American Finance Association

Discussion of

Is There a Replication Crisis in Finance?

Jensen, Kelly and Pedersen

Campbell R. Harvey
Duke University and NBER

Paper offers

- Contributes to the debate on replication crisis
- Proposes a Empirical Bayes model that “leads to different conclusions”, i.e., **there is no crisis**.

Big picture issues

2005

- Many fields have already gone through their “replication crises” others are late.

Why Most Published Research Findings Are False

John P. A. Ioannidis

Ioannidis JPA (2005) Why Most Published Research Findings Are False. PLoS Med 2(8): e124.

Big picture issues

- Fields like genome association studies and finance failed to correct inference for multiple testing

… and the Cross-Section of Expected Returns

Hundreds of papers and factors attempt to explain the cross-section of expected returns. Given this extensive data mining, it does not make sense to use the usual criteria for establishing significance. Which hurdle should be used for current research? Our paper introduces a new multiple testing framework and provides historical cutoffs from the first empirical tests in 1967 to today. A new factor needs to clear a much higher hurdle, with a *t*-statistic greater than 3.0. We argue that most claimed research findings in financial economics are likely false. (*JEL* C12, C52, G12)

Campbell R. Harvey, Yan Liu, Heqing Zhu, … and the Cross-Section of Expected Returns, *The Review of Financial Studies*, Volume 29, Issue 1, January 2016, Pages 5–68.

Big picture issues

But this is more than a multiple testing problem

- Editors want to publish impactful papers
- Papers with negative results (insignificant) have low cites
- Authors are strongly incented to publish – in many schools a single A-level is enough for tenure
- Authors need a “significant” result
- Incentivizes p-hacking expeditions

Big picture issues

Should finance be treated differently? Why are so many results "significant" in finance? (Harvey 2017)

1. **Do we have better theories in financial economics than, say, in particle physics?**
2. By observing phenomena, do we have better prior beliefs than other fields?
3. In financial economics, are our hypotheses often narrow and focused?
4. Is the connection between theories, hypotheses, and empirical findings more flexible in financial economics? [Think of testing the CAPM]
5. Is it more likely that there are interaction effects between the researcher and the effect being researched? [e.g., confirmation bias.]

Big picture issues

Should finance be treated differently? Why are so many results "significant" in finance? (Harvey 2017)

6. **Is there manipulation of data and results?** [Hard misconduct rare. However, p-hacker has a large toolbox which includes strategic data selection, outlier exclusion, winsorization rules, variable transformation, instrument selection, estimation method]
7. Is there a lack of a replication culture? [Until recently, replications were not considered A-level research in finance.]
8. **Is it hard to publish findings that do not "support" the hypothesis being tested?**

The evidence

McLean and Pontiff, JF 2016

"We study the out-of-sample and post-publication return predictability of 97 variables shown to predict cross-sectional stock returns. Portfolio returns are **26% lower out-of-sample and 58% lower post-publication**."

The evidence

Linnainmaa and Roberts, RFS 2018

"Using data spanning the twentieth century, we show that the **majority** of accounting-based return anomalies, including investment, **are most likely an artifact of data snooping**."

Juhani T Linnainmaa, Michael R Roberts, *The Review of Financial Studies*, Volume 31, Issue 7, July 2018, Pages 2606–2649,

The evidence

Bailey and Lopez de Prado, (RSS-Significance)

"Finance's inability to conduct controlled experiments makes it virtually impossible to debunk a false claim. One would hope that, in such a field, **researchers would be particularly careful when conducting statistical inference. Sadly, the opposite is true**."

