Why financial research is prone to false statistical discoveries

David H. Bailey
https://www.davidhbailey.com
Lawrence Berkeley National Lab (retired) and University of California, Davis
Marcos López de Prado
https://www.quantresearch.org
Cornell University and Abu Dhabi Investment Authority

This talk is available at:
http://www.davidhbailey.com/dhbtalks/dhb-risk-2022.pdf

January 25, 2022

Why financial research is prone to false statistical discoveries

- The chances of finding a truly profitable investment design or strategy is very low, due to intense competition.
- True findings are mostly short-lived, as a result of the non-stationary nature of most financial systems.
- Since it is rarely possible to rigorously test statistical findings with controlled experiments, it is often difficult to debunk a false claim.

One would hope that financial researchers would be particularly careful when conducting statistical inference. Sadly, the opposite appears more accurate.

- D. H. Bailey and M. Lopez de Prado, "How backtest overfitting in finance leads to false discoveries," *Significance*, Royal Statistical Society, Jan 2022, condensed from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3895330.

Reproducibility crises in biomedicine, psychology and economics

- In 2011, Bayer researchers reported that they were able to reproduce only 17 of 67 pharma studies.
- In 2012, Amgen researchers reported that they were able to reproduce only 6 of 53 cancer studies.
- In August 2015, the Reproduciblity Project in Virginia reported that they were able to reproduce only 39 of 100 psychology studies.
- In September 2015, the U.S. Federal Reserve was able to reproduce only 29 of 67 economics studies.
- In an updated 2018 study, the Reproducibility Project was able to replicate only 14 out of 28 psychology experimental studies.

Reproducibility Project staff

Credit: NY Times

Reproducibility crisis in high-performance computing

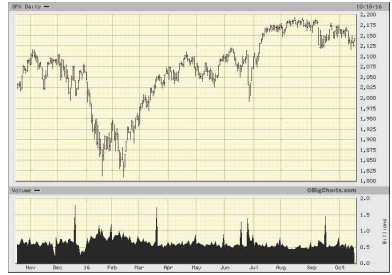
- In 1990, highly parallel computer arrays emerged for high-performance scientific computing.
- Some in the research community hyped the new technology with inflated claims of performance.
- I personally published two articles warning of such distortions, but such warnings were mostly ignored.
- The field's credibility declined; dozens of parallel computer firms went bankrupt. At least one CEO personally blamed me for his firm's demise.
- After improved benchmarks and performance standards, credibility was restored. Now virtually all scientific supercomputers feature a massively parallel design.

Japan's Fugaku supercomputer

Lesson: If a field has sloppy standards, with a public perception that does not match reality, it usually ends badly.

Is there a similar crisis in finance?

- Only a small fraction of investment funds beat the corresponding market averages over a 10-year window.
- Only a small fraction of financial forecasters do better than chance when their records are checked.
- Financial news is replete with pseudomathematical jargon: "Fibonacci ratios," "cycles," "waves," "golden ratios," "parabolic SARs," "pivot points," "rising wedges," "shoulders," etc.
- Many analysts (and brokerages) recommend discredited statistical methods (e.g., "technical analysis").
- Even in academic literature, few authors disclose the full extent of computer searches used in developing or tuning a model, and few if any journals require authors to disclose such information.



What should responsible finance researchers do?

Ensure that their own published research is mathematically and statistically sound.

Email exchange between DHB and a finance colleague

Email from DHB to finance colleague, 10 June 2013:

One thing that has always puzzled me about the financial world is the following sort of thing: [several specific examples cited]. Excuse me for being "dumb," but this sort of thing seems to me to be outright nonsense. ... [For example,] when people like those above say that they "know" where the stock market is heading, this cannot have any scientific basis. ...

So why doesn't somebody blow this whistle on this sort of thing? Am I missing something?

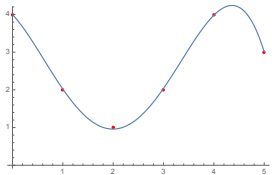
Response from finance colleague to DHB, 17 June 2013:

It is not a dumb question at all. ... I completely agree with your assessment. The amount of nonsense ... is incredible.

Backtest overfitting: Statistical overfitting of historical market data

Backtest overfitting is arguably the financial field's version of "p-hacking," namely the regrettable practice of only citing results from a subset of data or tests. Examples:

- Employing a theoretical model that inherently possesses a higher level of complexity than the backtest data.
- Using a computer to try millions or billions of variations of a model or strategy on historical market data, and then only presenting results from the variation that works best.
- Using a computer to explore millions or billions of weighting factors for a mutual fund, then only marketing the one that scores best on a backtest.

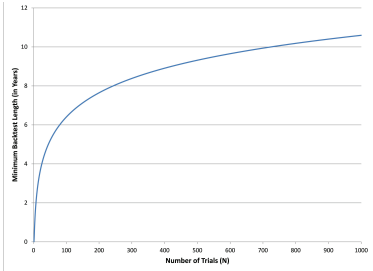


Fitting six data points almost perfectly with a fourth-degree function.

