

# Lecture 2: Fund Performance, Skill, and Luck

Institutional Investors

Calvin J. Chiou

2026-10-14

## Table of contents

|                                                                    |    |
|--------------------------------------------------------------------|----|
| Overview . . . . .                                                 | 2  |
| 1. Motivation: The Luck-vs.-Skill Problem . . . . .                | 3  |
| Two Fundamental Questions . . . . .                                | 4  |
| 2. Equilibrium Accounting (Fama & French 2010) . . . . .           | 4  |
| 2.1 The Zero-Sum Identity . . . . .                                | 4  |
| 2.2 Aggregate Portfolio Results . . . . .                          | 5  |
| 3. Bootstrap Simulation Methodology . . . . .                      | 6  |
| 3.1 The Problem with Standard Inference . . . . .                  | 6  |
| 3.2 The Bootstrap Procedure . . . . .                              | 6  |
| 3.3 Why Joint Sampling Matters . . . . .                           | 7  |
| 4. Bootstrap Results: Luck vs. Skill . . . . .                     | 7  |
| 4.1 Net Returns: Few Managers Cover Costs . . . . .                | 7  |
| 4.2 Gross Returns: Skill Exists, but Is Modest . . . . .           | 8  |
| 4.3 Graphical Summary: The CDF Comparison . . . . .                | 8  |
| 4.4 A Competing Route: The False Discovery Rate . . . . .          | 9  |
| 4.5 Reassessing the FDR: Skill, Luck, or Lack of Power? . . . . .  | 9  |
| 5. The Value-Added Approach (Berk & van Binsbergen 2015) . . . . . | 12 |
| 5.1 Why Alpha Fails as a Skill Measure . . . . .                   | 12 |
| 5.2 The Value-Added Measure . . . . .                              | 13 |
| 5.3 The Benchmark: Vanguard Index Funds . . . . .                  | 13 |
| 6. The Berk-Green (2004) Model . . . . .                           | 14 |
| 7. Key Results: Berk & van Binsbergen (2015) . . . . .             | 15 |
| 7.1 Average Fund Adds Positive Value . . . . .                     | 15 |
| 7.2 Persistence of Value Added . . . . .                           | 16 |
| 7.3 Investors Can Identify Skill . . . . .                         | 16 |
| 7.4 Where Does the Value Go? . . . . .                             | 17 |
| 8. Reconciling the Three Papers . . . . .                          | 17 |

|                                                              |    |
|--------------------------------------------------------------|----|
| 9. Implications for Practice . . . . .                       | 18 |
| 9.1 For Performance Evaluation . . . . .                     | 18 |
| 9.2 For Investors . . . . .                                  | 18 |
| 9.3 For Research . . . . .                                   | 19 |
| 10. Empirical Implementation . . . . .                       | 19 |
| Setup . . . . .                                              | 19 |
| Load Data from Lecture 1 . . . . .                           | 20 |
| Step 1 — Estimate Baseline Carhart Alphas . . . . .          | 21 |
| Step 2 — Impose Zero Alpha (Demean Returns) . . . . .        | 22 |
| Step 3 — Bootstrap Simulation . . . . .                      | 23 |
| Step 4 — Compare Actual vs. Simulated t-Statistics . . . . . | 25 |
| Step 5 — Summarize Luck vs. Skill Evidence . . . . .         | 27 |
| Simulating the Value-Added Measure . . . . .                 | 29 |
| Illustrating the Power Problem . . . . .                     | 30 |
| 11. Summary Statistics and Comparison . . . . .              | 35 |
| 12. Key Takeaways . . . . .                                  | 36 |
| Discussion Questions . . . . .                               | 37 |
| References . . . . .                                         | 38 |

## Overview

Lecture 1 established the **Carhart four-factor model** as the standard benchmark for evaluating equity fund performance. A natural follow-up question is: once we estimate alpha, how do we know whether it reflects *skill* or simply *luck*?

This lecture digs into three papers on this question:

- **Fama and French (2010)**: Use bootstrap simulations to separate luck from skill in the cross-section of mutual fund alphas. Their sobering conclusion: for net returns, few funds have sufficient skill to cover their costs.
- **Berk and van Binsbergen (2015)**: Reframe the question entirely. Instead of asking whether alpha is positive, they ask how much *dollar value* a fund extracts from markets. Their finding: the average fund adds roughly \$3.2 million per year—but competition drives net returns to investors toward zero.
- **Andrikogiannopoulou and Papakonstantinou (2019)**: Ask a prior question—*can these tests detect skill even when it is there?* They show that the false discovery rate (FDR) methodology, a leading alternative to the bootstrap, misclassifies most genuinely skilled managers as having zero alpha, because fund returns are too noisy for the underlying tests to have power.

Together, the first two papers reveal a fundamental tension in the fund performance literature:

**Skill exists** (Berk & van Binsbergen), **but investors cannot easily capture it** (Fama & French).

The third paper adds a methodological warning that cuts across both:

**A finding of “no skill” may be a finding about our instruments rather than about managers.**

We will:

1. Understand the **equilibrium accounting** framework for evaluating fund performance in aggregate
2. Master the **bootstrap simulation** methodology for distinguishing luck from skill
3. Examine the **false discovery rate** approach and why (andrikogiannopoulou2019?) argue it fails in this setting
4. Learn the **value-added** measure of skill and why it differs from alpha
5. Examine the **Berk-Green (2004)** model and its empirical predictions
6. Implement a simplified bootstrap exercise in R using CRSP-style mutual fund data, and simulate the **power problem** directly

---

## 1. Motivation: The Luck-vs.-Skill Problem

Suppose you observe a fund with a four-factor alpha of +2% per year over 10 years. Should you conclude the manager is skilled? Not necessarily. With thousands of funds, some will have high alphas purely by chance. The core challenge is:

$$\underbrace{\hat{\alpha}_i}_{\text{estimated}} = \underbrace{\alpha_i}_{\text{true skill}} + \underbrace{\text{noise}}_{\text{estimation error}}$$

The noise component is not small. The average standard error of a fund’s monthly alpha estimate is about 0.28% per month (roughly 3.4% annualized), as documented by Fama and French (2010). With only 10 years of data (120 monthly observations), detecting a true annual alpha of 2% at the 5% significance level requires:

$$t = \frac{\hat{\alpha}}{\hat{\sigma}/\sqrt{T}} = \frac{2\%/12}{3.4\%/\sqrt{120}} \approx 0.63$$

This is far below the conventional threshold of 2.0. **Statistical power is the central problem** in evaluating individual fund performance—a point that Wermers (2011) emphasizes and that Lecture 1 introduced.

! Important

Hold on to this calculation. In Section 4.5 we will see that (andrikogiannopoulou2019?) build an entire critique of the modern skill-measurement literature out of exactly this arithmetic. The low-power problem is not a caveat to be noted and set aside; it can reverse the economic conclusion.

## Two Fundamental Questions

Fama and French (2010) and Berk and Binsbergen (2015) answer *different* but complementary questions:

Table 1: Two perspectives on fund performance

| Question            | Fama & French (2010)         | Berk & van Binsbergen (2015)                   |
|---------------------|------------------------------|------------------------------------------------|
| Unit of analysis    | Return to investors          | Dollar value extracted from markets            |
| Measure of skill    | Net / gross alpha            | Value added = AUM $\times$ gross excess return |
| Key finding         | Few funds cover costs        | Average fund adds \$3.2M/year                  |
| Conclusion on skill | Largely absent (net of fees) | Widespread; investors recognize it             |

## 2. Equilibrium Accounting (Fama & French 2010)

### 2.1 The Zero-Sum Identity

Fama and French (2010) begin with a powerful accounting constraint called **equilibrium accounting**. The aggregate of all investors holds the market portfolio. If passive investors hold passive strategies with zero alpha, then the aggregate portfolio of *active* investors must also have zero alpha—before costs.

This is essentially Sharpe’s Arithmetic (Sharpe 1991):

$$\underbrace{\alpha_{\text{active, aggregate}}}_{\text{must be zero (before costs)}} = 0$$

After costs:

$$\alpha_{\text{active, aggregate}}^{\text{net}} = -\text{expense ratio} < 0$$

**Implication:** Active management is a *zero-sum* game before costs and a *negative-sum* game after costs. If some funds have positive net alpha, other funds must have an equally large negative net alpha.

## 2.2 Aggregate Portfolio Results

Fama and French (2010) estimate regression (1) for the value-weighted (VW) and equal-weighted (EW) portfolios of all actively managed U.S. equity mutual funds, 1984–2006:

$$R_{i,t} - R_{f,t} = a_i + b_i(R_{M,t} - R_{f,t}) + s_i\text{SMB}_t + h_i\text{HML}_t + m_i\text{MOM}_t + e_{i,t} \quad (1)$$

Key results from their Table II:

Table 2: Aggregate fund performance: net returns (Fama & French, 2010, Table II)

| Portfolio | Model        | Net (ann.) | t-stat |
|-----------|--------------|------------|--------|
| EW        | Three-factor | −0.93%     | −2.13  |
| EW        | Four-factor  | −0.92%     | −2.05  |
| VW        | Three-factor | −0.81%     | −2.50  |
| VW        | Four-factor  | −1.00%     | −3.02  |

The message is clear: **on net returns, aggregate active management destroys value by approximately the amount of expense ratios** (roughly 1% per year during this period). The VW four-factor alpha of −1.00% closely matches the average expense ratio of actively managed funds.

