Gods' Fight about ...

Andrew Chen vs. HLZ

Haotian Deng

Shanghai University of Finance and Economics

April 10, 2024

Contents

1. Author
2. Motivation
3. Why Most Claimed Statistical Findings Are False
4. Why Most Claimed Statistical Findings Are True

Campbell R. Harvey

Former president of the American Finance Association

- p-hacking and Bayesianized p-value
- Presidential Address: The Scientific Outlook in Financial Economics
- Bayesian Inference in Asset Pricing Tests
- . . . and the Cross-Section of Expected Returns
- False (and Missed) Discoveries in Financial Economics
- Tortured Data (in SFS), Lucky Factors (in Jacobs Levy Center's Conference)

Abstracting from the financial crisis, we conclude that active management of both equity and fixed income has significantly contributed to the returns of the fund.

Andrew Y. Chen

- Publication Bias in Asset Pricing Research
- Peer-reviewed theory does not help predict the cross-section of stock returns
- Zeroing In on the Expected Returns of Anomalies
- Do t-Statistic Hurdles Need to be Raised?
- In Full-Information Estimates, Long-Run Risks Explain at Most a Quarter of P/D Variance, and Habit Explains Even Less
- The Limits of p-Hacking: A Thought Experiment

ABSTRACT → 摘要？抽象！

Suppose that the 300+ published asset pricing factors are all spurious. How much p-hacking is required to produce these factors? If 10,000 researchers generate eight factors every day, it takes hundreds of years. This is because dozens of published t-statistics exceed 6.0, while the corresponding p-value is infinitesimal, implying an astronomical amount of p-hacking in a general model. More structure implies that p-hacking cannot address $\approx$100 published t-statistics that exceed 4.0, as they require an implausibly nonlinear preference for t-statistics or even more p-hacking. These results imply that mis-pricing, risk, and/or frictions have a key role in stock returns.

1. Author
2. **Motivation**
3. Why Most Claimed Statistical Findings Are False
4. Why Most Claimed Statistical Findings Are True

p-hacking: racing down the path in pursuit of *p*-value

- Once the null hypothesis that ”the expected return of the factor is zero” is rejected, people will readily assume that ”the expected return of the factor is significantly non-zero”, and that ”the lower the p-value, the more significant the expected return of the factor”, thus pursuing lower p-values.
- A sufficiently low p-value is only a necessary condition, not a sufficient condition, for the significance of factor expected returns being non-zero.
- Intentional or unintentional data snooping and data manipulation, multiple hypothesis testing problems.

If you torture the data long enough, it will confess.

Publication Bias

- From the perspective of driving scientific progress, ”ineffective factors” and ”effective factors” are equally important.
- If it can be conclusively proven that a certain factor cannot deliver excess returns, then people can confidently avoid that factor, making it still very valuable to investment practice.
- Driven by utilitarian motives, scholars are more willing to spend their research time and energy on factors with low p-values that can be found through various means, only willing to publish the ”most significant” research findings, rather than taking the risk to study ”ineffective factors”.

Soft Science vs. Hard Science

Why there are so many low p-value factors?

- Hard science: Unaffected by researchers' personal preferences, conclusions can be directly drawn from the data, and are highly generalizable.
- Soft science: Easily influenced by the researcher's personal preferences since research results depend on how hypotheses are formulated, how data are processed, and how results are interpreted.

p–value and multiple hypothesis testing

- p–value $\equiv \mathbf{prob}(D \mid H_0) \neq \mathbf{prob}(H_0 \mid D)$
- multiple hypothesis testing
- type I error and false discovery rate

$$\mathrm{FDR} = \mathbb{E}(\frac{N_{0|r}}{R})$$

Table 1: Contingency table in testing M hypotheses

	Unpublished	Published	Total
Truly insignificant	800($N_{0\|a}$)	10($N_{0\|r}$)	810
Truly significant	100($N_{1\|a}$)	90($N_{1\|r}$)	190
Total	900($M - R$)	100(R)	1000(M)

1. Author
2. Motivation
3. **Why Most Claimed Statistical Findings Are False**
4. Why Most Claimed Statistical Findings Are True

