero. Thus, this assumption implies a kind of worst case for the missing t-stats below 1.96.

4.1.2. A conservative gamma estimator

For robustness, I examine an even more conservative distributional assumption. I assume $f(\cdot|\lambda)$ is a gamma distribution with shape < 1.0 and scale parameter λ offers a natural robustness check. The gamma distribution nests the exponential with shape = 1.0, and a smaller shape parameter implies a larger mode near zero. The larger mode, in turn, implies a smaller value for $Pr\left(|t_1| > \bar{t}|\hat{\lambda}\right)$ and a larger $\widehat{\text{FDR}}_{\text{semi}}\left(\bar{t};\hat{\lambda}\right)$ in Equation (25).

I choose a shape parameter of 0.5. This choice implies a skewness of $2/\sqrt{0.5} = 2.8$, about 1 unit larger than the skewness of the exponential assumption. We will see that this assumption implies a very large spike in t-stats at zero.

exponential with scale parameter λ.

4.2. Simulation Verification

An advantage of the semi-parametric estimates is that, unlike Proposition 1, Equation (25) can be automatically implemented and adapted to any dataset. As a result, I can check the validity of these estimates in simulated data, and check that the convergence assumptions (6)-(8) hold.

To focus on these convergence assumptions, I design the simulation to fully capture the empirical correlation structure, but otherwise make the simulation as simple as possible.

4.2.1. Simulation Model

The simulation assumes there are strategies $i = 1,2,...,N$ and months $t = 1,2,...,T$. Each strategy-month has a return

$$r_{i,t} = \mu_i + \epsilon_{i,t},$$

where μ_i is the population mean return and $\epsilon_{i,t}$ is a mean zero residual.

$\epsilon_{i,t}$ is drawn by "extrapolating" from the empirical de-meaned long-short returns. To construct these simulated residuals, let $\hat{r}_{j,\tau}$ be the return for predictor j in month τ in the balanced panel version of the empirical data (Table 1), with $j = 1,...,N_e$ and $\tau = 1,2,...,T_e$. Define de-meaned empirical returns as $\hat{e}_{j,\tau} = \hat{r}_{j,\tau} - T_e^{-1}\sum_{\tau'=1}^{T_e}\hat{r}_{j,\tau'}$, then the simulated residual is constructed as

$$\epsilon_{i,t} = 0.9\hat{e}_{\tilde{i}(i),\tilde{t}(t)} + 0.1\delta_{i,t} \tag{28}$$

where, $\tilde{i}(i)$ is a random integer between 1 and N_e, $\tilde{t}(t)$ is a random integer between 1 and T_e, and $\delta_{i,t} \sim N(0,5)$ i.i.d. In other words, I cluster-bootstrap residuals from empirical data, where the clustering preserves cross-sectional correlations, but I mix in 10% random noise of volatility similar to the empirical data (see Table 1).

Equation (28) is a simple way to address the problem of extrapolating an observed N_e-dimensional distribution (with correlations) to a $N-$dimensional distribution, where $N \gg N_e$. This approach ensures that the resulting distribution of pairwise correlations for $\epsilon_{i,t}$ is close to the distribution of pairwise correlations for $\hat{e}_{j,\tau}$, while ensuring that $\epsilon_{i,t}$ does not consist of copies of identical strategies. Alternatively, one could extrapolate by parameterizing the distribu-

tion of correlations, but this would significantly complicate the simulation (e.g., Chen (2020)). For robustness, Appendix A.3 examines alternative extrapolations including a pure cluster bootstrap without noise and a block-independent bootstrap. Both of these alternatives lead to similar results.[10]

The population mean return μ_i takes on one of two values

$$\mu_i = \begin{cases} 0 & \text{with prob } p_F \\ \gamma & \text{with prob } \left(1 - p_F\right) \end{cases}$$

where γ is the mean return of true predictors and p_F is the (unconditional) probability a predictor is false. t-stats are calculated the standard way

$$t_i = \frac{\bar{r}_i}{\sigma_i / \sqrt{T}},$$

where the sample mean $\bar{r}_i = T^{-1} \sum_{i=1}^{T} r_i$ and sample volatility $\sigma_i = \sqrt{T^{-1} \sum_{i=1}^{T} \left(r_{i,t} - \bar{r}_i\right)^2}$ are also standard.

Strategies are selected as findings (the event S_i occurs) according to a staircase function

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

where s_{marginal} is the probability a marginal t-stat is selected, t_{good} is the hurdle beyond which t-stats are no longer marginal, and $\bar{s}$ is the maximum probability of selection

Throughout the simulations, I set $N = 10{,}000$, $T = 200$, and $\bar{s} = 1$. I set $\bar{s} = 1$ for computational expedience. A smaller $\bar{s}$ would lead to identical results with a larger N.

4.2.2. Simulation Results

Before examining the simulated estimations, I first check that the simulated correlations are close to the empirical data. Figure 1 makes this comparison. The figure plots the distribution of pairwise correlations for $\epsilon_{i,t}$ implied by the simu-

[10] Code for these alternatives can be found at https://github.com/chenandrewy/mostly-true.

lation (solid) against the empirical distribution (dashed). The two distributions line up very closely. This close match shows that the simulation effectively captures the dependence structure in the data, and provides a relevant test of the crucial convergence assumptions (6)-(8).