When returns are measured *gross* (before expense ratios), the alpha is close to zero:

Table 3: Aggregate fund performance: gross returns (Fama & French, 2010, Table II)

| Portfolio | Model        | Gross (ann.) | t-stat |
|-----------|--------------|--------------|--------|
| EW        | Three-factor | +0.36%       | +0.85  |
| EW        | Four-factor  | +0.39%       | +0.90  |
| VW        | Three-factor | +0.13%       | +0.40  |
| VW        | Four-factor  | −0.05%       | −0.15  |

The near-zero gross alpha is consistent with equilibrium accounting: in aggregate, active managers are no better or worse than passive investors at picking stocks—but they charge fees.

**i** Note

**Key insight from equilibrium accounting:** The aggregate results say nothing about whether *individual* fund managers have skill. It is possible (and, as we will see, likely) that skilled and unskilled managers coexist, with the talented ones roughly offsetting the incompetent ones in aggregate.

---

### 3. Bootstrap Simulation Methodology

The central methodological contribution of Fama and French (2010) is the **bootstrap simulation** approach to distinguishing luck from skill in the cross-section of fund alphas.

#### 3.1 The Problem with Standard Inference

With thousands of funds, standard t-tests for individual fund alphas are unreliable for two reasons:

1. **Multiple comparisons:** With 3,000+ funds, many will have extreme alpha estimates purely by chance.
2. **Cross-fund correlation:** Fund returns are correlated (they hold similar stocks), so the cross-sectional distribution of t-statistics differs from the t-distribution assumed by standard tests.

#### 3.2 The Bootstrap Procedure

Fama and French (2010) implement the following procedure:

**Step 1 — Demean to impose zero true alpha:**

For each fund  $i$ , subtract its estimated alpha from its monthly returns to create “zero-alpha” returns:

$$r_{i,t}^* = r_{i,t} - \hat{\alpha}_i$$

**Step 2 — Bootstrap:**

Draw a random sample (with replacement) of 273 months from the actual calendar months. Apply the *same* random draw to all funds simultaneously. This preserves: - The cross-correlation

structure of fund returns - The time-series properties of explanatory returns (market, SMB, HML, MOM)

**Step 3 — Estimate alphas on simulated data:**

Run the four-factor regression for each fund on the simulated return series. Record the t-statistic  $t(\hat{\alpha})$  for each fund.

**Step 4 — Repeat 10,000 times:**

Build the distribution of  $t(\hat{\alpha})$  under the null hypothesis of zero true alpha.

**Step 5 — Compare:**

Compare the actual cross-section of  $t(\hat{\alpha})$  from real fund data to the simulated distribution. If the actual right tail is significantly fatter than the simulated right tail, this is evidence of genuine skill.

### 3.3 Why Joint Sampling Matters

A crucial feature of the Fama and French (2010) approach is that they draw the *same* random months for all funds. This is in contrast to Kosowski et al. (2006), who simulate each fund independently. Joint sampling matters because:

- It preserves the correlation of alpha estimates across funds
- It captures any correlated heteroskedasticity
- It prevents false evidence of skill that arises from ignoring cross-fund dependence

Fama and French (2010) show that when Kosowski et al.'s stronger results are re-examined using joint sampling, the evidence for skill in the right tail weakens considerably.

---

## 4. Bootstrap Results: Luck vs. Skill

### 4.1 Net Returns: Few Managers Cover Costs

Fama and French (2010)'s main result on net returns is sobering. The cross-section of  $t(\ )$  estimates from actual fund returns lies almost entirely to the *left* of the simulated distribution:

- Below the 80th percentile, actual net return  $t(\ )$  beats the simulated values in less than **1% of simulation runs**
- Even at the 90th percentile, **90% of simulations beat** the actual  $t(\ )$

**Interpretation:** Most funds do not have sufficient skill to cover their costs. A fund at the 90th percentile of historical  $t()$ —which superficially looks very good—is in fact consistent with luck, not skill.

The one glimmer of hope: at the **97th–99th percentiles**, the actual  $t()$  estimates are roughly in line with the simulated distribution. This suggests a small number of truly skilled managers exist, but they are hidden among lucky underperformers.

#### 4.2 Gross Returns: Skill Exists, but Is Modest

When returns are measured gross of expenses, evidence of both positive and negative skill emerges:

- **Left tail:** Actual  $t()$  is more extreme than simulated  $\rightarrow$  evidence of *truly unskilled* managers who produce negative alpha before fees
- **Right tail:** Actual  $t()$  exceeds simulated values  $\rightarrow$  evidence of *genuinely skilled* managers who outperform passive benchmarks before fees

Fama and French (2010) quantify the cross-sectional dispersion of true alpha using additional simulations with *injected* alpha. Their estimate:

$$\sigma(\alpha^{\text{true}}) \approx 1.25\% \text{ per year}$$

This is a normal distribution with mean zero. Implications:

- Roughly **16%** of funds have true gross alpha  $> 1.25\%$  per year ( 10 bp/month)
- Only **2.4%** have true gross alpha  $> 2.50\%$  per year ( 21 bp/month)

! Important

**The skill that exists is not large enough to cover costs for most managers.** A true gross alpha of 1.25% barely covers a 1% expense ratio. After fees, most skilled managers produce approximately zero alpha for investors.

#### 4.3 Graphical Summary: The CDF Comparison

The key figure in Fama and French (2010) compares the cumulative distribution functions (CDFs) of actual vs. simulated  $t()$ :

- For **net returns:** The actual CDF lies to the *left* of the simulated CDF almost everywhere  $\rightarrow$  negative skill overall
- For **gross returns:** The actual CDF crosses the simulated CDF around the 60th–70th percentile  $\rightarrow$  positive skill in the right tail, negative skill in the left tail

#### 4.4 A Competing Route: The False Discovery Rate

The bootstrap tells us whether the *tails* of the cross-section look unusual, but it does not directly tell us *what fraction* of funds have positive, zero, or negative alpha. (barras2010?)—hereafter BSW—propose an approach that does: the **false discovery rate** (FDR), imported from the multiple-testing literature in statistics (benjamini1995?).

The logic is elegant. Run a separate alpha test for every fund and collect the  $I$  p-values  $\hat{p}_i$ . Under the null of zero alpha, a p-value is uniform on  $[0, 1]$ . So if we pick a high threshold  $\lambda$ —say 0.5—and assume that *every* p-value above  $\lambda$  comes from a genuinely zero-alpha fund, we can count those p-values and extrapolate to the whole interval:

$$\hat{\pi}^0(\lambda) = \frac{\frac{1}{1-\lambda} \cdot \#\{\hat{p}_i : \hat{p}_i > \lambda\}}{I} \quad (2)$$

Knowing  $\hat{\pi}^0$ , we can then take the funds that *did* reject at some significance level  $\gamma$ , subtract the number we expect to have rejected by luck alone ( $\hat{\pi}^0 \cdot \gamma$ ), and recover the proportions of truly negative- and positive-alpha funds:

$$\hat{\pi}^-(\gamma, \lambda) = \frac{\#\{(\hat{\alpha}_i, \hat{p}_i) : \hat{\alpha}_i < 0, \hat{p}_i < \gamma\}}{I} - \hat{\pi}^0(\lambda) \cdot \frac{\gamma}{2} \quad (3a)$$

$$\hat{\pi}^+(\gamma, \lambda) = \frac{\#\{(\hat{\alpha}_i, \hat{p}_i) : \hat{\alpha}_i > 0, \hat{p}_i < \gamma\}}{I} - \hat{\pi}^0(\lambda) \cdot \frac{\gamma}{2} \quad (3b)$$

Applied to U.S. equity mutual funds, BSW report that **75% of funds have zero alpha net of expenses, 24.4% have negative alpha, and only 0.6% have positive alpha**. These numbers have been enormously influential. They are read as direct confirmation of Berk and Green (2004): if almost every fund sits at exactly zero net alpha, that is what a competitive equilibrium with rational capital reallocation looks like.

The method spread well beyond mutual funds—to hedge funds, trading strategies, takeover announcements, and jump detection in asset returns.

#### 4.5 Reassessing the FDR: Skill, Luck, or Lack of Power?

(andrikogiannopoulou2019?) ask whether the FDR can actually do what it claims in this setting. Their answer is no, and the reason is the arithmetic from Section 1.

Look again at equation (2). It rests on an assumption that is easy to miss: *every p-value above  $\lambda$  belongs to a zero-alpha fund*. That holds only if nonzero alphas are large relative to their standard errors—in other words, only if the individual tests have **power**. If a fund has a real alpha of 1% per year but its alpha is estimated with a standard error of comparable size, its

p-value is essentially a draw from the uniform distribution too. It lands above  $\lambda$  roughly as often as a genuinely zero-alpha fund's would, and the FDR counts it as one.

The bias runs in one direction. Because equation (2) divides by  $1 - \lambda$ , each skilled fund that fails to reject contributes  $1/(1 - \lambda) = 2$  funds to the zero-alpha tally when  $\lambda = 0.5$ . **The FDR systematically overstates the proportion of zero-alpha funds and understates the proportion with genuine skill.**

! The headline result

(andrikogiannopoulou2019?) find that, at the sample sizes actually available in mutual fund data, the FDR misclassifies as zero-alpha:

- **90%** of funds whose true alpha is  $\pm 1\%$  per year
- **65%** of funds whose true alpha is  $\pm 2\%$  per year
- roughly **30%** of funds whose true alpha is as large as  $\pm 3.5\%$  per year

A 2% annual alpha is not a rounding error. It is roughly twice the average expense ratio, and a manager who reliably delivered it would be a star. The methodology cannot see two-thirds of them.

### Why BSW's own simulation looked so convincing

BSW validated the FDR with a simulation, and in that simulation it performed beautifully. (andrikogiannopoulou2019?) show why: the simulation was run under conditions far more favorable than the real data.

