Publication Bias in Asset Pricing Research

Written by Andrew Y. Chen and Tom Zimmerman
Presented by Reese Alexander

What’s The Finance?

- We all want to be published, especially those of us without tenure.
- As a field we have standards for what’s significant, mostly t-stats greater than a certain value
- This leads to us only publishing “notable” findings
- Very few results are publishable with a t-stat close to zero.

How should we study publication bias?

- A meta-study of meta-studies
- The equation we need to study is:

$$\text{Reported effect} = \text{True effect} + \text{Author error} + \text{Sampling error}$$

- Since publication bias selects for larger results, we can assume that

$$E(\text{Reported effect} \mid \text{Published}) > E(\text{True effect} \mid \text{Published})$$

- Thus we need to estimate Author error and sampling error.

Author versus Sampling error

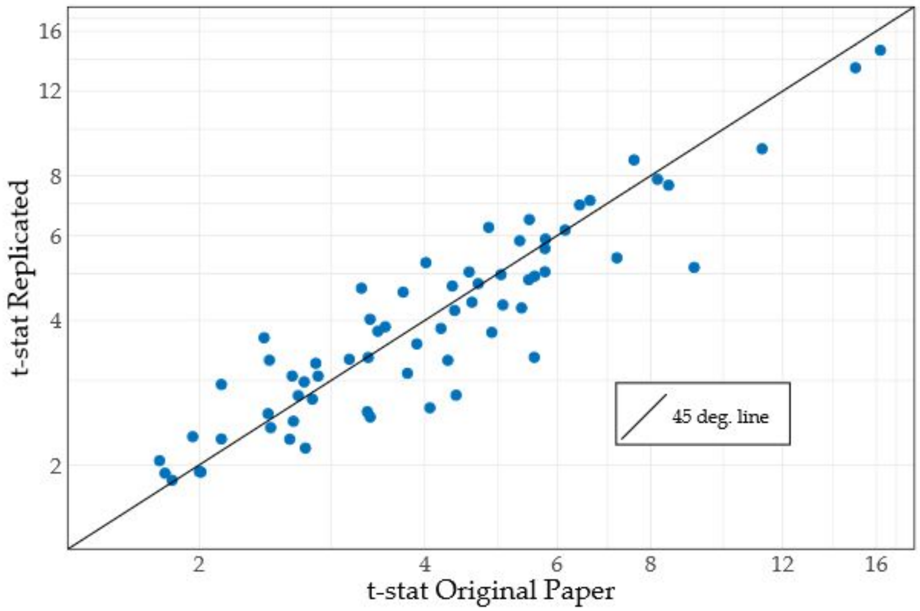
- Author error is effectively negligence (or p-hacking/data mining)
- “Shoot a gun and call what you hit the target” approach
- Would potentially be irreplicable due to author choices, and likely not hold out of sample
- Sampling error is just that, noisiness of sampling leading to false discoveries
- High returns, and high t-stats would indicate a lack of sampling error

Four key facts of the literature on prediction

1. Almost all predictability results are replicable
2. Predictability persists out of sample
3. Empirical t-stats are higher than the standard 2.0
4. Predictors have minimal correlation

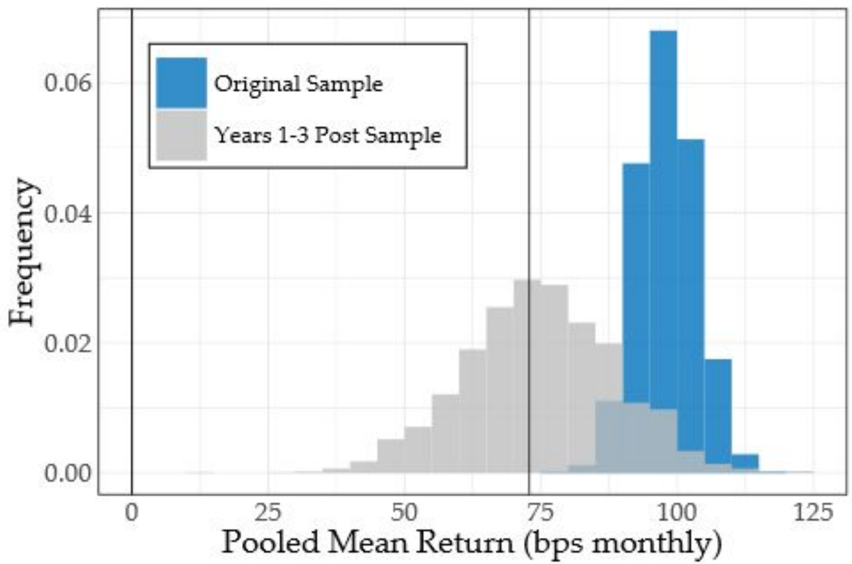
1. Almost all predictability results are replicable