[Figure 1 about here]

The simulated estimations are shown in Figure 2. The figure shows estimates of the upper bound on the FDR (Equation (25)) compared to the actual FDR, for various t-stat hurdles $\bar{t}$. Each panel shows the results of a different set of parameter values. I examine four sets of parameters, intended to span the range of realistic values.

[Figure 2 about here]

The top panels both examine a "moderate bias" setting, in which $t_{\text{good}} = 2.60$ and $s_{\text{marginal}} = 0.5$. These parameters imply that predictors that fail to meet the 1% significance threshold are considered marginal, and that marginal predictors are half as likely to be selected as predictors with $|t_i| > 2.60$. I call this bias moderate, as these parameter values are equivalent to the bias assumed in HLZ's "model with correlations."

The top panels differ in the assumed magnitude of actual false discoveries. Panel (a) assumes a "moderate FDR," with $p_F = 0.50$ and $\gamma = 0.50$. The moderate FDR setting is chosen to be close to estimates from HLZ's model with correlations. Panel (b) assumes a "huge FDR," with $p_F = 0.990$ and $\gamma = 0.250$. This second parameter set is chosen to imply the highest FDR possible while maintaining good behavior of the simulation and the intuition that a monthly return of $\gamma < 0.25\%$ is economically insignificant. The choice of 0.25% per month is motivated by the fact that 90 percent of CZ's reproductions produce mean returns that exceed 0.27% per month (Table 1).

Panels (a) and (b) show that the semi-parametric FDR estimates are quite conservative. The exponential (line) and conservative (dashed) FDRs upper bounds are far above the actual FDR (dotted) in both panels. Indeed, in the huge FDR setting, The estimated FDR bounds remain above 50% for $\bar{t}$ up to 5.0, though the actual FDR has plummeted to zero past $\bar{t} = 4.0$. These results suggest that the semi-parametric estimates are safe to use if one is especially concerned about the worst case: a setting in which both p_F is close to 1 and γ is close to 0.

The bottom panels of Figure 2 examine a "extreme bias" setting, in which $t_{\text{good}} = 5.0$ and $s_{\text{marginal}} = 0.25$. This setting implies that predictors with a t-stat above 5.0 are 4× more likely to be published than predictors with a t-stat between 2.0 and 5.0. Once again, I examine two choices for parameters that control the FDR. Panel (c) assumes a moderate FDR (similar to HLZ's estimates) while Panel (d) assumes a huge FDR.

As in the top panels, the bottom panels show that the semi-parametric FDR estimates consistently place an upper bound on the actual FDR. Overall, the FDR bound estimates are somewhat lower than in the top panels, but they are relatively unaffected by the extreme bias.

This robustness is likely because of the fact that the student's t-distribution decays extremely quickly for values larger than 2.0. In the large degrees of freedom limit, this decay is a exponential-quadratic. As a result, even if t-stats of 5.0 are 4-times more likely to be published, the decay in the right tail of the observed distribution is still quantitatively similar to the full distribution.

An interesting result of these simulations is that it is quite difficult to create a realistic simulation that implies an actual FDR > 50% for t-stat hurdles of 3.0 or more. Even if 99% of predictors are false and the true predictors have mean returns of only 25 bps per month, one still finds that a t-stat hurdle of 3.0 is sufficient to imply that most discoveries are true. Appendix A.3 shows that changing the bootstrap to consist of independent blocks can lead to an FDR of 75% for a t-stat hurdle of 3.0, suggesting that some of these limits are due to the negative correlations in the data. These results are reminiscent of Chen (2021)'s thought experiments and suggest an alternative argument for the idea that most findings are true, though a formal demonstration is beyond the scope of this paper.

4.3. Empirical Results

Having verified that the estimator works, I now show the empirical estimates.

Figure 3 begins by showing how the parametric component fits the empirical t-stats. The top panel shows the distribution implied by the exponential estimator. As expected, the exponential assumption (line) fits the right tail of the data (bars) very well. Consistent with HLZ and Harvey (2017), the estimated $\hat{\lambda} = 2.12$ is right in the middle of HLZ's estimates of between 1.93 and 2.22. This estimate implies that $Pr\left(|t_1| > \hat{t}_{\text{good}}\right) = \exp(-2.6/2.12) = 29\%$, consistent with the lower

bound of 15% implied by the Yan-Zheng data (Section 3.2). Indeed, this estimate suggests that the non-parametric estimates are too conservative by a factor of roughly 2.

[Figure 3 about here]

Panel (b) shows the fit of the conservative gamma estimator. The fit implies a very large mass of t-stats near zero. Indeed, the density near zero is so large that it is difficult to see the data (bars) in the same scale. This model implies $Pr\left(|t_1| > \hat{t}_{\text{good}}\right) = 15.6\%$, similar to the Yan-Zheng bound. Thus, Panel (b) offers one interpretation of the Yan-Zheng bound (assuming that it is binding).