Table 4: Why the BSW simulation overstates FDR accuracy

| BSW simulation assumption                                  | Reality in mutual fund data                          |
|------------------------------------------------------------|------------------------------------------------------|
| Nonzero alphas at $-3.2\%$ and $+3.8\%$ per year           | Large share of funds have alphas between 1% and 2.5% |
| $T = 384$ months per fund (the <i>maximum</i> in the data) | Median 150, mean 180 observations per fund           |
| $N = 1,400$ funds, balanced panel                          | 2,076 funds, unbalanced panel                        |
| Little or no cross-sectional error correlation             | Substantial correlation from common strategies       |

Each of these pushes in the same direction. The result is that a methodology which looks nearly unbiased in its original validation becomes badly biased once the data-generating process resembles the data it is applied to.

### Three findings that make this hard to fix

**More funds do not help.** This is the most counterintuitive result. Increasing  $N$  leaves the bias essentially unchanged, because  $N$  affects neither the power of each individual test nor the distribution of p-values. (andrikogiannopoulou2019?) put it memorably: it is like estimating a population’s mean weight by weighing more and more people on a biased scale. Only  $T$ —observations per fund—matters, and convergence is glacial. They find that **even 200 years of monthly data would not be enough** to get accurate estimates for realistic alpha distributions.

**Confidence intervals are not a safety net.** For roughly half the data-generating processes they examine, the nominal 95% confidence intervals around the FDR estimates **never** contain the true proportions across 1,000 simulated samples. Realistic cross-sectional correlation in residuals inflates the estimator’s variability by about a factor of five, widening 90% intervals by 4.6 to 7.5 times.

**The correction may not beat no correction.** For alphas of 1% or 2% per year, the FDR does not outperform the naive “no-luck” approach of simply counting significant alphas without any multiple-testing adjustment—and can do worse, because its bias is so large.

### The connection back to Fama and French

There is a result here that speaks directly to Section 4.2. Recall that Fama and French (2010) estimate the cross-sectional dispersion of true alpha at about 1.25% per year. (andrikogiannopoulou2019?) simulate a data-generating process calibrated close to that estimate—alphas drawn from a normal with  $\sigma_\alpha = 1.5\%$ , with only **1%** of funds having exactly zero alpha—and apply the FDR to it. The FDR reports that **70%** of funds have zero alpha.

In other words: take the world that Fama and French’s own estimates describe, and the FDR will tell you it is the world that BSW found.

#### **i** What this does and does not overturn

(andrikogiannopoulou2019?) is a critique of a **methodology**, not a claim that mutual fund managers are secretly skilled. They do not estimate skill themselves in this paper. What they establish is that the widely cited “75% of funds have zero alpha” figure cannot bear the economic weight placed on it—in particular, it is not evidence for the Berk and Green (2004) equilibrium, because an estimator with this much bias would report approximately that number whether or not Berk-Green were true.

A test that returns the same answer under the hypothesis and its negation is not evidence about the hypothesis.

## What to do instead

The authors are constructive. Their recommendations:

1. **Check power before trusting the estimator.** Simulate your own data-generating process and verify that individual tests have enough power for the FDR's assumptions to hold. This is cheap—we do it in Section 10.
2. **Pool cross-sectional information** rather than testing each fund in isolation. Approaches that estimate the whole cross-sectional distribution of alpha, rather than running  $I$  separate underpowered tests, extract far more from the same data. They point to (harvey2018?) and their own related work in this vein.
3. **Be wary of the FDR wherever the signal-to-noise ratio is low.** The problem is not specific to mutual funds. Any application with large  $N$  and small  $T$ —hedge funds, trading strategies, event studies—inherits it.

---

## 5. The Value-Added Approach (Berk & van Binsbergen 2015)

### 5.1 Why Alpha Fails as a Skill Measure

Berk and Binsbergen (2015) make a fundamental theoretical point: **neither net alpha nor gross alpha correctly measures managerial skill**. The argument proceeds as follows.

Under standard neoclassical assumptions (rational investors, competitive markets, optimizing managers with decreasing returns to scale):

**Net alpha** measures market efficiency, not skill. If investors are rational and markets competitive, the net alpha is driven to zero by capital flows—regardless of skill. A fund with high skill attracts capital until its net alpha is competed away.

**Gross alpha** equals the fee the manager charges. To see this, note:

$$\alpha^{\text{gross}} = \alpha^{\text{net}} + f_i = 0 + f_i = f_i$$

where  $f_i$  is the fund's percentage fee. The gross alpha is a *choice variable* of the manager, not a measure of skill. A manager who indexes more of the portfolio can achieve any desired gross alpha by adjusting the fee accordingly.

This leads to a striking **proposition** (Berk & van Binsbergen 2015, Proposition 1):

*The only condition under which gross alpha measures managerial skill is if all managers set their fees to ensure that the AUM of all funds is exactly \$1.*

Since AUM varies enormously across funds, gross alpha is an unreliable proxy for skill in practice (their Table 8: cross-sectional correlation between value added and gross alpha is only 23%, with  $R^2 = 5\%$ ).

## 5.2 The Value-Added Measure

The correct measure of managerial skill, according to Berk and Binsbergen (2015), is the **value added** of the fund—the dollar value extracted from financial markets:

$$V_{i,t} = q_{i,t-1} \cdot (R_{i,t}^g - R_{i,t}^B) \quad (4)$$

where: -  $q_{i,t-1}$  = assets under management (AUM) at the end of the prior period -  $R_{i,t}^g$  = gross excess return of fund  $i$  at time  $t$  -  $R_{i,t}^B$  = benchmark return (the investor's next best alternative)

The fund's **skill** is the time-series expectation of value added:

$$S_i \equiv \mathbb{E}[V_{i,t}] \quad (5)$$

Intuitively: a manager who adds 1% alpha on a \$10 billion fund adds *more value* (\$100 million) than a manager who adds 10% alpha on a \$1 million fund (\$100,000), even though the former has a much lower gross alpha.

### Tip

Notice that the value-added measure sidesteps part of the problem (**andrikogiannopoulou2019?**) identify. Dollar value added is estimated from the *level* of AUM times realized excess return, and its cross-sectional dispersion is enormous, so the signal-to-noise ratio is far better than it is for percentage alpha. This is one reason Berk and Binsbergen (2015) can document persistence over ten years where alpha-based studies find almost none.

## 5.3 The Benchmark: Vanguard Index Funds

A critical design choice in Berk and Binsbergen (2015) is the **benchmark**. The standard approach (Carhart four factors) uses factor-mimicking portfolios that:

1. Were not actually available to investors in the early part of the sample (e.g., small-cap and value index funds were not widely available until the early 1990s)
2. Do not account for transaction costs
3. May contain hindsight bias

Berk and Binsbergen (2015) instead use **Vanguard index funds** as the benchmark—the actual investment opportunity available to investors at each point in time. This benchmark:

- Is guaranteed to be tradable and marketed
- Includes transaction costs
- Reflects the dynamic evolution of available passive strategies

The FFC (Fama-French-Carhart) factor portfolios themselves produce significantly positive alpha relative to Vanguard benchmarks:

Table 5: Net alpha of FFC factor portfolios vs. Vanguard benchmark (Berk & van Binsbergen 2015, Table 2)

| Factor | Alpha (bp/month) | t-stat | Adjusted $R^2$ |
|--------|------------------|--------|----------------|
| Market | 2                | 0.83   | 99%            |
| SMB    | 22               | 2.80   | 74%            |
| HML    | 35               | 3.37   | 52%            |
| UMD    | 70               | 3.38   | 15%            |

The high alphas of SMB, HML, and UMD reflect the hindsight bias embedded in using these factors: investors in the early sample period could not cheaply access these strategies.

---

## 6. The Berk-Green (2004) Model

Both papers reference the theoretical framework of Berk and Green (2004), which provides an equilibrium model of the mutual fund industry. The key predictions of the model:

1. **Fund managers have skill** (positive gross alpha on the first dollar managed)
2. **Decreasing returns to scale:** the more capital a manager actively manages, the lower the alpha per dollar
3. **Rational investors:** capital flows to funds with positive expected alpha until expected net alpha is driven to zero
4. **In equilibrium:** every fund has zero net alpha, but managers capture all economic rents through fees

This model has three testable predictions that the papers address:

Table 6: Testing the Berk-Green (2004) model

| Prediction                            | Fama & French (2010)                              | Berk & van Binsbergen (2015)                                       | Andrikogiannopoulou & Papakonstantinou (2019)                                            |
|---------------------------------------|---------------------------------------------------|--------------------------------------------------------------------|------------------------------------------------------------------------------------------|
| Zero net alpha in aggregate           | <b>Rejected:</b> aggregate net alpha $-1\%$ /year | Not directly tested; focus on value added                          | Not tested directly                                                                      |
| Investors identify and reward skill   | Not examined                                      | <b>Confirmed:</b> current compensation predicts future value added | Not examined                                                                             |
| Most individual funds have zero alpha | Consistent with bootstrap evidence                | Not the right question                                             | <b>Cannot be inferred from FDR estimates:</b> the estimator would report this regardless |

Fama and French (2010) explicitly state that their results “reject most of [Berk and Green’s] predictions about mutual fund returns.” The aggregate negative net alpha contradicts the model’s prediction that all managers have zero expected net alpha in equilibrium.

Berk and Binsbergen (2015) find results more consistent with Berk-Green: investors *do* appear to identify skilled managers (compensation predicts future value added), and the cross-sectional distribution of skill is persistent for up to ten years.

(andrikogiannopoulou2019?) add a third position that is neither support nor rejection: the FDR evidence most often cited *in favor* of Berk-Green is uninformative about it, so the model’s empirical status rests on the other two literatures rather than on the “75% zero alpha” statistic.