- The authors replicate t-stats from 153 different papers
- This figure implies that author error is minimal
- Contrasting to prior meta studies which deemphasized microcap stocks and failed to replicate



2. Predictability persists out of sample

- 74% of in sample returns persist 1-3 years after.
- A drop in returns is expected by investors trading on research
- 26% decrease due to sampling error is likely an upper bound



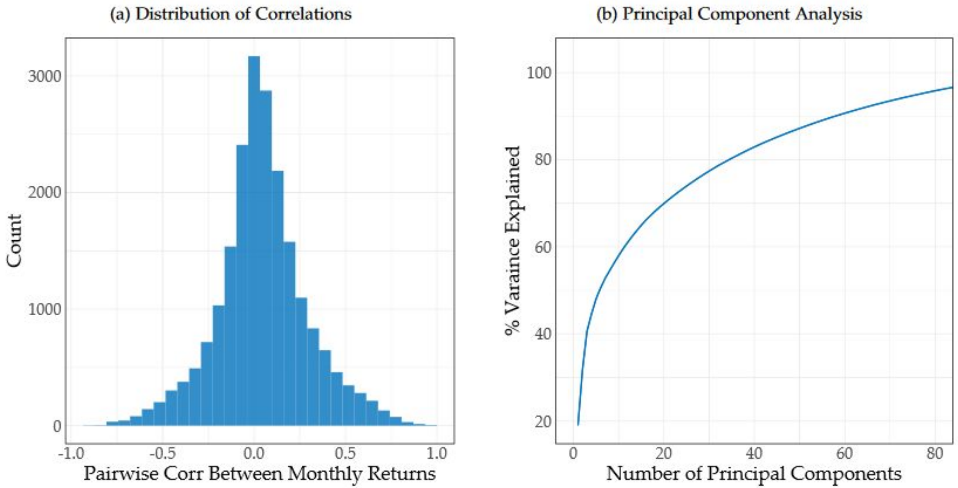
3. Empirical t-stats are higher than the standard 2.0

	t-stat minimum						
	2.0	3.0	4.0	5.0	6.0	7.0	8.0
	(a) Number of Predictors that Meet Minimum						
Published and Replicated	183	121	74	48	26	18	12
Systematically Data-Mined	5,464	2,837	1,522	832	374	185	76
	(b) Percent of Signals that Meet Minimum						
Published and Replicated	88.4058	58.4541	35.7488	23.1884	12.5604	8.6957	5.7971
Systematically Data-Mined	30.1662	15.6628	8.4028	4.5934	2.0648	1.0214	0.4196

Data-mined estimates are from Yan and Zheng (2017)

4. Predictors have minimal correlation

- Predictors are generally between -.5 and .5 which is important in the context of multiple testing.
- It takes 60 principal components to explain 90% of the variance.
- This is expected due to referee’s beliefs in a factor structure for stock returns.



A model for publication bias

- Authors generate ideas which have a t-stat, with two components.

$$t_i = \theta_i + Z_i$$
$$Z_i \sim Normal(0,1)$$

- The primary variable is theta, a randomly distributed true return, and Z represents scaled sampling error.
- Publication is a distribution based on t_i, weakly increasing in theta