In both panels of Figure 3, the models fit the published data well. Despite the strong fit, the panels show very different implications for the distribution of t-stats, consistent with Chen (2020)'s finding that t-stat hurdles are weakly identified in a setting with publication bias. However, unlike in Chen (2020) (and HLZ's model with correlations), the estimators in this paper aim only to place an upper bound on the FDR. I do not attempt to estimate $Pr(F_1)$, nor do I try to find a point estimate of the FDR. Moreover, by bringing in external data in the form of the Yan-Zheng strategies, I place intuitive limits on these upper bounds.

Having shown the model fits, Figure 4 shows the key implications: estimated FDR upper bounds. The figure plots the exponential (line) and conservative (dotted) upper bounds along with the non-parametric estimate (dashed) for comparison. The non-parametric estimate ends at $\bar{t} = 2.60$ (vertical line) as Proposition 1 is only valid to the right of this line. The semi-parametric estimates, then, extend the FDR upper bounds to smaller t-stat hurdles.

[Figure 4 about here]

Both estimators imply that predictors with $|t_i| > 2.0$ have FDRs of at most 25%. That is, predictors that meet the traditional hurdle are at least 75% likely to be true. As a consequence, if one is willing to believe the model fits in Figure 3 (or Harvey (2017)'s Figure 1), then one should believe that not only are most claimed statistical findings likely true, but that most claimed findings are true. Indeed, the exponential estimate, which offers a more natural fit to the data, implies that at least 88% of traditionally-significant predictors are likely to be true.

To emphasize this point, Table 4 shows FDR upper bound estimates for various hurdle $\bar{t}$, along with the share of CZ's reproductions that meet $\bar{t}$. 88% of CZ's

reproductions meet the traditional hurdle of $\bar{t}$ =2.0, and the FDR estimates imply that at least 68% of these predictors are true. Lowering $\bar{t}$ to 1.60 includes 95% of CZ's reproductions, but even for this low bar the estimators still imply that most predictors are true, with the exponential estimate implying that at least 75% of these predictors are true. Even if the remaining 5% of predictors are classified as false discoveries (these are arguably failed reproductions), these numbers imply that the majority of claimed findings are true.

[Table 4 about here]

5. Conclusion

I revisit the conflicting findings in HLZ. Though HLZ argue that most claimed research findings are false, their numerical estimates imply that most statistical findings are true.

I verify HLZ's primary numerical findings, though this verification is due to offsetting effects. HLZ's use of the Benjamini and Yekutieli (2001) correlation penalty is unnecessarily conservative, but this penalty happens to be similar in size to a publication bias penalty that is not used in HLZ's main figure. Excluding the correlation penalty and including the publication bias penalty results in an FDR of at most 10% for the 70% of predictors that meet the t-stat hurdle of 2.6. I verify in simulations that these estimates are valid for the type of dependence we find in predictor data.

I also provide simple, closed-form expressions for FDR bounds that are valid under publication bias. These expressions provide the intuition behind more complicated estimations that find similar results (e.g. Chen (2020)). In short, the FDR in the literature is small because even trading on random accounting-based strategies leads to large t-stats with some regularity.

An important caveat is that my study does not examine trading costs, nor does it examine whether the statistical tests satisfactorily address the research claims in the original papers. Thus, HLZ's argument that most claimed research findings are false may still be correct. However, verifying this claim seems to be outside of the realm of multiple testing statistics.

A. Appendix

A.1. A "Microfoundation" for the two-step selection model

[TBC]

Probability of selection depends simultaneously on both t-stat t and a signal θ in the following form

$$
\begin{aligned}
P(S|t,\theta) &= g(t)\,h(\theta) \\
t &= \mu + \epsilon \\
\theta &= \mu + \delta \\
\epsilon &\sim f_{\text{null}}
\end{aligned}
$$

This form means we can interpret $g(t)$ and $h(\theta)$ as probabilities. Define events S_t and S_θ such that

$$
\begin{aligned}
P(S_\theta|\theta) &= h(\theta) \\
P(S_t|t,S_\theta) &= s(t)
\end{aligned}
$$

then I can rewrite $P(S|t,\theta)$ as

$$P(S|t,\theta) = P(S_t)\,P(S_\theta),$$

Assuming that θ and ϵ are independent, then

$$P(\epsilon|S_\theta) \sim f_{\text{null}}$$

and as a result, $t|S_\theta$ is unbiased in the sense that

$$t|S_\theta, \mu = 0 \sim f_{\text{null}}.$$

Thus, we can think of the first step in Section 2.2 as the t-stats conditional on the event S_θ. Then if both S_θ and S_t occur, the predictor is selected, and we have in

general

$$t|S_\theta, S_t, \mu = 0 \sim f_{\text{null}}$$

A.2. Proofs

Proof of Proposition 1

Proof. For $\bar{t} > t_{\text{good}}$, $Pr\left(|t_1| > \bar{t}|S_1\right) = \bar{s}Pr\left(|t_1| > \bar{t}\right)/Pr\left(S_1\right)$. Thus,

$$\begin{aligned}\frac{Pr\left(|t_1| > \bar{t}|S_1\right)}{Pr\left(|t_1| > \bar{t}\right)} &= \frac{\bar{s}}{Pr\left(S_1\right)} \\ &= \frac{\bar{s}}{Pr\left(S_1 \cap |t_1| < t_{\text{good}}\right) + \bar{s}Pr\left(|t_1| > t_{\text{good}}\right)} \\ &\leq \frac{1}{Pr\left(|t_1| > t_{\text{good}}\right)}.\end{aligned}$$