## 7. Key Results: Berk & van Binsbergen (2015)

### 7.1 Average Fund Adds Positive Value

Using the Vanguard benchmark, the **average mutual fund adds approximately \$3.2 million per year** (time-weighted mean: \$270,000/month in Y2000 dollars). The t-statistic is 5.74—strong statistical evidence against the no-skill hypothesis.

Using the FFC risk adjustment, the result is similar: \$250,000/month,  $t = 3.94$ .

However, the distribution is highly skewed:

Table 7: Distribution of value added across funds (Berk & van Binsbergen 2015, Table 3)

| Percentile                  | Value Added (Y2000 \$/month) |
|-----------------------------|------------------------------|
| 1st                         | −\$3.60M                     |
| 10th                        | −\$590K                      |
| 50th (median)               | −\$20K                       |
| 90th                        | +\$750K                      |
| 99th                        | +\$7.82M                     |
| % with positive value added | 43%                          |

Most funds destroy value; the *average* is positive because the minority of skilled managers manages most of the capital.

## 7.2 Persistence of Value Added

The most compelling evidence for skill is the **persistence** of value added for up to 10 years. Berk and Binsbergen (2015) construct a non-parametric test: do funds in the top decile (by skill ratio) continue to beat funds in the bottom decile in future periods?

The null hypothesis is Binomial( $n$ ,  $\frac{1}{2}$ ): under no skill, top decile funds beat bottom decile funds 50% of the time. The actual results:

- Top decile beats bottom decile at every horizon from 3 to 10 years
- The null is rejected at the 95% confidence level at most horizons
- The top decile controls approximately 25% of total AUM (investors reward skill with capital)

This persistence is far longer than anything documented using alpha measures, which Carhart (1997) and others have found to be largely unpredictable.

## 7.3 Investors Can Identify Skill

A surprising finding: **current managerial compensation predicts future value added better than past value added does**. This means investors are not “dumb money” chasing past returns—they appear to genuinely infer managerial skill and allocate capital accordingly.

Managerial compensation = AUM  $\times$  fee rate. Since fee rates vary little across funds, it is primarily investors’ capital allocation decisions that determine compensation. The finding implies that the flow-performance relationship reflects rational information processing, consistent with Berk and Green (2004).

## 7.4 Where Does the Value Go?

Value added can be decomposed into:

$$V_{i,t} = \underbrace{q_{i,t-1} \cdot f_{i,t-1}}_{\text{fees (compensation)}} + \underbrace{q_{i,t-1} \cdot \varepsilon_{i,t}}_{\text{returns to investors}}$$

Berk and Binsbergen (2015) find that **the net alpha to investors is approximately zero**—managers capture essentially all economic rents through fees. Investors receive no abnormal return from investing in skilled funds because competition drives net alpha to zero. This is the central prediction of Berk and Green (2004), and it is confirmed by the data.

Furthermore, roughly **half of total value added comes from diversification services** (the value Vanguard itself charges for) and half from active stock picking and market timing.

---

## 8. Reconciling the Three Papers

A reader might be confused: Fama & French say skill is rare; Berk & van Binsbergen say the average fund is skilled; Andrikogiannopoulou & Papakonstantinou say we may not be able to tell. Who is right?

All three are right—they are answering different questions:

Table 8: Reconciling the three papers

| Dimension                             | Fama & French (2010)                  | Berk & van Binsbergen (2015)                  | Andrikogiannopoulou & Papakonstantinou (2019) |
|---------------------------------------|---------------------------------------|-----------------------------------------------|-----------------------------------------------|
| Skill exists?                         | Modest evidence (gross returns only)  | Yes—average fund adds \$3.2M/year             | Unanswerable from FDR estimates               |
| Do investors benefit?                 | No—net returns are negative           | No—net alpha = 0 (rents captured by managers) | Not addressed                                 |
| Can investors identify skilled funds? | Not examined                          | Yes—compensation predicts future value added  | Not addressed                                 |
| Right benchmark?                      | FFC factors                           | Vanguard index funds (preferred)              | Not addressed                                 |
| Can the <i>test</i> detect skill?     | Bootstrap has some power in the tails | Value added has far better signal-to-noise    | <b>No</b> —FDR misses 65% of $\pm 2\%$ alphas |

The substantive resolution between the first two:

**Skill exists and is measurable in dollars. But competition among investors drives the net return to investors to zero. Skilled managers capture all rents through fees and fund size. The *dollar* evidence of skill is clear; the *return* evidence is weak because rational flows eliminate the return opportunity.**

And the methodological caution the third paper adds:

**Before interpreting any “no skill” finding, ask what the test would have reported had skill been present. If the answer is “the same thing,” the finding is about the test.**

---

## 9. Implications for Practice

### 9.1 For Performance Evaluation

1. **Do not rely on alpha alone.** The cross-section of alpha estimates is dominated by noise. Only the extreme tails (99th percentile and beyond) provide any reliable signal of genuine skill after accounting for multiple comparisons.
2. **Consider fund size.** Berk & van Binsbergen demonstrate that small funds can have high alpha but low value added. A large fund with modest alpha may be adding more value in dollar terms.
3. **Use the right benchmark.** FFC factors contain hindsight bias. Traded alternatives (passive index funds) are a more appropriate benchmark for evaluating whether a manager beats the investor’s actual opportunity set.

### 9.2 For Investors

4. **The expected net return from active management is approximately zero or slightly negative.** Fama & French’s equilibrium accounting tells us this must be true on average. The zero or negative net alpha is not a measurement artifact.
5. **Identifying skilled managers is difficult but not impossible.** Berk & van Binsbergen show that investors in aggregate appear to do so (compensation predicts performance). But individual investors face a severe information disadvantage relative to the institutional investors who dominate capital flows.

6. **Passive investing dominates for most investors.** Even if skill exists and is persistent, skilled managers capture their own rents. Net alpha to investors is approximately zero for the best funds—and negative for most funds.

### 9.3 For Research

7. **Bootstrap simulation is the gold standard** for detecting skill in cross-sections of fund returns. Standard t-tests are inadequate due to multiple comparisons and cross-fund correlation.
  8. **Statistical power is severely limited** with typical mutual fund sample sizes. A fund needs decades of monthly data to reliably distinguish a 2% annual alpha from noise. For shorter-lived funds, inference requires simulation-based methods.
  9. **Validate your estimator on a realistic DGP, not a convenient one** (andrikogiannopoulou2019?). BSW's validation simulation was internally correct; it simply assumed alphas far larger and panels far longer than the data contain. Before adopting any estimator, simulate it under parameters calibrated to the data you actually have.
  10. **Adding observations in the cross-section does not fix a power problem.** Power depends on  $T$  per unit, not  $N$ . This is worth internalizing: it applies to panel settings throughout empirical finance, not just fund performance.
- 

## 10. Empirical Implementation

In this section we implement a simplified version of the Fama and French (2010) bootstrap procedure. We use the same 10 US ETFs and factor data from Lecture 1 to illustrate the methodology. Note that this exercise is pedagogical—the full Fama and French (2010) analysis uses thousands of actively managed mutual funds.

### Setup

```
library(tidyverse)
library(tidyquant)
library(frenchdata)
library(broom)
library(gt)
library(scales)
```

```
set.seed(42) # reproducibility
```

## Load Data from Lecture 1

```
# --- ETF list and labels ---
etf_tickers <- c("SPY", "QQQ", "IWM", "VTV", "VUG", "ARKK", "XLE", "XLF", "VNQ", "IWD")
etf_labels <- tibble(
  symbol    = etf_tickers,
  name      = c("S&P 500", "Nasdaq-100", "Russell 2000", "Vanguard Value",
                "Vanguard Growth", "ARK Innovation", "Energy Sector",
                "Financials Sector", "Real Estate (REIT)", "Russell 1000 Value"),
  category  = c("Broad market", "Tech/Growth", "Small cap", "Large value",
                "Large growth", "Active growth", "Sector", "Sector", "REIT", "Large value")
)

# --- Download prices ---
start_date <- "2015-01-01"
end_date   <- "2024-12-31"

prices_raw <- tq_get(etf_tickers, from = start_date, to = end_date,
                     get = "stock.prices")

# --- Monthly returns ---
returns_monthly <- prices_raw |>
  group_by(symbol) |>
  tq_transmute(select = adjusted, mutate_fun = periodReturn,
               period = "monthly", type = "arithmetic",
               col_rename = "ret") |>
  ungroup() |>
  left_join(etf_labels, by = "symbol")

# --- French factors ---
ff3_raw <- download_french_data("Fama/French 3 Factors")
mom_raw <- download_french_data("Momentum Factor (Mom)")

ff3_monthly <- ff3_raw$subsets$data[[1]] |>
  mutate(date = lubridate::ym(date),
         across(c(`Mkt-RF`, SMB, HML, RF), ~ . / 100)) |>
  rename(RMRF = `Mkt-RF`) |>
  filter(date >= as.Date(start_date), date <= as.Date(end_date))
```

```

mom_monthly <- mom_raw$subsets$data[[1]] |>
  mutate(date = lubridate::ym(date), Mom = Mom / 100) |>
  rename(UMD = Mom) |>
  filter(date >= as.Date(start_date), date <= as.Date(end_date))

factors <- left_join(ff3_monthly, mom_monthly, by = "date")

# --- Merge and compute excess returns ---
data_merged <- returns_monthly |>
  mutate(ym = lubridate::floor_date(date, "month")) |>
  left_join(factors |> mutate(ym = lubridate::floor_date(date, "month")) |>
    select(ym, RMRF, SMB, HML, UMD, RF),
    by = "ym") |>
  mutate(excess_ret = ret - RF) |>
  filter(!is.na(RMRF))

data_merged |>
  select(symbol, ym, excess_ret, RMRF, SMB, HML, UMD) |>
  slice_head(n = 5)