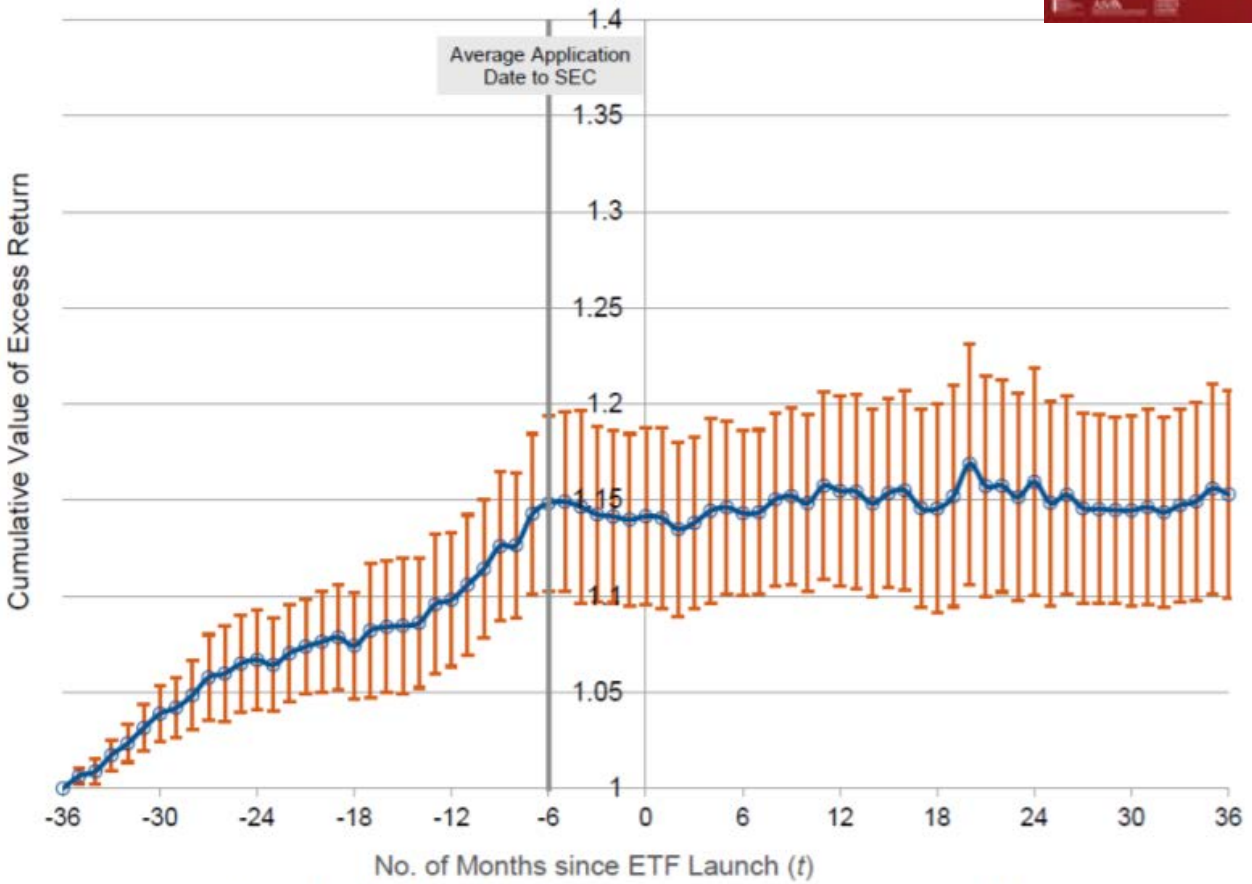


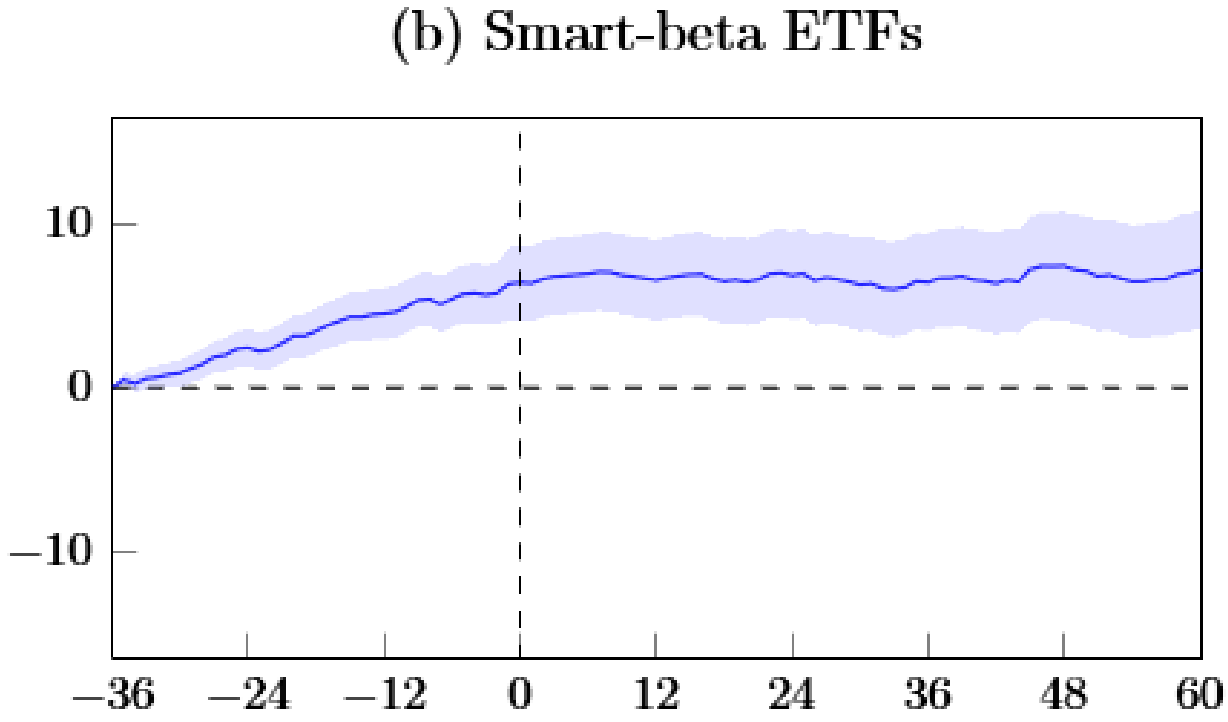
Figure 3 – Backtested performance vs. performance out-of-sample. Reproduced from [BrightLi2015], Figure 3, with permission.