Thus,

$$\begin{aligned}\operatorname*{plim}_{N\longrightarrow\infty}\left[\hat{c}\widehat{\text{FDR}}_{\text{naive}}\left(\bar{t}\right)\right] &= \frac{1}{Pr\left(|t_1| > t_{\text{good}}\right)}\frac{Pr\left(|t_1| > \bar{t}|F_1\right)}{Pr\left(|t_1| > \bar{t}\right)}\left[\frac{Pr\left(|t_1| > \bar{t}|S_1\right)}{Pr\left(|t_1| > \bar{t}\right)}\right]^{-1} \\ &\geq \frac{Pr\left(|t_1| > \bar{t}|F_1\right)}{Pr\left(|t_1| > \bar{t}\right)} \geq \text{FDR}\left(\bar{t}\right).\end{aligned}$$

□

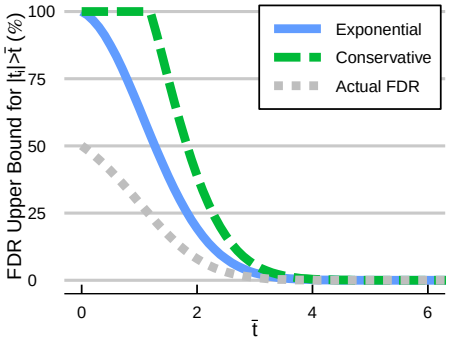
A.3. Alternative Simulation Results

[TBC]

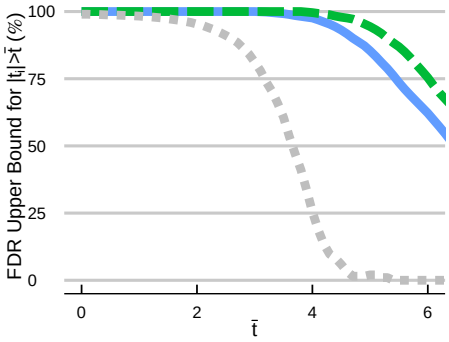
Figure A.1: Estimations on an Alternative Simulation

I follow the caption in Figure 2, but draw residuals $\epsilon_{i,t}$ in independent blocks from the empirical residuals $\hat{e}_{j,\tau}$, instead of following Equation (28). **Interpretation:** The FDR bounds are robust to this alternative correlation structure. Correlations closer to zero seem to allow for larger actual FDRs compared to the empirically-motivated correlations in Figure 2.

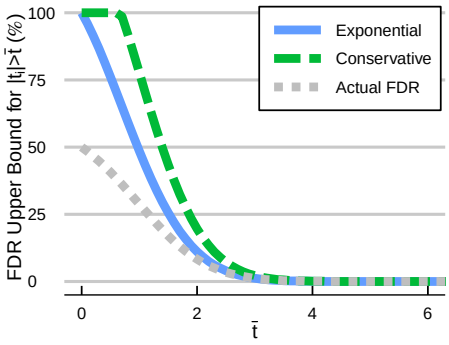
(a) Moderate Bias, Moderate FDR



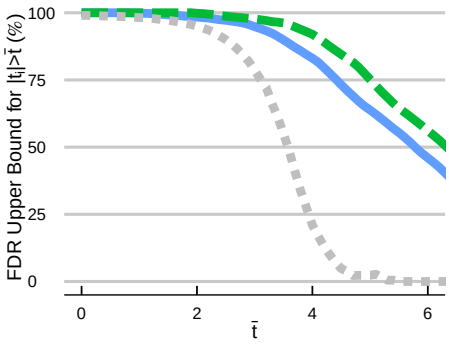
(b) Moderate Bias, Extreme FDR



(c) Extreme Bias, Moderate FDR



(d) Extreme Bias, Extreme FDR



References

Andrews, I., Kasy, M., 2019. Identification of and correction for publication bias. American Economic Review 109, 2766–94.

Benjamini, Y., 2008. Comment: Microarrays, empirical bayes and the two-groups model. Statistical Science 23.

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

Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. Annals of statistics , 1165–1188.

Chen, A.Y., 2020. Do t-stat hurdles need to be raised? Available at SSRN: https://papers.ssrn.com/abstract=3254995 .

Chen, A.Y., 2021. The limits of p-hacking: Some thought experiments. The Journal of Finance .

Chen, A.Y., Velikov, M., 2021. Zeroing in on the expected returns of anomalies. Working Paper .