```

```

# A tibble: 5 x 7
  symbol ym      excess_ret  RMRF  SMB  HML  UMD
  <chr> <date>      <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
1 SPY   2015-01-01   -0.0291 -0.031 -0.0061 -0.0348  0.0372
2 SPY   2015-02-01    0.0562  0.0613  0.0064 -0.0175 -0.0287
3 SPY   2015-03-01   -0.0157 -0.0111  0.0306 -0.0038  0.0272
4 SPY   2015-04-01    0.00983  0.0059 -0.0298  0.0181 -0.0726
5 SPY   2015-05-01    0.0129  0.0137  0.0096 -0.011  0.0576

```

## Step 1 — Estimate Baseline Carhart Alphas

```

# Estimate four-factor model for each ETF
carhart_base <- data_merged |>
  group_by(symbol, name) |>
  do(tidy(lm(excess_ret ~ RMRF + SMB + HML + UMD, data = .))) |>
  filter(term %in% c("(Intercept)", "RMRF", "SMB", "HML", "UMD")) |>
  mutate(term = recode(term,
    "(Intercept)" = "alpha", "RMRF" = "b_mkt",
    "SMB" = "b_smb", "HML" = "b_hml", "UMD" = "b_umd"
  ))

```

```

)) |>
pivot_wider(names_from = term,
             values_from = c(estimate, std.error, statistic, p.value),
             names_glue = "{term}_{.value}") |>
mutate(alpha_ann = alpha_estimate * 12) |>
ungroup()

# Extract t-statistics for alpha
alpha_t_actual <- carhart_base |>
  select(symbol, name, alpha_estimate, alpha_statistic, alpha_ann)

alpha_t_actual |>
  arrange(desc(alpha_statistic)) |>
  gt() |>
  fmt_number(columns = c(alpha_estimate, alpha_statistic, alpha_ann), decimals = 3) |>
  cols_label(
    symbol      = "Ticker",
    name        = "ETF",
    alpha_estimate = "Monthly ",
    alpha_statistic = "t()",
    alpha_ann    = "Ann.  "
  ) |>
  tab_header(
    title      = "Baseline Carhart Alphas",
    subtitle   = "These are the 'actual' t-statistics that bootstrap will benchmark against"
  ) |>
  tab_style(
    style      = cell_fill(color = "#e8f5e9"),
    locations  = cells_body(rows = alpha_statistic > 2)
  ) |>
  tab_style(
    style      = cell_fill(color = "#ffebee"),
    locations  = cells_body(rows = alpha_statistic < -2)
  )

```

## Step 2 — Impose Zero Alpha (Demean Returns)

```

# Subtract estimated alpha from each fund's excess returns
# This creates the population from which we bootstrap
data_zero_alpha <- data_merged |>

```

## Baseline Carhart Alphas

These are the 'actual' t-statistics that bootstrap will benchmark against

| Ticker | ETF                | Monthly | t( )   | Ann.   |
|--------|--------------------|---------|--------|--------|
| QQQ    | Nasdaq-100         | 0.003   | 1.980  | 0.033  |
| VUG    | Vanguard Growth    | 0.001   | 1.117  | 0.012  |
| SPY    | S&P 500            | 0.000   | 0.214  | 0.001  |
| XLF    | Financials Sector  | 0.000   | -0.088 | -0.002 |
| ARKK   | ARK Innovation     | -0.001  | -0.245 | -0.014 |
| XLE    | Energy Sector      | -0.002  | -0.424 | -0.024 |
| VTV    | Vanguard Value     | -0.001  | -0.768 | -0.009 |
| VNQ    | Real Estate (REIT) | -0.004  | -1.306 | -0.052 |
| IWD    | Russell 1000 Value | -0.002  | -2.327 | -0.025 |
| IWM    | Russell 2000       | -0.002  | -3.049 | -0.024 |

```
left_join(carhart_base |> select(symbol, alpha_estimate), by = "symbol") |>
mutate(excess_ret_zero = excess_ret - alpha_estimate)

# Quick check: every fund's alpha on demeaned returns should be ~0.
# Report the worst case across funds, not an average - an average could
# hide two large errors of opposite sign.
data_zero_alpha |>
  group_by(symbol) |>
  do(tidy(lm(excess_ret_zero ~ RMRF + SMB + HML + UMD, data = .))) |>
  filter(term == "(Intercept)") |>
  ungroup() |>
  summarise(max_abs_intercept = max(abs(estimate))) |>
  pull(max_abs_intercept)    # should be ~1e-18
```

```
[1] 6.532995e-18
```

### Step 3 — Bootstrap Simulation

We draw 1,000 random samples of months (with replacement), applying the **same draw to all ETFs** to preserve cross-correlation. For each draw, we estimate the four-factor model and record the t-statistic for alpha.

The implementation below arranges the data as two matrices once, then resamples *rows*. This is worth doing carefully, because the structure of the code is the methodological point of Fama and French (2010): a single index vector `idx` is applied to the factor matrix and the fund matrix

simultaneously, so every fund sees the same calendar months in the same order. Independent per-fund sampling — the Kosowski et al. (2006) approach — would require drawing a separate `idx` inside the fund loop, which is exactly the one-line change that destroys the cross-correlation structure.

```
# Funds as columns, months as rows. Keep only months where every fund
# has data, so the joint draw is well defined.
wide_rets <- data_zero_alpha |>
  select(symbol, ym, excess_ret_zero) |>
  pivot_wider(names_from = symbol, values_from = excess_ret_zero) |>
  arrange(ym) |>
  drop_na()

wide_facs <- data_zero_alpha |>
  distinct(ym, RMRF, SMB, HML, UMD) |>
  arrange(ym) |>
  semi_join(wide_rets, by = "ym")

stopifnot(identical(wide_facs$ym, wide_rets$ym))

fund_cols <- setdiff(names(wide_rets), "ym")

Y <- as.matrix(wide_rets[, fund_cols])          # T x N fund returns
X <- cbind(`(Intercept)` = 1,
          as.matrix(wide_facs[, c("RMRF", "SMB", "HML", "UMD")])) # T x 5

cat("Bootstrap panel:", nrow(Y), "months x", ncol(Y), "funds\n")
```

Bootstrap panel: 120 months x 10 funds

```
# t(alpha) for ALL funds at once, given a design matrix and return matrix.
# Vectorised across funds: one solve() per simulation rather than one
# lm() per fund. Verified to reproduce summary(lm(...)) to ~1e-12.
alpha_tstats <- function(X, Y) {
  XtXinv <- solve(crossprod(X))
  B      <- XtXinv %*% crossprod(X, Y)          # 5 x N coefficients
  resid  <- Y - X %*% B
  sigma2 <- colSums(resid^2) / (nrow(X) - ncol(X))
  B[1, ] / sqrt(XtXinv[1, 1] * sigma2)         # intercept t-stats
}
```

```

n_sim <- 1000          # full paper uses 10,000; 1,000 runs in about a second
n_obs <- nrow(Y)

sim_t_stats <- matrix(NA_real_, nrow = n_sim, ncol = ncol(Y),
                      dimnames = list(NULL, colnames(Y)))

for (b in seq_len(n_sim)) {
  # ONE index vector, applied to factors and every fund alike.
  # This is what "joint sampling" means in Fama-French (2010).
  idx <- sample.int(n_obs, size = n_obs, replace = TRUE)
  sim_t_stats[b, ] <- alpha_tstats(X[idx, , drop = FALSE],
                                   Y[idx, , drop = FALSE])
}

sim_t_df <- as_tibble(sim_t_stats) |>
  mutate(sim = row_number()) |>
  pivot_longer(-sim, names_to = "symbol", values_to = "t_sim")

```

#### Step 4 — Compare Actual vs. Simulated t-Statistics

```

# Compute percentile of actual t-stat in simulated distribution
pct_actual <- sim_t_df |>
  left_join(alpha_t_actual |> select(symbol, t_actual = alpha_statistic),
            by = "symbol") |>
  group_by(symbol) |>
  summarise(
    pct_beat = mean(t_sim < t_actual, na.rm = TRUE), # fraction of sims below actual
    t_actual = first(t_actual),
    .groups = "drop"
  ) |>
  left_join(etf_labels, by = "symbol")

# Plot: density of simulated t-stats + actual t-stat as vertical line
# Focus on a representative subset of 4 ETFs
selected <- c("SPY", "ARKK", "IWM", "QQQ")

sim_t_df |>
  filter(symbol %in% selected) |>
  left_join(alpha_t_actual |> select(symbol, t_actual = alpha_statistic), by = "symbol") |>
  left_join(etf_labels, by = "symbol") |>

```

```

ggplot(aes(x = t_sim)) +
  geom_histogram(aes(y = after_stat(density)), bins = 50,
                 fill = "steelblue", alpha = 0.6, color = "white") +
  geom_vline(aes(xintercept = t_actual), color = "tomato",
             linewidth = 1.2, linetype = "dashed") +
  geom_vline(xintercept = c(-2, 2), color = "gray40",
             linewidth = 0.5, linetype = "dotted") +
  facet_wrap(~name, scales = "free_y", nrow = 2) +
  labs(
    title      = "Bootstrap Distribution of t() Under Null of Zero True Alpha",
    subtitle   = "Dashed red line = actual t(); dotted lines =  $\pm 2$  conventional threshold",
    x          = "Simulated t()",
    y          = "Density"
  ) +
  theme_minimal()