A model for publication bias

- Due to publication bias, observed ideas are given by:

$$\Pr(pub_i | t_i, \theta_i) = p(t_i, \theta_i | \sigma_{pub})$$

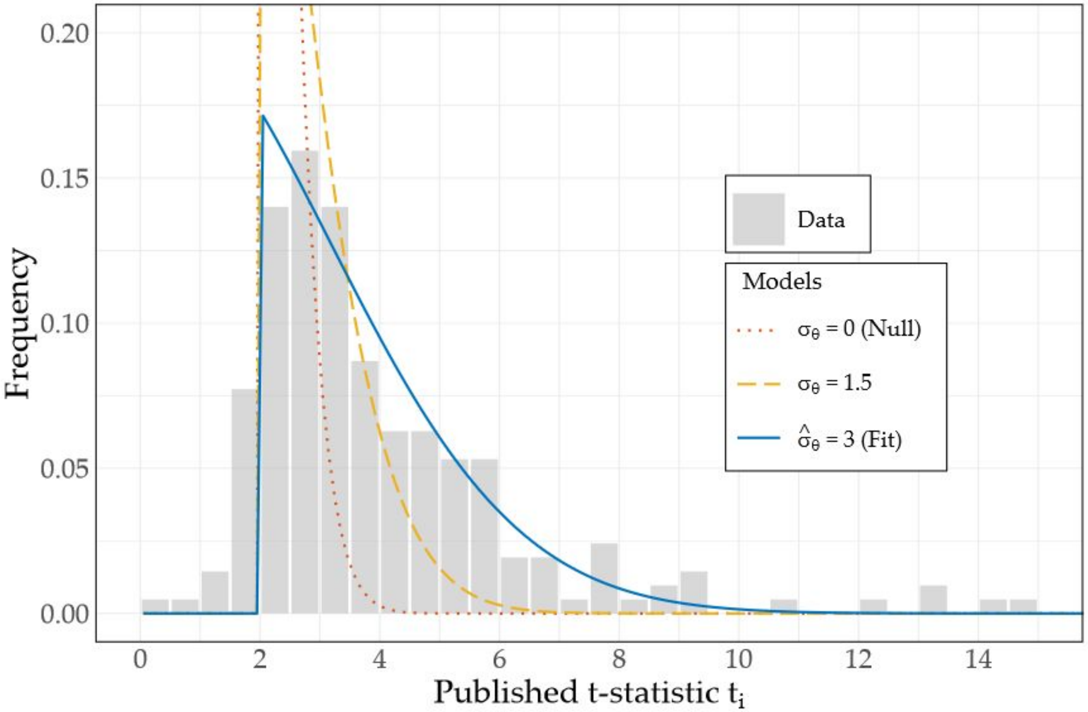
$$\Pr(\mathrm{pub}_i | t_i, \theta_i) = \begin{cases} \bar{p}, & t_i > 2 \\ 0, & \text{otherwise.} \end{cases}$$

- If we assume Z is positive for published research, this implies the discovered returns of most published research are too high, and that we should shrink them towards zero to approximate the truth.

Fitting the model

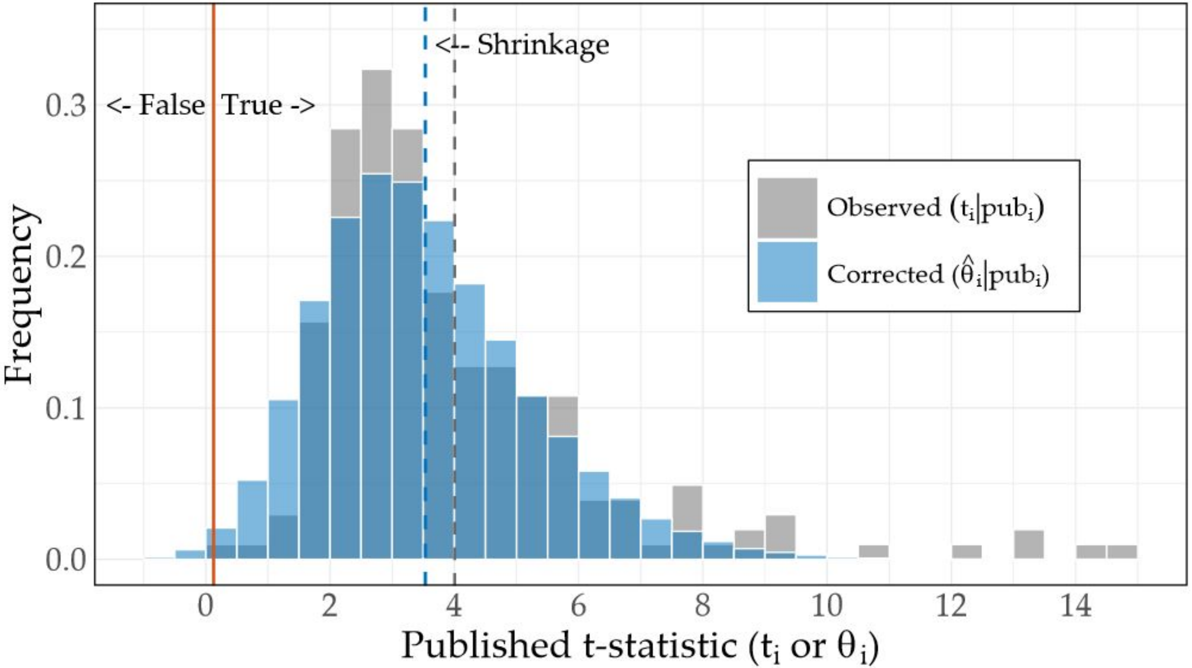
We only need to estimate the variance of theta to match the distribution of published t-statistics.