Chen, A.Y., Zimmermann, T., 2020. Publication bias and the cross-section of stock returns. The Review of Asset Pricing Studies 10, 249–289.

Chen, A.Y., Zimmermann, T., Forthcoming. Open source cross sectional asset pricing. Critical Finance Review .

Chordia, T., Goyal, A., Saretto, A., 2020. Anomalies and false rejections. The Review of Financial Studies 33, 2134–2179.

Efron, B., 2012. Large-scale inference: empirical Bayes methods for estimation, testing, and prediction. volume 1. Cambridge University Press.

Efron, B., Tibshirani, R., 2002. Empirical bayes methods and false discovery rates for microarrays. Genetic epidemiology 23, 70–86.

Efron, B., Tibshirani, R., Storey, J.D., Tusher, V., 2001. Empirical bayes analysis of a microarray experiment. Journal of the American statistical association 96, 1151–1160.

Fama, E.F., MacBeth, J.D., 1973. Risk, return, and equilibrium: Empirical tests. The Journal of Political Economy , 607–636.

Farcomeni, A., 2007. Some results on the control of the false discovery rate under dependence. Scandinavian Journal of Statistics 34, 275–297.

Ferreira, J., Zwinderman, A., 2006. On the benjamini–hochberg method. The Annals of Statistics 34, 1827–1849.

Genovese, C.R., Roeder, K., Wasserman, L., 2006. False discovery control with p-value weighting. Biometrika 93, 509–524.

Giglio, S., Liao, Y., Xiu, D., 2021. Thousands of alpha tests. The Review of Financial Studies 34, 3456–3496.

Harvey, C.R., 2017. Presidential address: The scientific outlook in financial economics. The Journal of Finance 72, 1399–1440.

Harvey, C.R., Liu, Y., 2021. Uncovering the iceberg from its tip: A model of publication bias and p-hacking. Available at SSRN 3865813 .

Harvey, C.R., Liu, Y., Zhu, H., 2016. ... and the cross-section of expected returns. The Review of Financial Studies 29, 5–68.

Jensen, T.I., Kelly, B.T., Pedersen, L.H., 2021. Is There A Replication Crisis In Finance? Technical Report. National Bureau of Economic Research.

Kozak, S., Nagel, S., Santosh, S., 2018. Interpreting factor models. The Journal of Finance 73, 1183–1223.

Lewellen, J., Nagel, S., Shanken, J., 2010. A skeptical appraisal of asset pricing tests. Journal of Financial Economics 96, 175–194.

McLean, R.D., Pontiff, J., 2016. Does academic research destroy stock return predictability? The Journal of Finance 71, 5–32.

Reiner-Benaim, A., 2007. Fdr control by the bh procedure for two-sided correlated tests with implications to gene expression data analysis. Biometrical Journal 49, 107–126.

Storey, J.D., 2002. A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64, 479–498.

Storey, J.D., Taylor, J.E., Siegmund, D., 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, 187–205.

Tucker, H.G., 1959. A generalization of the glivenko-cantelli theorem. The Annals of Mathematical Statistics 30, 828–830.

Van der Vaart, A.W., 2000. Asymptotic statistics. volume 3. Cambridge university press.

Yan, X.S., Zheng, L., 2017. Fundamental analysis and the cross-section of stock returns: A data-mining approach. The Review of Financial Studies 30, 1382–1423.

Zhu, M., 2021. Appropriate t-stat hurdles in large-scale testing. Available at SSRN 3866076 .

Figure 1: Simulated Correlations

I simulate monthly long-short return residuals by mixing a cluster-bootstrap of de-meaned returns from the CZ data with noise ($\epsilon_{i,t}$ in Equation (28)) with $N = 10{,}000$ and $T = 200$. The residuals use balanced panel data for simplicity (see Table 1). I then plot the distribution of pairwise correlations (solid) and compare with the distribution from the CZ data (dashed). **Interpretation:** The cluster bootstrap simulation mimics empirical correlations well.

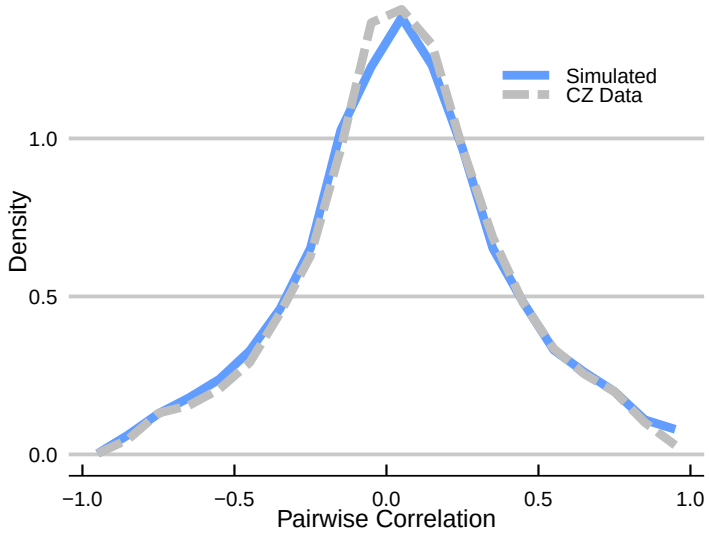
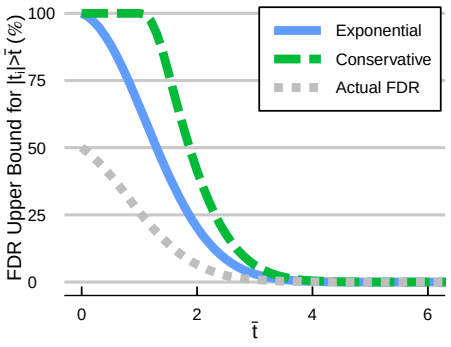