```

## Bootstrap Distribution of $t(\cdot)$ Under Null of Zero True Alpha

Dashed red line = actual  $t(\cdot)$ ; dotted lines =  $\pm 2$  conventional threshold

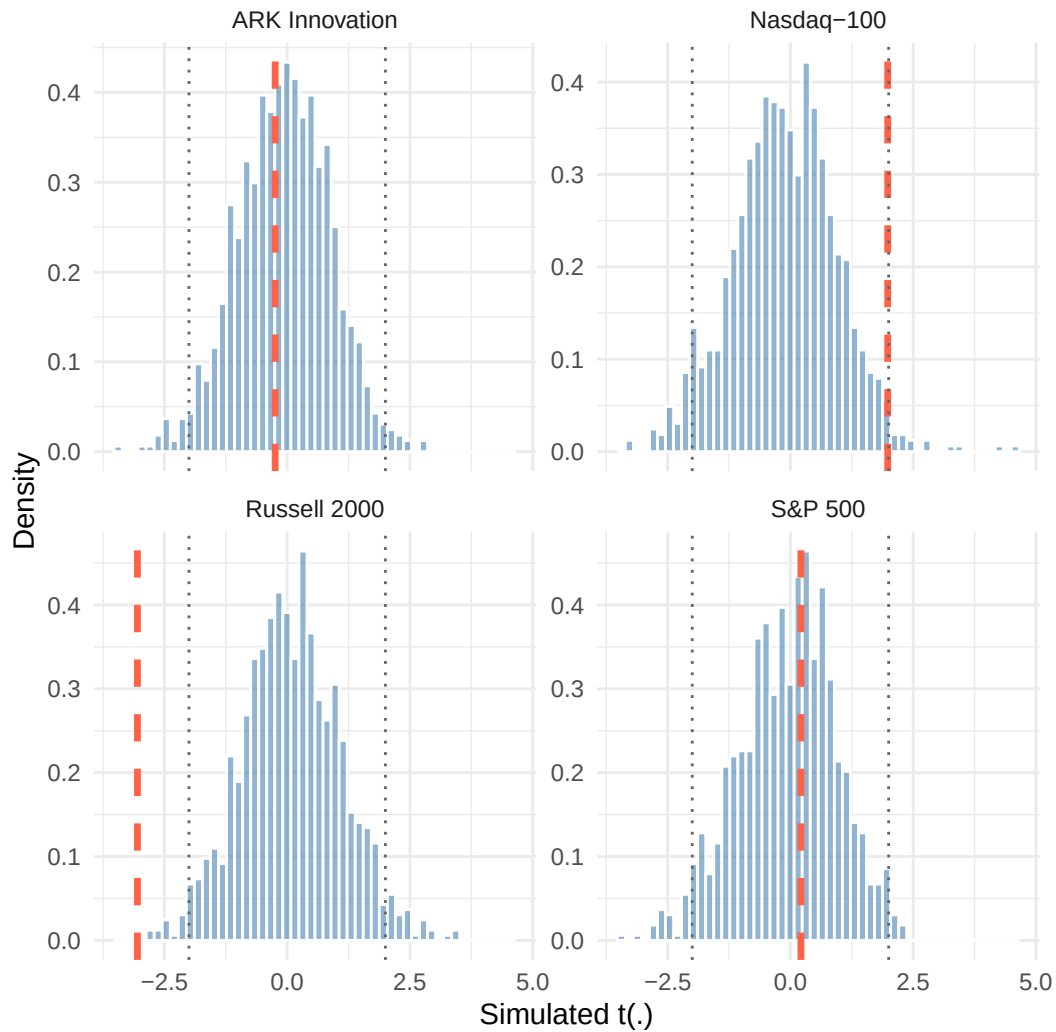


Figure 1: Bootstrap distribution of  $t(\cdot)$  vs. actual  $t(\cdot)$  for selected ETFs

### Step 5 — Summarize Luck vs. Skill Evidence

```
pct_actual |>
  arrange(desc(pct_beat)) |>
  select(name, t_actual, pct_beat) |>
  mutate(
```

```

interpretation = case_when(
  pct_beat > 0.90 ~ "Right-tail evidence of skill",
  pct_beat < 0.10 ~ "Left-tail evidence of inferior performance",
  TRUE          ~ "Consistent with luck"
)
) |>
gt() |>
fmt_number(columns = t_actual, decimals = 3) |>
# Format pct_beat for display but keep it NUMERIC in the underlying
# data, so the conditional fills below compare numbers to numbers.
# (Overwriting it with percent() would make it a character vector, and
# "95.0%" > 0.90 is a silent string comparison, not a numeric test.)
fmt_percent(columns = pct_beat, decimals = 1) |>
cols_label(
  name          = "ETF",
  t_actual      = "Actual t()",
  pct_beat      = "% Simulations Below",
  interpretation = "Inference"
) |>
tab_header(
  title      = "Bootstrap Inference: Luck vs. Skill",
  subtitle   = paste0("Based on ", scales::comma(n_sim),
    " bootstrap simulations; joint monthly resampling")
) |>
tab_footnote(
  footnote    = "Percent of simulations that produce t() below the actual value. Values >90%
  locations   = cells_column_labels(columns = pct_beat)
) |>
tab_style(
  style       = cell_fill(color = "#e3f2fd"),
  locations   = cells_body(rows = pct_beat > 0.90)
) |>
tab_style(
  style       = cell_fill(color = "#fce4ec"),
  locations   = cells_body(rows = pct_beat < 0.10)
)

```

### **i** Note

**Interpretation reminder:** Our ETF sample differs from Fama and French (2010) in an important way. ETFs are passively managed (except ARKK), so their “alpha” mainly

Table 9

## Bootstrap Inference: Luck vs. Skill

Based on 1,000 bootstrap simulations; joint monthly resampling

| ETF                | Actual $t()$ | % Simulations Below <sup>1</sup> | Inference                                  |
|--------------------|--------------|----------------------------------|--------------------------------------------|
| Nasdaq-100         | 1.980        | 98.2%                            | Right-tail evidence of skill               |
| Vanguard Growth    | 1.117        | 88.5%                            | Consistent with luck                       |
| S&P 500            | 0.214        | 58.8%                            | Consistent with luck                       |
| Financials Sector  | -0.088       | 46.1%                            | Consistent with luck                       |
| ARK Innovation     | -0.245       | 41.6%                            | Consistent with luck                       |
| Energy Sector      | -0.424       | 32.5%                            | Consistent with luck                       |
| Vanguard Value     | -0.768       | 21.4%                            | Consistent with luck                       |
| Real Estate (REIT) | -1.306       | 9.4%                             | Left-tail evidence of inferior performance |
| Russell 1000 Value | -2.327       | 0.4%                             | Left-tail evidence of inferior performance |
| Russell 2000       | -3.049       | 0.0%                             | Left-tail evidence of inferior performance |

<sup>1</sup>Percent of simulations that produce  $t()$  below the actual value. Values  $>90\%$  suggest skill;  $<10\%$  suggest inferior performance.

reflects factor model misspecification or expense ratios, not stock-picking skill. The bootstrap exercise here is purely pedagogical. For genuine luck-vs.-skill inference, the procedure should be applied to a large cross-section of *actively managed* open-end mutual funds.

### Simulating the Value-Added Measure

To illustrate the Berk and Binsbergen (2015) approach, we compute a simplified value-added measure for each ETF. We proxy AUM using the ETF's market price  $\times$  shares outstanding—not directly available via `tidyquant`, so we demonstrate with hypothetical but plausible AUM data.

```
# Use hypothetical (approximate) average AUM for each ETF (in billions USD)
# These are rough approximations for pedagogical illustration
aum_proxy <- tibble(
  symbol = etf_tickers,
  avg_aum_bn = c(500, 250, 70, 100, 120, 10, 25, 40, 65, 55) # approximate 2015–2024 avg AUM
)

# Compute value added: AUM  $\times$  gross excess return above RMRF benchmark
```

```

# Here we use RMRF as the naive one-factor benchmark for illustration
value_added <- data_merged |>
  left_join(aum_proxy, by = "symbol") |>
  mutate(
    # Simplified value added: AUM * (fund return - market return benchmark)
    # In practice, BvB use the Vanguard multi-fund benchmark
    gross_excess_vs_mkt = excess_ret - RMRF,
    value_added_mn = avg_aum_bn * 1000 * gross_excess_vs_mkt # in $millions
  ) |>
  group_by(symbol, name) |>
  summarise(
    mean_va_ann = mean(value_added_mn, na.rm = TRUE) * 12, # annualized avg
    mean_alpha = mean(excess_ret - RMRF, na.rm = TRUE) * 12,
    avg_aum_bn = first(avg_aum_bn),
    .groups = "drop"
  ) |>
  arrange(desc(mean_va_ann))

value_added |>
  gt() |>
  fmt_number(columns = c(mean_va_ann, avg_aum_bn), decimals = 1) |>
  fmt_percent(columns = mean_alpha, decimals = 2) |> # it is a return, not a count
  cols_label(
    name = "ETF",
    mean_va_ann = "Ann. Value Added ($M)",
    mean_alpha = "Ann. Alpha (vs. Market)",
    avg_aum_bn = "Approx. AUM ($B)"
  ) |>
  tab_header(
    title = "Simplified Value-Added Illustration",
    subtitle = "Pedagogical example: AUM * (fund return - market return)"
  ) |>
  tab_footnote(
    footnote = "AUM values are approximate. True BvB value added uses Vanguard fund benchmark",
    locations = cells_title()
  )

```

## Illustrating the Power Problem

The exercises above take the estimator's accuracy for granted. This one checks it. Following (andrikogiannopoulou2019?), we build a world where we *know* the truth, run the FDR