Estimates suggest that expected returns three standard errors from zero are common. (60 basis points on average)



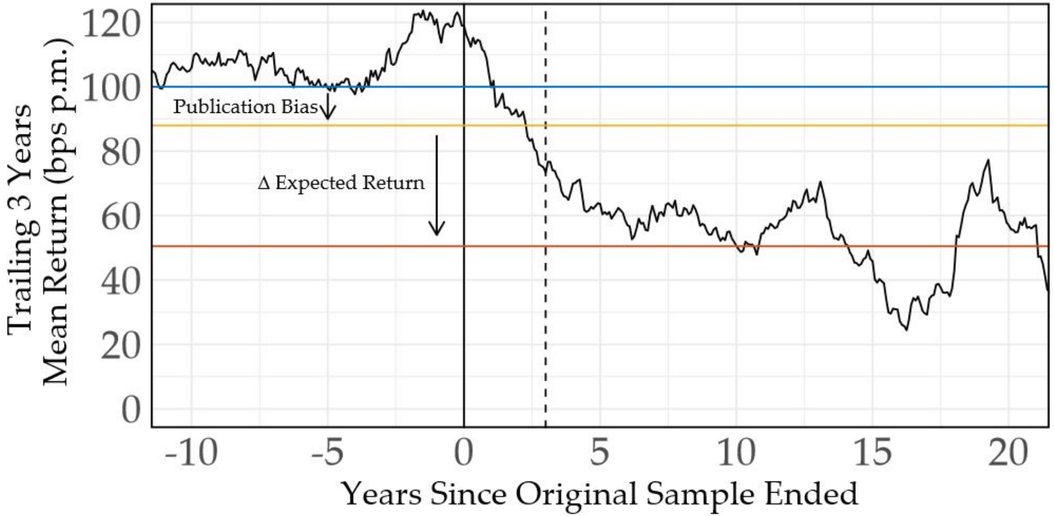
So what should shrinkage be?

By simulating this theta parameter, they show that the average t statistic should shrink by 10%, and that the false discovery rate is extremely negligible at .4%.



Post publication drop-off of predictors

- Predictability drops by 50% after publication in the full sample.
- For the three years following publication, they only drop 12%
- The remaining 38% is likely due to a decline in expected return.



A final critique the meta studies

- Loose language causes different interpretations
- “False findings” of Harvey et al. (2016) could just as easily be insignificant predictors
- Failed replications are not failures if the original paper never claimed to meet the bar for significance.
- Hou et al. (2020) replicate a variety of studies, and assess if they meet $|t|>1.96$, when the original papers didn’t claim that high of a t-stat.

Taking ideas to other fields: Equity premium predictability

- How replicable is premium predictability? Goyal et al. (2021) claim that a majority of predictors fail to replicate (or even have the same sign)
- This implies author or sampling error plays a larger role in that context
- Very few of the equity premium p-values are much smaller than .05
- They are mostly uncorrelated which implies that these p-values are fairly accurate (multiple testing is not so necessary)
- All together these paint a picture that publication bias may play a larger role in this field.