Benjamini, Hochberg, and Yekutieli's (BHY) Adjustment

1. Order the original p-values such that

$$p_{(1)} \leq p_{(2)} \leq \cdots p_{(b)} \leq \cdots \leq p_{(M)}$$

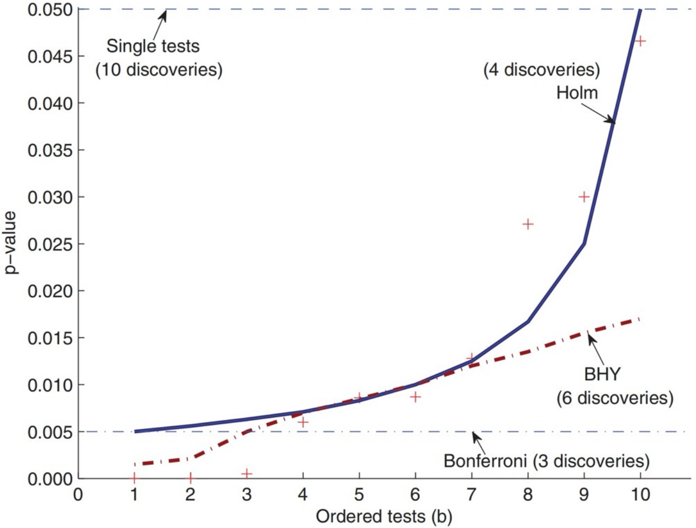
2. Let k be the maximum index such that

$$p_{(b)} \leq \frac{b}{M \times c(M)} \alpha_d$$

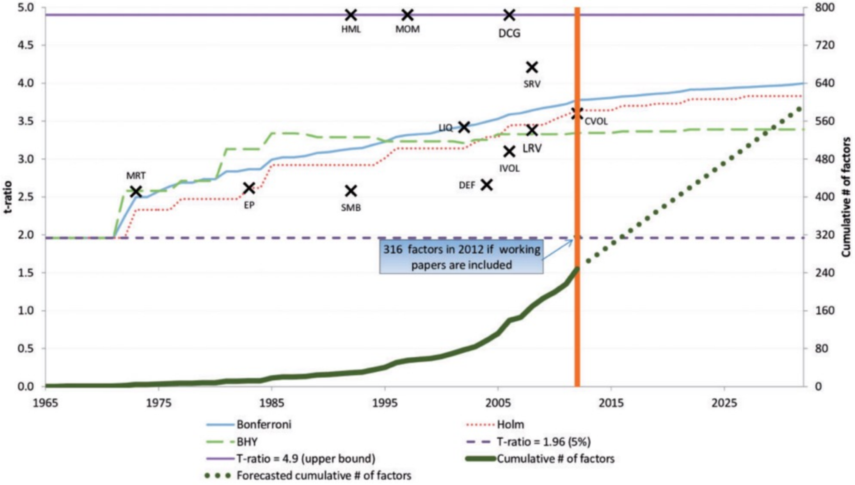
3. Reject null hypotheses $H_{(1)} \cdots H_{(k)}$, but not others

$$p_{(i)}^{\text{BHY}} = \begin{cases} p_{(M)} & \text{if } i = M, \\ \min\left[p_{(i+1)}^{\text{BHY}}, \frac{M \times c(M)}{i} p_{(i)}\right] & \text{if } i \leq M-1, \end{cases}$$

Multiple Test Thresholds



Adjusted t-statistics, 1965–2032



An important Philosophical Issue and Conclusion

Why should we have a higher threshold for today's data mining than for data mining in the 1980s?

1. The rate of discovering a true factor has decreased.
2. There is a limited amount of data (CRSP database).
3. The cost of data mining has dramatically decreased.

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 Bonferonni, 142 under Holm, 132 under BHY (1%), and 80 under BHY (5%)