# Simplified Value-Added Illustration<sup>1</sup>

Pedagogical example:  $\text{AUM} \times (\text{fund return} - \text{market return})$ <sup>1</sup>

| symbol | ETF                | Ann. Value Added (\$M) | Ann. Alpha (vs. Market) | Approx. AUM (\$B) |
|--------|--------------------|------------------------|-------------------------|-------------------|
| QQQ    | Nasdaq-100         | 13,124.9               | 5.25%                   | 250               |
| VUG    | Vanguard Growth    | 3,531.5                | 2.94%                   | 120               |
| ARKK   | ARK Innovation     | 452.5                  | 4.53%                   | 10                |
| SPY    | S&P 500            | 187.9                  | 0.04%                   | 500               |
| XLF    | Financials Sector  | -298.9                 | -0.75%                  | 40                |
| XLE    | Energy Sector      | -1,087.0               | -4.35%                  | 20                |
| IWD    | Russell 1000 Value | -2,329.8               | -4.24%                  | 50                |
| IWM    | Russell 2000       | -2,638.1               | -3.77%                  | 70                |
| VTI    | Vanguard Value     | -2,793.6               | -2.79%                  | 100               |
| VNQ    | Real Estate (REIT) | -4,602.6               | -7.08%                  | 60                |

<sup>1</sup>AUM values are approximate. True BvB value added uses Vanguard fund benchmarks and precise historical AUM.

estimator on it, and see what it reports.

The setup is deliberately generous to the FDR. We assume alphas are estimated with no factor-model misspecification and no cross-sectional correlation in residuals—both of which, in the paper, make matters worse.

```
# Parameters follow Andrikogiannopoulou & Papakonstantinou (2019):
# monthly residual SD of 0.021 in their baseline calibration, lambda = 0.5.
sigma_e <- 0.021
lambda <- 0.5

# Simulate one cross-section of funds with KNOWN alphas, then apply the
# FDR estimator of equation (2) and return its estimate of pi^0.
fdr_pi0 <- function(n_funds, T_obs, pi0, pi_neg, pi_pos, alpha_bar) {
  se <- sigma_e / sqrt(T_obs) # SE of the monthly alpha estimate
  a_m <- alpha_bar / 12 # annual alpha -> monthly
  truth <- sample(c(0, -a_m, a_m), n_funds, replace = TRUE,
                 prob = c(pi0, pi_neg, pi_pos))
  ahat <- rnorm(n_funds, mean = truth, sd = se)
  pval <- 2 * (1 - pnorm(abs(ahat) / se))
  mean(pval > lambda) / (1 - lambda) # equation (2)
}
```

```

set.seed(2019)

fdr_grid <- expand_grid(
  alpha_bar = c(0.010, 0.020, 0.035),
  T_obs      = c(180, 384)
) |>
  mutate(
    pi0_hat = map2_dbl(alpha_bar, T_obs, \(a, t)
      mean(replicate(200, fdr_pi0(2000, t, 0.75, 0.23, 0.02, a)))),
    bias = pi0_hat - 0.75
  )

fdr_grid |>
  mutate(T_lab = if_else(T_obs == 180,
    "T = 180 (data median/mean)",
    "T = 384 (BSW simulation)")) |>
  select(T_lab, alpha_bar, pi0_hat, bias) |>
  arrange(T_lab, alpha_bar) |>
  gt(groupname_col = "T_lab") |>
  fmt_percent(columns = c(alpha_bar, pi0_hat, bias), decimals = 1) |>
  cols_label(
    alpha_bar = "True | | of nonzero funds (ann.)",
    pi0_hat    = "FDR estimate of ",
    bias        = "Bias (truth = 75%)"
  ) |>
  tab_header(
    title      = "The FDR Estimator Under Known Truth",
    subtitle    = "75% zero-alpha, 23% negative, 2% positive - BSW's own mix"
  ) |>
  tab_style(
    style       = cell_fill(color = "#ffebee"),
    locations   = cells_body(rows = bias > 0.05)
  )

```

Read the bottom row of each block first. With  $T = 384$  and alphas of  $\pm 3.5\%$  per year—BSW’s simulation design—the FDR recovers a  $\pi^0$  of about 76% against a truth of 75%. The estimator looks excellent, which is exactly what BSW reported.

Now read upward. Holding everything else fixed and shrinking the true alpha to  $\pm 1\%$  per year, the same estimator reports that roughly 94% of funds have zero alpha. Nothing changed about the method; only the world it was applied to.

Table 10: FDR estimate of the zero-alpha proportion when the truth is 75%

## The FDR Estimator Under Known Truth

75% zero-alpha, 23% negative, 2% positive — BSW's own mix

| True    of nonzero funds (ann.) | FDR estimate of | Bias (truth = 75%) |
|---------------------------------|-----------------|--------------------|
| T = 180 (data median/mean)      |                 |                    |
| 1.0%                            | 97.0%           | 22.0%              |
| 2.0%                            | 90.4%           | 15.4%              |
| 3.5%                            | 80.5%           | 5.5%               |
| T = 384 (BSW simulation)        |                 |                    |
| 1.0%                            | 94.0%           | 19.0%              |
| 2.0%                            | 83.6%           | 8.6%               |
| 3.5%                            | 76.1%           | 1.1%               |

```

misclass <- function(alpha_bar, T_obs) {
  se <- sigma_e / sqrt(T_obs)
  # P(p-value > lambda) for a fund whose true alpha is alpha_bar,
  # scaled by 1/(1-lambda) because equation (2) extrapolates.
  p_above <- pnorm(qnorm(1 - lambda / 2) - (alpha_bar / 12) / se)
  pmin(1, p_above / (1 - lambda))
}

expand_grid(
  alpha_bar = seq(0.005, 0.05, by = 0.001),
  T_obs      = c(120, 180, 384)
) |>
  mutate(
    rate = misclass(alpha_bar, T_obs),
    T_lab = factor(T_obs, levels = c(120, 180, 384),
                  labels = c("T = 120", "T = 180 (realistic)", "T = 384"))
  ) |>
  ggplot(aes(alpha_bar, rate, color = T_lab)) +
  geom_line(linewidth = 0.8) +
  geom_vline(xintercept = c(0.01, 0.02), linetype = "dotted", color = "grey40") +
  annotate("text", x = 0.0205, y = 0.05, hjust = 0, size = 3,
          label = "±1% and ±2% per year") +
  scale_x_continuous(labels = percent_format()) +
  scale_y_continuous(labels = percent_format(), limits = c(0, 1)) +

```

```
labs(
  x = "True annual alpha of a nonzero-alpha fund",
  y = "Probability of being counted as zero-alpha",
  color = NULL,
  title = "Where the FDR loses skilled managers",
  subtitle = "Economically large alphas are invisible at realistic sample lengths"
) +
theme_minimal() +
theme(legend.position = "bottom")
```

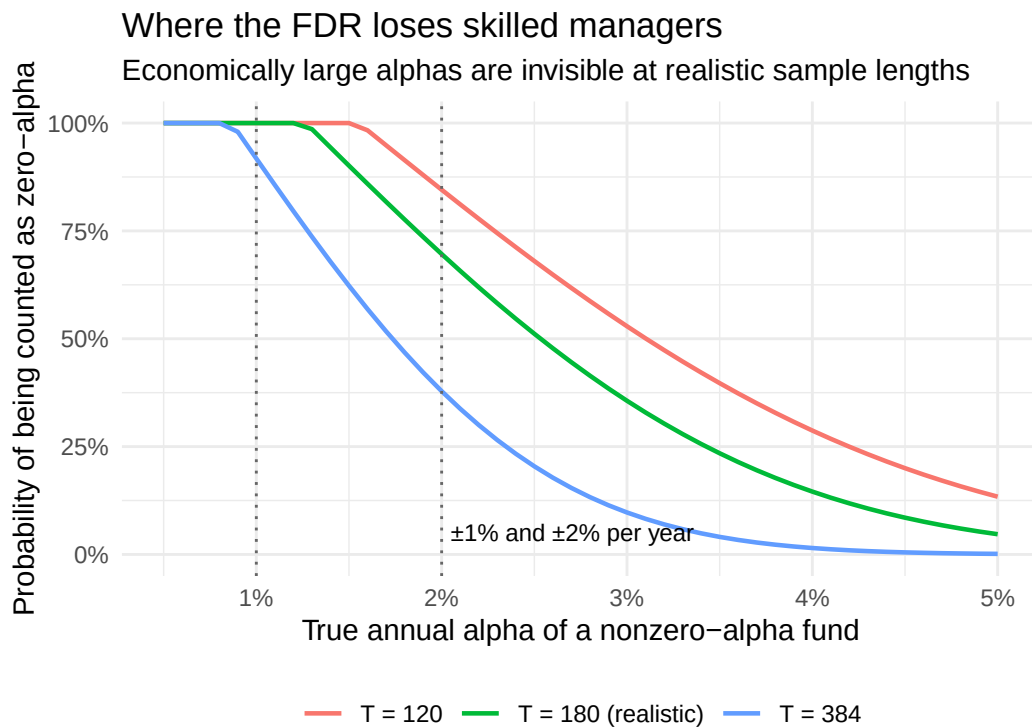


Figure 2: Share of genuinely nonzero-alpha funds that the FDR classifies as zero-alpha

#### ⚠ How closely does this match the paper?

Reasonably, but not exactly, and the gap is worth understanding.