Key Takeaways

- Publication bias is not as strong in asset pricing research
- Most published results hold, however they may become weaker due to liquidity or decline in expected return
- Weak correlations suggest that predictors are explaining unique variation.
- Asset pricing researchers are likely not datamining to find a t-stat just above 1.96 (or referees want higher than that)

Replication of figure 3a

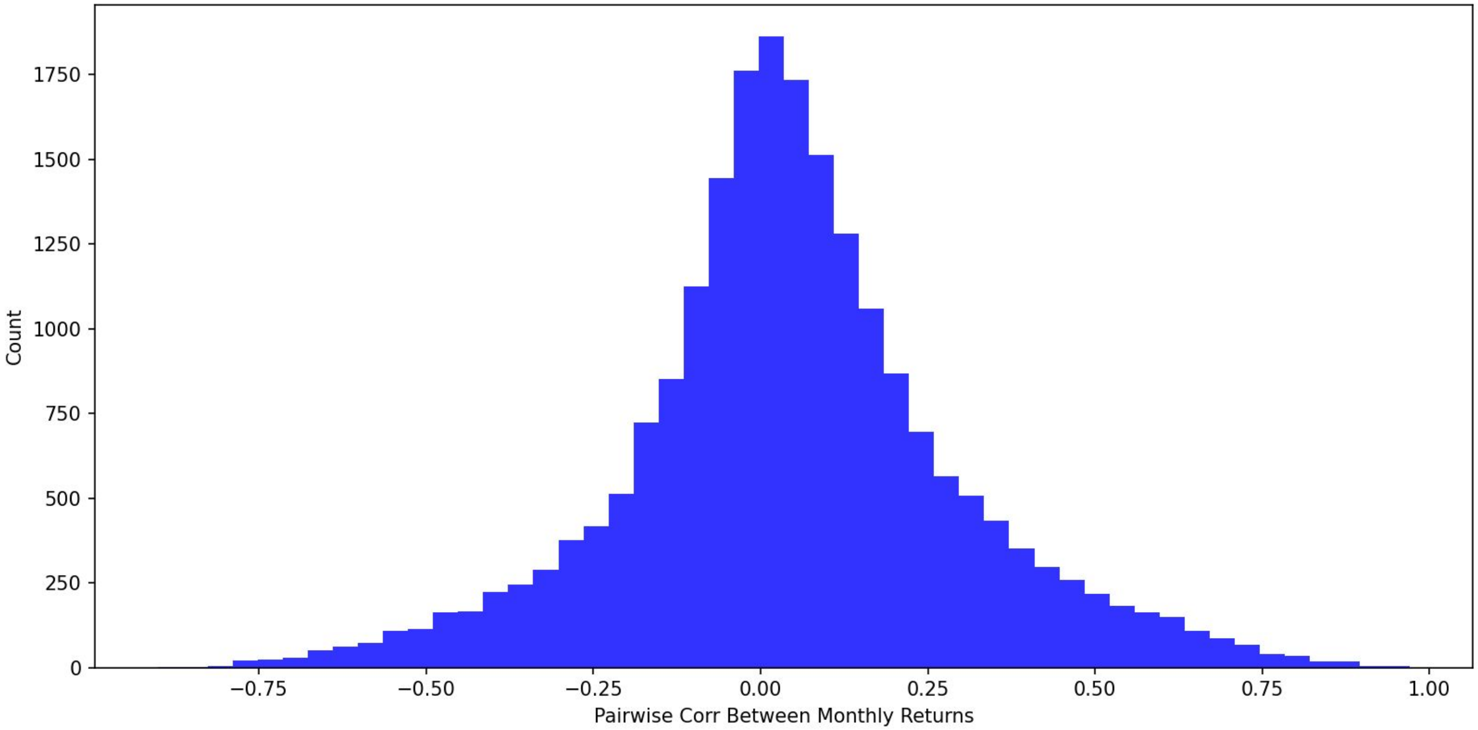
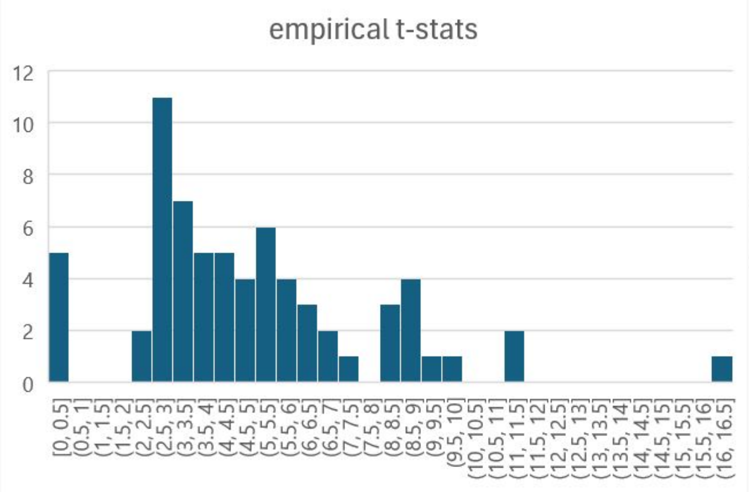
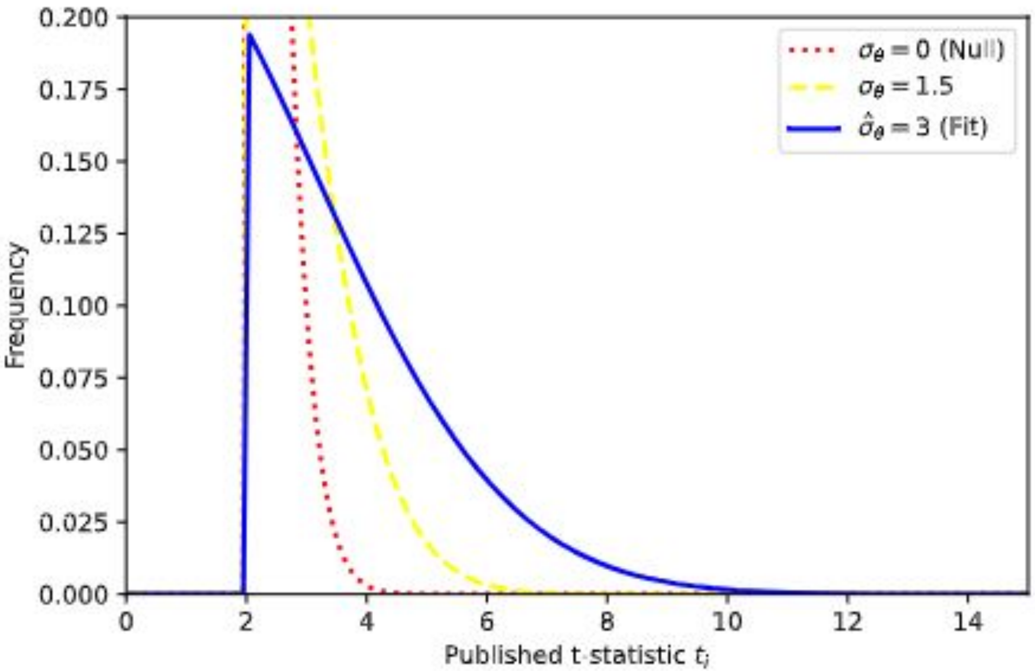