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3895330

The evidence

Ben-David, Franzoni, Kim and Moussawi (2021, SSRN)

"The authors argue that portfolios of smart-beta ETFs are designed by overfitting the data (p-hacking) to generate indexes that outperform before launch (a similar argument is made in Harvey, 2021), but **deliver zero outperformance after launch.** We confirm these results…"



https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3765063

Replicating Jensen, Kelly and Pedersen

I thank the authors for providing code and data before publication! **The results replicate.**

- Focus on **key parameter** π_1 which is the prior fraction of true anomalies among tests (true fraction for DGP in JKP Empirical Bayes). Note π_0 is the fraction of true nulls (1- π_1)
- Simulate 130 strategies over 68 years - successfully reproduce JKP.
- Simulate 10,000 times generating 10,000 anomaly populations.

[24] We define the true discovery rate to be the number of significantly positive alphas according to, respectively, OLS, BY, and EB divided by the number of truly positive alphas. Given our simulation structure, half of the alphas are expected to be positive in any simulation. Some of these will be small (i.e., economically insignificant) positives, so a testing procedure would require a high degree of statistical power to detect them. This is why the true discovery rate is below one even for high values of τ_c.

Sensitivity Analysis: Jensen, Kelly and Pedersen

Compare: Standard multiple testing adjustment BH, Empirical Bayes (EB) and unadjusted OLS

- Big question: **What is a reasonable value for π_1?**
- JKP assume π_1 = 50%, we examine other values, e.g., what if π_1 = 25% or π_1 = 5% true?
- Is the prior of 50% for true anomalies reasonable? (I personally dealt with thousands of submissions to the JF claiming a significant factor.)
- My 400 factor list is only for top journals. There are 111 finance journals in Clarivate's *Journal Citation Reports*

Sensitivity Analysis: Jensen, Kelly and Pedersen

Results are sensitive

- Assume we are targeting a 5% False Discovery Rate (FDR)
- π_1 = 50%, EB does well achieving 5.7% FDR. BH is a bit conservative at 2.6%. Even OLS does well at 7.0%
- π_1 = 25%, EB has 14.0% FDR. BH is close at 3.8%.
- π_1 = 5%, EB does poorly with 33.6% FDR. BH is accurate at 4.5%.
- π_1 = 1%, EB does very poorly with 56.4% FDR. BH is accurate at 4.7%. OLS horrible at 77.0%.

Storey 2003

The Annals of Statistics
2003, Vol. 31, No. 6, 2013–2035

THE POSITIVE FALSE DISCOVERY RATE: A BAYESIAN INTERPRETATION AND THE q-VALUE[1]

BY JOHN D. STOREY

None of this is surprising!

- Storey (2003) Theorem #1 shows that the positive false discovery rate* can be written:

$$pFDR = \frac{Type\ I\ error}{Type\ I\ error + \frac{\pi_1}{\pi_0}}$$

- In JKP, the true-false ratio $\frac{\pi_1}{\pi_0}$ = 1, so

*p*FDR ≈ Type I error, e.g., at 5%, *p*FDR= 4.8%

**p*FDR is related to the FDR. It is sometimes called the posterior Bayesian Type I error. See Storey (2003).

Storey 2003

None of this is surprising!

- If $\frac{\pi_1}{\pi_0}$ << 1, which is the case for most applications of multiple testing, *p*FDR will be much larger than the Type I error
- This implies the need for a higher t-statistic cutoff
- It is not clear the EB approach is a useful approach for multiple testing

Should priors be based on published anomalies?

Other issues

- JKP elicit their priors based on published anomalies (or more precisely, anomalies they constructed by themselves) alone.
- This has a large impact on their conclusion.
- Other papers in this literature strive to back out the anomaly population, e.g., Harvey, Liu and Zhu (2016).

Conclusions

- I am encouraged by research about how we do research. It is about time.
- I welcome debate about the “replication crisis”
- It is very hard to find anomalies that “beat the market”
- **There is no reason that the field of finance should get a “free pass”.**

Conclusions

- The results in this paper are consistent with Storey (2003)
- Key results in the paper depend on the assumption that 50% of anomalies are true.
- **That's not my prior and I doubt it is yours.**

Conclusions

- **The first step toward the resolution of the replication crisis in finance is to acknowledge it**

Soon to be released:

Harvey, Liu and Sancetta (2022)

"Yes, There is Replication Crisis in Finance"

Appendix

Suppose only one published anomaly with t=5

- Consider two possibilities: A) 10 trials and B) 1000 trials to find the anomaly
- With multiple testing adjustments, there would be a moderate adjustment to the significance level for A) and a substantial adjustment for B)
- With EB, nothing would happen.