Our simplified version draws  $\hat{\alpha}$  directly from a normal rather than estimating it from a factor regression, and it fixes  $\lambda = 0.5$  where (andrikogiannopoulou2019?) select  $\lambda$  optimally by bootstrap and average over 1,000 repetitions. It reproduces their central case closely—at  $T = 384$  with  $\pm 3.5\%$  alphas we get  $\hat{\pi}^0 \approx 76\%$  against their reported

76.3%—but our misclassification rates for small alphas run somewhat higher than their simulated 90% and 65% figures.

Treat the exact numbers here as illustrative and the paper's as authoritative. The mechanism, the sign of the bias, and its order of magnitude are what this exercise is for. The fact that a back-of-envelope reproduces the qualitative result is itself the point: the bias is not subtle, and anyone could have checked.

---

## 11. Summary Statistics and Comparison

```
# Combine Carhart results with bootstrap pct and AUM
final_tbl <- carhart_base |>
  select(name, alpha_ann, alpha_statistic, b_mkt_estimate) |>
  left_join(pct_actual |> select(name, pct_beat), by = "name") |>
  arrange(desc(alpha_ann))

final_tbl |>
  gt() |>
  fmt_number(columns = c(alpha_ann, alpha_statistic, b_mkt_estimate), decimals = 3) |>
  fmt_percent(columns = pct_beat, decimals = 1) |>
  cols_label(
    name           = "ETF",
    alpha_ann       = "Carhart (Ann.)",
    alpha_statistic = "t()",
    b_mkt_estimate  = "(MKT)",
    pct_beat        = "Bootstrap %ile"
  ) |>
  tab_header(
    title      = "Fund Performance Summary with Bootstrap Inference",
    subtitle   = "Bootstrap %ile = fraction of simulations with t() below actual"
  ) |>
  data_color(
    columns = alpha_ann,
    fn = scales::col_numeric(
      palette = c("tomato", "white", "steelblue"),
      domain  = NULL
    )
  )
)
```

Table 11

## Fund Performance Summary with Bootstrap Inference

Bootstrap %ile = fraction of simulations with  $t(\cdot)$  below actual

| ETF                | Carhart (Ann.) | $t(\cdot)$ | (MKT) | Bootstrap %ile |
|--------------------|----------------|------------|-------|----------------|
| Nasdaq-100         | 0.033          | 1.980      | 1.082 | 98.2%          |
| Vanguard Growth    | 0.012          | 1.117      | 1.074 | 88.5%          |
| S&P 500            | 0.001          | 0.214      | 0.981 | 58.8%          |
| Financials Sector  | -0.002         | -0.088     | 1.061 | 46.1%          |
| Vanguard Value     | -0.009         | -0.768     | 0.886 | 21.4%          |
| ARK Innovation     | -0.014         | -0.245     | 1.518 | 41.6%          |
| Russell 2000       | -0.024         | -3.049     | 1.032 | 0.0%           |
| Energy Sector      | -0.024         | -0.424     | 1.093 | 32.5%          |
| Russell 1000 Value | -0.025         | -2.327     | 0.915 | 0.4%           |
| Real Estate (REIT) | -0.052         | -1.306     | 0.857 | 9.4%           |

## 12. Key Takeaways

1. **Equilibrium accounting constrains aggregate performance** (Fama and French 2010). Before costs, aggregate active management must have zero alpha. After costs, it must be negative. This is arithmetic, not an empirical finding.
2. **The bootstrap simulation is the right tool** for separating luck from skill in large fund cross-sections. Standard t-tests fail due to multiple comparisons and cross-fund correlation. Joint resampling of months preserves the correlation structure of fund returns.
3. **Few funds cover costs on net returns** (Fama and French 2010). The cross-section of  $t(\cdot)$  for net returns is almost entirely to the left of the simulated null distribution. Skill is detectable only in the extreme right tail (97th percentile) and even then the evidence is mixed.
4. **Skill exists in gross returns**, with true alpha dispersion of roughly  $\pm 1.25\%$  per year. But this barely covers average expense ratios of 1%.
5. **The FDR estimate that “75% of funds have zero alpha” does not survive scrutiny** (andrikogiannopoulou2019?). At realistic sample lengths the methodology misclassifies 65% of funds with alphas of  $\pm 2\%$  per year, and 90% of those with  $\pm 1\%$ . The bias is upward in  $\pi^0$  by construction, because equation (2) extrapolates every failure to reject.

6. **Low power is not a caveat; it can invert the conclusion.** Under a data-generating process calibrated to Fama and French's own estimate of alpha dispersion—where only 1% of funds truly have zero alpha—the FDR reports 70% zero-alpha funds. An estimator that returns BSW's answer whether or not BSW's world obtains is not evidence for that world, and therefore not evidence for Berk and Green (2004) either.
7. **Power comes from  $T$ , not  $N$ .** Adding funds does not improve the FDR estimator at all; only lengthening each fund's return history does, and convergence is so slow that even 200 years of monthly data would be insufficient. This generalizes well beyond fund performance.
8. **Alpha is the wrong measure of skill** (Berk and Binsbergen 2015). Under standard neoclassical assumptions, neither net alpha (driven to zero by competition) nor gross alpha (equals the manager's chosen fee) measures skill. Value added—AUM times gross excess return—is the correct measure.
9. **The average fund adds \$3.2 million per year** in dollar terms (Berk and Binsbergen 2015). Skill is widespread but concentrated among large funds. Cross-sectional differences in skill persist for up to 10 years—far longer than persistence in alpha.
10. **Investors appear to identify skill:** managerial compensation ( $\text{AUM} \times \text{fee}$ ) predicts future value added. Capital flows to skilled managers, consistent with Berk and Green (2004).
11. **Skilled managers capture all rents:** net alpha to investors is approximately zero even for the best funds. There is no free lunch in active management—the benefit of skill accrues to the manager, not the investor.

---

## Discussion Questions

1. Fama and French (2010) find that the VW portfolio of active funds has a net alpha of approximately  $-1\%$  per year. Is this evidence that active managers *lack* skill, or evidence that *investors* rationally drive net alphas to zero? How would you distinguish between these two interpretations empirically?
2. Berk and Binsbergen (2015) argue that gross alpha is not a valid measure of skill because it equals the fee the manager charges. Does this imply that a manager who charges 2% in fees is more skilled than one who charges 0.5%? What does the theory actually predict about the relationship between fees and skill?
3. The Fama and French (2010) bootstrap preserves cross-correlation by jointly sampling months for all funds. Why is this important? What would happen to the simulation results if each fund's months were sampled independently?

4. Berk and Binsbergen (2015) use Vanguard index funds rather than FFC factors as the benchmark. Their Table 2 shows that HML and UMD have significant positive alpha relative to Vanguard funds. What does this imply about alpha estimates computed using FFC factors for the early part of the sample?
5. Suppose a fund manager has a true annual alpha of 2% and manages \$100 million. Under the Berk-Green (2004) model, what happens to the fund's alpha and AUM over time as investors observe the fund's performance? Is this story consistent with Berk and Binsbergen (2015)'s finding that value-added persistence lasts 10 years?
6. Fama and French (2010) find that the CAPM (single-factor) bootstrap produces much stronger apparent evidence of skill than the four-factor bootstrap. Why does adding size, value, and momentum factors to the benchmark reduce the apparent evidence of skill? What does this tell you about the source of "alpha" in a CAPM regression?
7. Write down the assumption behind equation (2) in words, then state precisely when it fails. (**andrikogiannopoulou2019?**) show that the resulting bias in  $\hat{\pi}^0$  is always *upward*. Explain why the bias cannot go the other way, and why the factor  $1/(1 - \lambda)$  makes it worse rather than better.
8. (**andrikogiannopoulou2019?**) find that increasing the number of funds  $N$  does not reduce the FDR estimator's bias, while increasing the number of observations per fund  $T$  does. Explain the asymmetry. Then consider: does the Fama and French (2010) bootstrap suffer from the same problem? Which features of the bootstrap make it more or less vulnerable than the FDR, and does it escape the low-power critique entirely?
9. Suppose you are refereeing a paper that applies the FDR methodology to a panel of 3,000 hedge funds with an average of 48 monthly return observations each. Using the reasoning in (**andrikogiannopoulou2019?**), what would you ask the authors to demonstrate before you would believe their estimated proportions? Be specific about the simulation you would request.

---

## References

- Berk, Jonathan B., and Jules H. van Binsbergen. 2015. "Measuring Skill in the Mutual Fund Industry." *Journal of Financial Economics* 118 (1): 1–20. <https://doi.org/10.1016/j.jfineco.2015.05.002>.
- Berk, Jonathan B., and Richard C. Green. 2004. "Mutual Fund Flows and Performance in Rational Markets." *Journal of Political Economy* 112 (6): 1269–95. <https://doi.org/10.1086/424739>.

- Carhart, Mark M. 1997. "On Persistence in Mutual Fund Performance." *Journal of Finance* 52 (1): 57–82.
- Fama, Eugene F., and Kenneth R. French. 2010. "Luck Versus Skill in the Cross-Section of Mutual Fund Returns." *Journal of Finance* 65 (5): 1915–47. <https://doi.org/10.1111/j.1540-6261.2010.01598.x>.
- Kosowski, Robert, Allan Timmermann, Russ Wermers, and Hal White. 2006. "Can Mutual Fund 'Stars' Really Pick Stocks? New Evidence from a Bootstrap Analysis." *Journal of Finance* 61 (6): 2551–95. <https://doi.org/10.1111/j.1540-6261.2006.01015.x>.
- Sharpe, William F. 1991. "The Arithmetic of Active Management." *Financial Analysts Journal* 47 (1): 7–9.
- Wermers, Russ. 2011. "Performance Measurement of Mutual Funds, Hedge Funds, and Institutional Accounts." *Annual Review of Financial Economics* 3: 537–74.