Figure 4 (almost)





Appendix

Figure 6: Distribution of Corrected t-stats (θ_i) from the Literature. Harvey, Liu, and Zhu (2016) uses their baseline SMM estimate (their Table 5, Panel A, $\rho = 0.2$). Chen and Zimmermann (2020) uses their baseline (Table 3, "All"). Jensen, Kelly, and Pedersen (2022) uses their baseline publication bias adjustment (Figure 9, $\tau_c = 0.29\%$). "Simple Normal" uses Section 3.2 (based on Chen and Zimmermann (2020)'s appendix). The literature differs in the modeling of the null ($\theta_i \leq 0$) but for $\theta_i > 0$ the distributions are similar to the simple normal model (Figure 4). All find that expected returns three standard errors from zero are common.

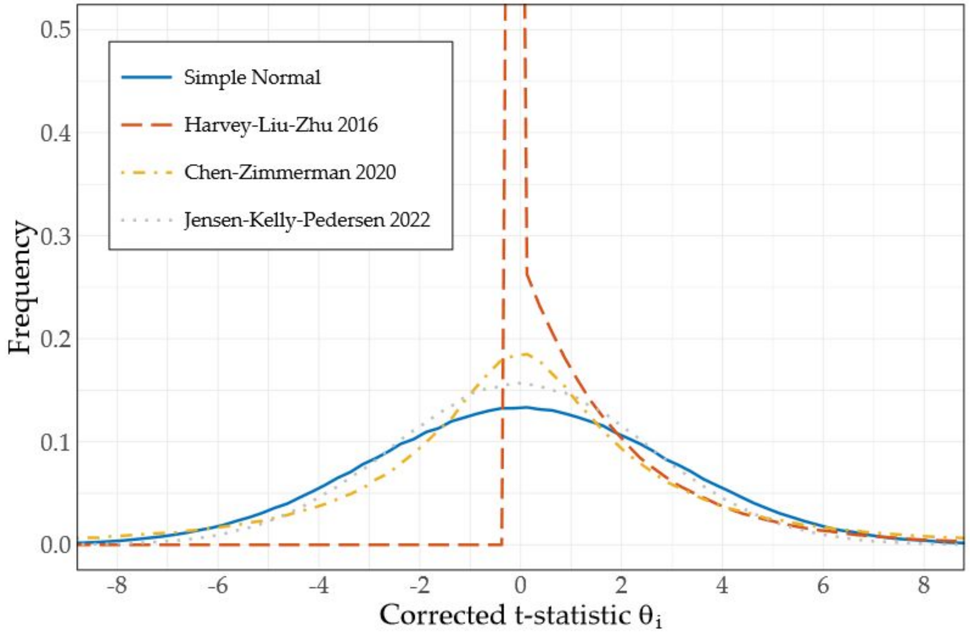


Figure 7: Shrinkage and FDR Estimates in Harvey et al. (2016). Plot compares the distribution of published t-stats (grey) to the distribution of standardized expected returns (blue) implied by Harvey et al. (2016)'s baseline estimate (Table 5, Panel A, $\rho = 0.2$). Dashed lines show the means of each distribution. The distance between these lines implies $\text{Shrinkage}_{\text{pub}} = 13\%$. $\text{FDR}_{\text{pub}} = 6.3\%$ is the mass of expected returns to the left of the solid line. These corrections are similar to the simple model (Figure 5) because the right tail of θ_i is similar (Figure 6).