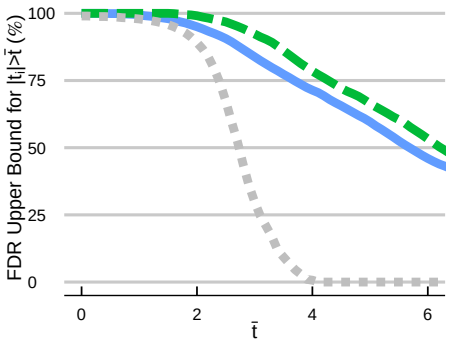
Figure 2: Estimations on Simulated Data

I simulate models of biased predictability findings (Section 4.2.1) and apply the semi-parametric FDR estimators (Section 4.1). All estimates use $\hat{t}_{\text{good}} = 2.6$. All models use $N = 10,000$, $T = 200$. "Moderate Bias" is $t_{\text{good}} = 2.6$, $s_{\text{marg}} = 0.50$; "Extreme Bias" is $t_{\text{good}} = 5.0$, $s_{\text{marg}} = 0.25$; "Moderate FDR" is $p_F = 0.5$, $\gamma = 0.5$; and "Extreme FDR" is $p_F = 0.99$, $\gamma = 0.25$. Moderate parameter values are close to HLZ's values, and the other parameters are intended to span the range of reasonable values. Panels show the average FDR across 200 simulations. **Interpretation:** The semi-parametric estimators consistently place an upper bound on the actual FDR, even in simulations with huge FDRs or with an extreme amount of publication bias.

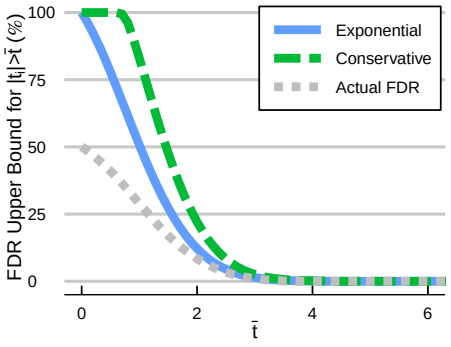
(a) Moderate Bias, Moderate FDR



(b) Moderate Bias, Extreme FDR



(c) Extreme Bias, Moderate FDR



(d) Extreme Bias, Extreme FDR

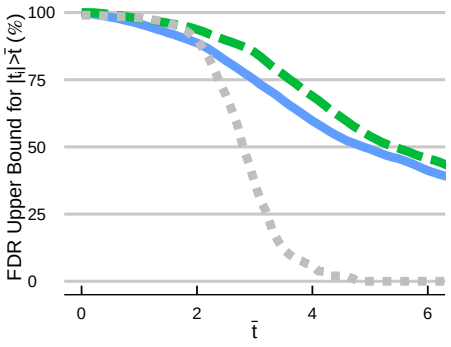
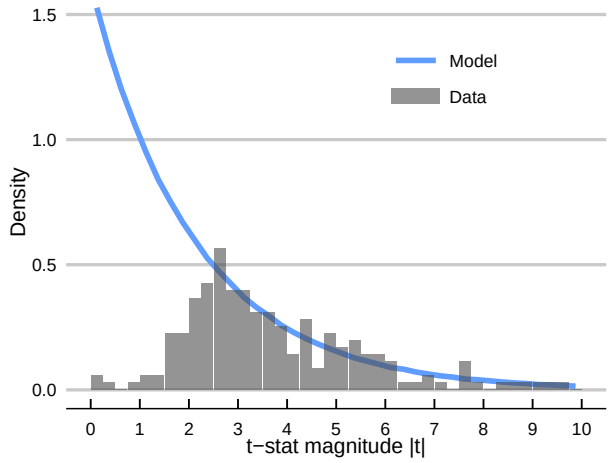


Figure 3: Semi-Parametric Model Fits

I apply semi-parametric FDR upper bound estimators (Equation (25)) to predictors from Chen and Zimmermann (Forthcoming). Panel (a) assumes unbiased t-stats follow an exponential distribution, as in HLZ. Panel (b) assumes a gamma distribution with shape parameter 1/2. Figures compare the distribution of unbiased t-stats (lines) to the published data (bars). All densities are normalized so that $Pr(|t_1| > 2.6) = 1$ for ease of comparison. **Interpretation:** The exponential estimate implies about 30% of $|t_i| > 2.6$, consistent with the Yan-Zheng lower bound of 15%. The conservative estimate implies a huge spike in $|t_i|$ near zero and 16% of $|t_i| > 2.6$. Both estimates fit the data well for $|t_1| > 2.6$.