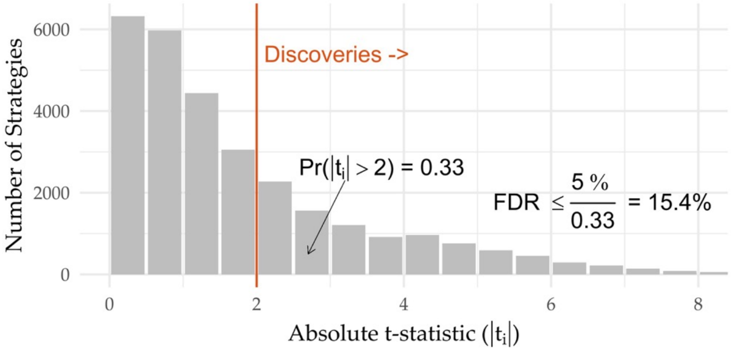
“easy bound” on $\text{FDR}_{|t|>2}$

$$\text{FDR}_{|t|>2} \approx \underbrace{\Pr\left(F_i \mid |t_i| > 2\right)}_{\text{the predictor is false when } t>2} = \frac{\Pr\left(|t_i| > 2 \mid F_i\right)}{\Pr\left(|t_i| > 2\right)} \Pr\left(F_i\right)$$

$$\leq \frac{\underbrace{\Pr\left(|t_i| > 2 \mid F_i\right)}_{t>2 \text{ when the predictor is false}}}{\Pr\left(|t_i| > 2\right)} \approx \frac{5\%}{\Pr\left(|t_i| > 2\right)}$$

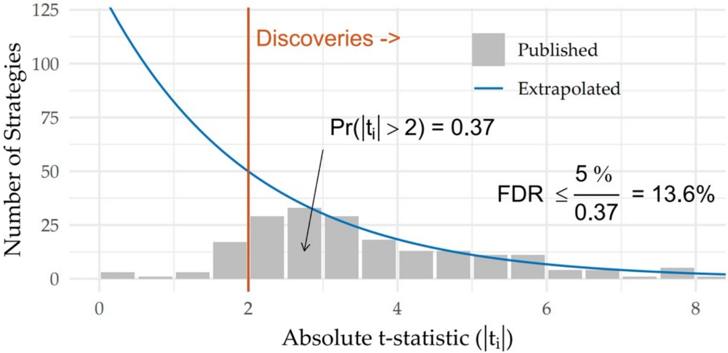
- One might estimate $\Pr\left(|t_i| > 2\right)$ by just counting the share of $|t_i| > 2$.
- Since publication bias, $\Pr\left(|t_i| > 2\right)$ may be overstated.
- This problem can be addressed by considering worst-case scenarios.

Bounding the FDR using Data-Mining as a Worst-Case



Thus, at least 84.6% of published predictors are true.

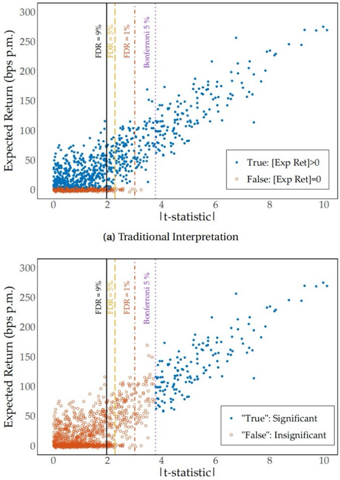
Bounding the FDR with Conservative Extrapolation



$$\mathrm{FDR}_{|t|>2} \leq \frac{5\%}{\exp\left[-2/\mathbb{E}\left(|t_i|\right)\right]} = \frac{5\%}{0.37} = 13.6\% \text{ (exponential CDF)}$$

$$\mathbb{E}\left(|t_i|\right) = \mathbb{E}\left(|t_i| \mid |t_i| > 2\right) - 2 \quad \text{(Due to the memoryless property)}$$

Two Interpretations of “False Factor”



(a) Traditional Interpretation

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

- HLZ’s estimates imply $\text{FDR}_{|t|>2} \approx 9\%$
- why do they (HLZ) argue that “most … are more likely false?”
- “false factor” vs “insignificant factor” (158/296=53%)
- “most” vs “many” (80/296=27%)
- “null” vs “insignificant”

From “Factor War” to “Test War”

- Lo and MacKinlay (1990): Data snooping can cause problems in testing asset pricing models.
- Sullivan et al. (1999): How to correct the impact of data snooping bias
- Harvey et al. (2016): After excluding the impact of multiple hypothesis testing, “the vast majority” fail to achieve significant excess returns.
- Linnainmaa and Roberts (2018): Constructe new out-of-sample data to test.
- Chordia et al. (2020): Creative use of simulation methods to infer how statistical characteristics of study-based factor sets can eliminate the effects of multiple hypothesis testing。

Thanks!