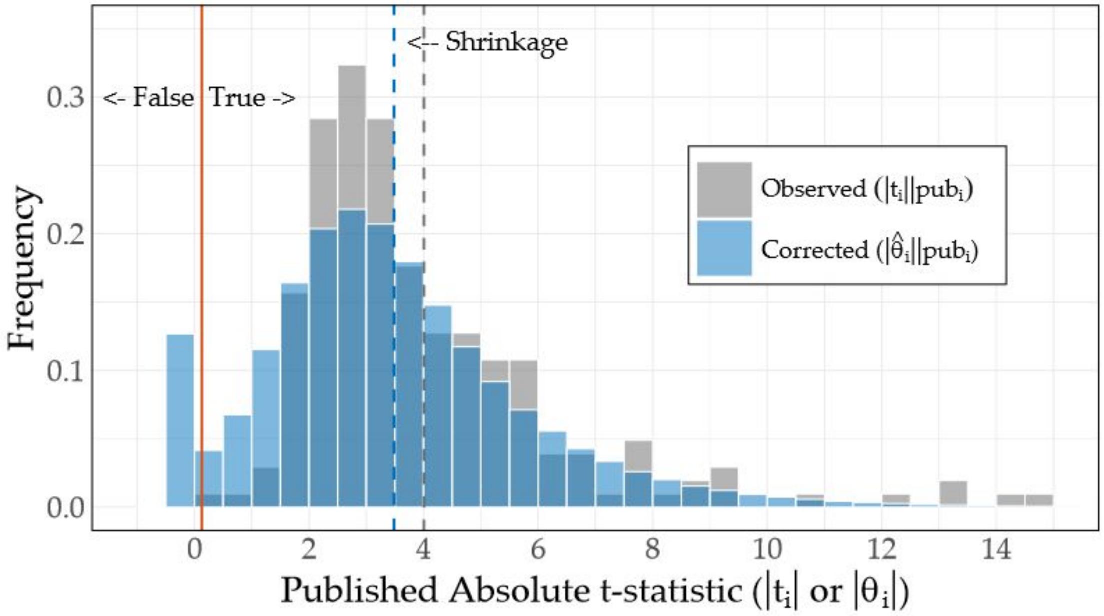


Figure 8: Multiple Testing vs Conservatism in Harvey, Liu, and Zhu (2016). We simulate predictors following Harvey et al. (2016)'s baseline estimate (Table 5, Panel A, $\rho = 0.2$). Normal$(0, 0.1)$ noise is added to the false predictors for ease of viewing. We plot a random sample of 1,172 total predictors, which implies roughly 300 predictors with $|t_i| > 1.96$. The classical 1.96 hurdle implies an FDR of 8.8% (share of hollow markers to the right of solid line). FDR = 5% requires raising the hurdle a bit, to 2.3 (dashed line). Harvey et al. (2016) recommend a more a conservative FDR = 1% (dotted line), or using the Holm (1979) algorithm at FWER $\leq$ 5% (dotted line), among other more conservative methods. Holm is computed using the 1,383 predictors shown. The significant raising of hurdles is due to conservatism, not multiple testing effects.

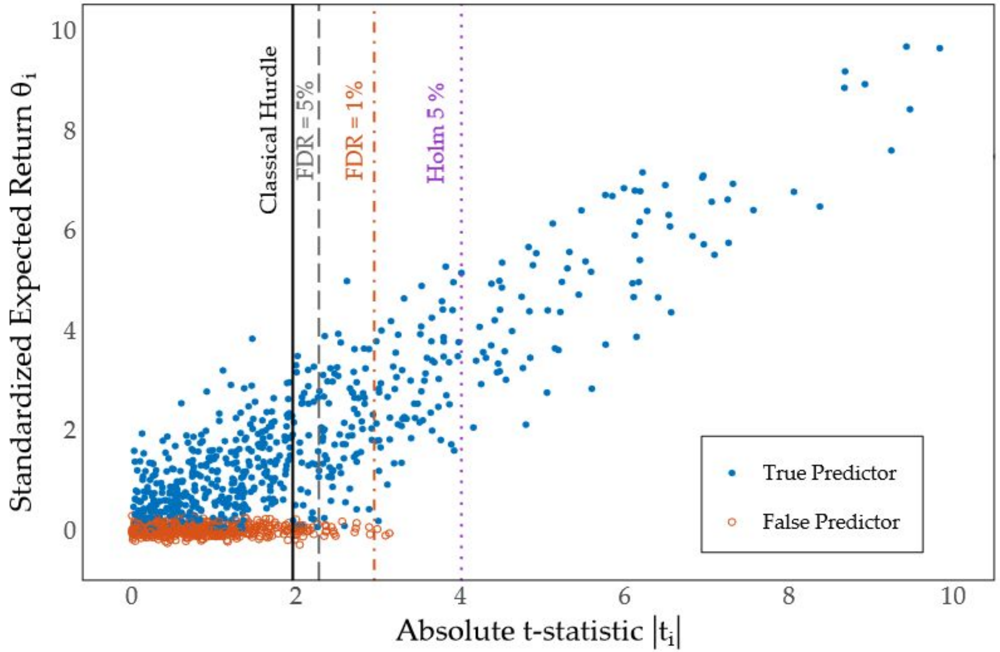


Figure 10: Liquidity Adjustments Decrease Returns by 30%. Data are 207 predictors from CZ22. Grand mean return averages across in-sample months and then averages across predictors. Error bars show to standard errors, approximated by the standard deviation across predictors divided by $\sqrt{207}$. Original implementations follows the original papers. Annual rebalancing updates signal data each year in June. ME $>$ NYSE 20 Pct excludes stocks that fall below the 20th percentile of NYSE market equity. Value-weighted weights stocks by their market equity. Liquidity adjustments robustly decrease expected returns by roughly 30%. Robust effects related to economics should not be equated with data snooping.