(a) Exponential



(b) Conservative Gamma

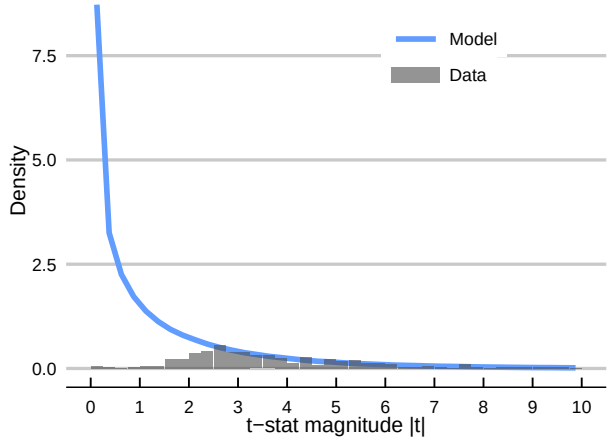


Figure 4: Semi-Parameteric FDR estimates

I apply semi-parametric FDR upper bound estimators (Section 4.1) to predictors from Chen and Zimmermann (Forthcoming). "Exponential" assumes unbiased t-stats follow an exponential distribution, as in HLZ. "Conservative" assumes a gamma distribution with shape parameter 1/2. The non-parametric estimates are shown for comparison, though these are missing below $\bar{t} = 2.60$ because these estimates are not valid in this range (Proposition 1). **Interpretation:** Most predictors with $|t_i| > 2.0$ are true. The figure provides a visualization of the Table 4, for ease of comparison with the simulations (Figure 2).

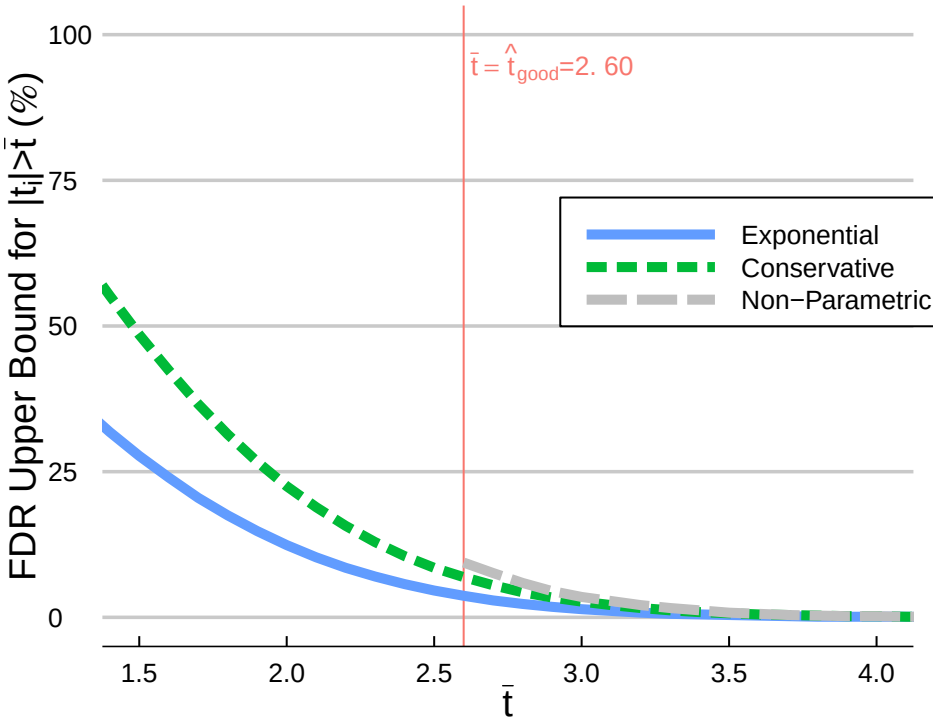


Table 1: Summary statistics for published cross-sectional predictors

Data comes from the Chen-Zimmermann (2020) open-source reproductions of 205 cross-sectional predictors with code and data available at www.openassetpricing.com. I use the "original paper" data, which forms long-short portfolios based on the instructions in the original papers. $|t_i|$ = |[mean return]|/[volatility]$\times\sqrt{[\text{number of months}]}$. These 205 predictors are selected by Chen and Zimmermann to be likely to produce t-stats > 1.96 in long-short portfolios based on results in the original papers, and are signed to have positive mean returns based on the original results. All statistics use the original sample periods, except for "Balanced Panel," which begins with all time periods, then limits the months to those with more than 150 predictors, and also requires predictors which have at least 200 months of data before imposing complete cases at the month-predictor level. **Interpretation:** The t-statistics shown here will be used to infer FDRs for the cross-sectional literature. The majority of these t-stats exceed 2.60. Correlations cluster around zero, in contrast to the common belief that there is a strong factor structure.

	Percentile								
	10	20	30	40	50	60	70	80	90
Univariate Statistics									
$\|t_i\|$	1.92	2.31	2.61	2.95	3.29	3.78	4.38	5.27	6.39
Mean Return (%)	0.27	0.33	0.40	0.50	0.56	0.66	0.79	1.00	1.30
Volatility (%)	1.50	1.89	2.31	2.57	2.98	3.38	3.91	4.43	5.68
Num of Months	164	209	252	288	336	384	454	468	528
Pairwise Correlations									
Overlapping In-Sample	−0.25	−0.13	−0.06	−0.01	0.04	0.08	0.14	0.22	0.36
Balanced Panel	−0.55	−0.34	−0.17	−0.04	0.09	0.21	0.34	0.50	0.68

Table 2: Non-Parametric FDR Upper Bound Estimates

I estimate publication bias adjusted FDR upper bounds (Proposition 1) for the Chen-Zimmermann dataset of published predictors. The FDR bounds are defined as

$$\text{Naive FDR Bound} \equiv \frac{p\text{-value for } \bar{t}}{\text{Share of Findings w/ } |t_i| > \bar{t}}$$
$$\text{Bias-Adjusted FDR Bound} \equiv \frac{1}{0.15}\,[\text{Naive FDR Bound}]\,,$$

where p-values are computed from 2-sided t-tests with $\nu = 100$, and 0.15 is a lower bound on $Pr\,(|t_1| > 2.6)$ that comes from assuming the unbiased research process is more effective at drawing large t-stats than Yan and Zheng's (2017) algorithm for generating strategies from random Compustat variables. **Interpretation:** For 70% of published predictors, the FDR is at most 10%. Thus, most claimed statistical findings in cross-sectional predictability are likely true. For t-stats above 4.0, the FDR is negligible, consistent with Chen (2021)'s thought experiments.

	t-hurdle $\bar{t}$			
	2.60	3.00	3.40	3.80
p-value for $\bar{t}$ (%)	1.1	0.3	0.1	0.0
Share of Findings w/ $\lvert t_i \rvert > \bar{t}$ (%)	70.7	58.0	46.8	39.0
Naive FDR Bound (%)	1.5	0.6	0.2	0.1
Bias-Adjusted FDR Bound (%)	10.2	3.9	1.4	0.4

Table 3: Conditional mean t-stat and estimation of exponential $|t_1|$

Table shows the mean $|t_i|$ conditioning on predictors having $|t_i|$ that exceed various t-stat hurdles $\bar{t}$. Data is the CZ predictors. Implied exponential $\hat{\lambda} = [\text{Mean } |t_i| \text{ for } |t_i| > \bar{t}] - \bar{t}$ (implied by Equation (27)). **Interpretation:** Mean $|t_i|$ for $|t_i| > \bar{t}$ increases by roughly 1.0 with every 1.0 increase in $\bar{t}$, consistent with the memoryless property of the exponential distribution. Thus, the exponential model provides a good fit to the data as well as a transparent estimation. The scale parameter $\hat{\lambda}$ is estimated to be about 2.1 if one chooses a hurdle between 2.0 and 5.0.

	t-stat cutoff $\bar{t}$				
	2.0	3.0	4.0	5.0	6.0
Mean $\|t_i\|$ for $\|t_i\| > \hat{t}_{\text{good}}$	4.2	5.1	6.2	7.1	8.6
Implied exponential $\hat{\lambda}$	2.2	2.1	2.2	2.1	2.6

Table 4: Semi-Parametric FDR Upper Bound Estimates

I apply semi-parametric FDR upper bound estimators (Equation (25)) to predictors from Chen and Zimmermann (Forthcoming). "Exponential" assumes unbiased t-stats follow an exponential distribution, as in HLZ. "Conservative" assumes a gamma distribution with shape parameter 1/2. The non-parametric estimates are shown for comparison for $\bar{t} \geq 2.60$ (the range in which Proposition 1 applies). **Interpretation**: The exponential estimate implies that at least 70% of claimed findings are true. Even the conservative estimate implies that most claimed statistical findings are true.

$\bar{t}$	Share of $\lvert t_i \rvert > \bar{t}$	FDR Upper Bound for $\lvert t_i \rvert > \bar{t}$		
		Semi-Parametric		Non-Parametric
		Exponential	Conservative	(for $\bar{t} > 2.60$)
1.60	94.6	24.0	42.4	
1.80	91.7	17.5	31.4	
2.00	88.3	12.4	22.5	
2.20	83.4	8.5	15.7	
2.40	77.1	5.7	10.5	
2.60	70.7	3.7	6.9	9.4
2.80	63.9	2.3	4.3	5.9
3.00	58.0	1.4	2.7	3.5
3.20	51.7	0.8	1.6	2.1
3.40	46.8	0.5	0.9	1.2
3.60	43.4	0.3	0.5	0.6
3.80	39.0	0.1	0.3	0.3
4.00	36.1	0.1	0.2	0.2