When a computer can analyze many variations of a model, fund or strategy on a fixed dataset, it is virtually certain that the "optimal" selection will be statistically overfit.

How easy is it to overfit a backtest? Very!

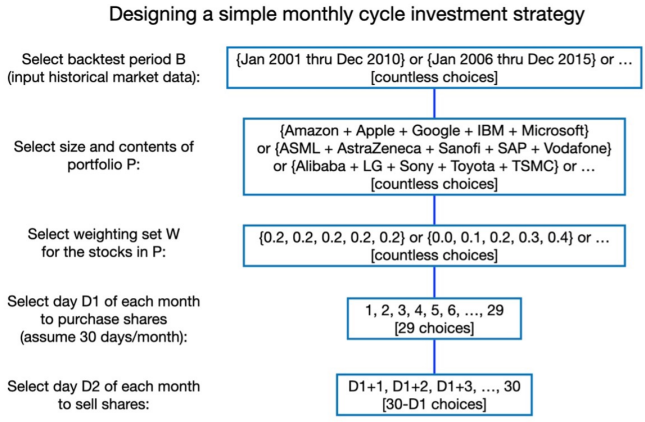
- If only 2 years of daily market data are available, then no more than 7 variations should be tried.
- If only 5 years of daily market data are available, then no more than 45 variations should be tried.



- D. H. Bailey, J. M. Borwein, M. Lopez de Prado and J. Zhu, "Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance," *Notices of the American Mathematical Society*, May 2014, 458–471, https://www.ams.org/notices/201405/rnoti-p458.pdf.

Parameter variations in a simple investment strategy

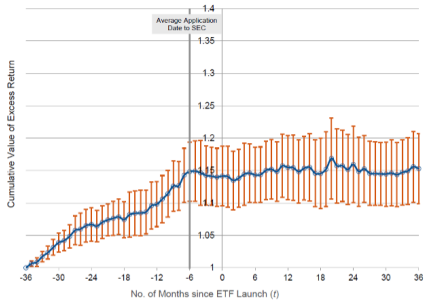
- Consider a very simple strategy, wherein one buys a set of stocks of stocks on one day of the month then sells them on another day.
- Even with this simple example, there are thousands of choices and parameter variations.
- There are 435 choices just for the start and end dates.



Searching by computer over all of the possible parameters for an "optimal" strategy is virtually certain to produce an overfit result.

Proliferation of new exchange-traded stock funds (ETFs)

- A 2015 study computed the performance of all ETFs that were launched in the U.S. market from 1993 to 2014.
- Researchers found that the investment strategies underlying those ETFs delivered average annual excess returns of approx. 5% prior to their launch (i.e., by using backtests).
- After the launch, annual excess returns dropped to approx. 0%.



Such disappointing behavior is entirely consistent with a design process that involves extensive computer exploration of parameters and weights, but selecting only the "optimal" set for an index fund subsequently fielded in the market.

- C. Brightman, F. Li and X. Liu, "Chasing performance with ETFs," *Fundamentals*, November 2015.

How difficult is it to design a stock portfolio to achieve a desired performance profile?

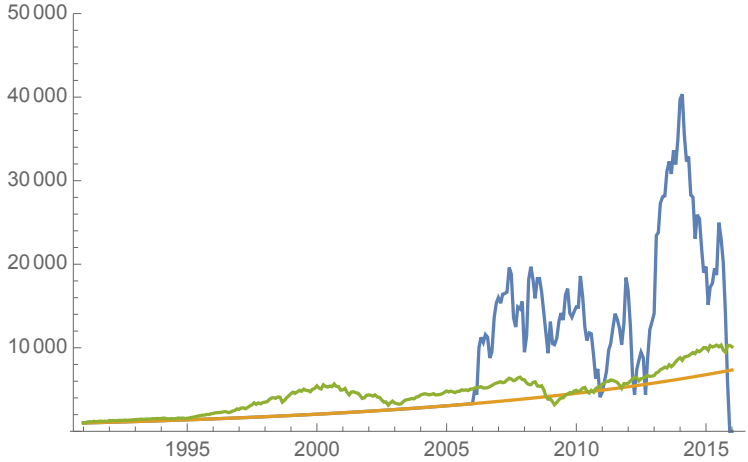
- Given some desired performance time series profile (e.g., a steady 1% per month growth), we devised a computer program that constructs a weighted subset of S&P500 stocks whose performance exactly matches the profile over the specified backtest time period (1991–2005).
- But most of these portfolios fail miserably when presented with new data (2006–2015).

These erratic and often catastrophic results on new (out-of-sample) data are symptomatic of statistical overfitting.

Technical details are given here:

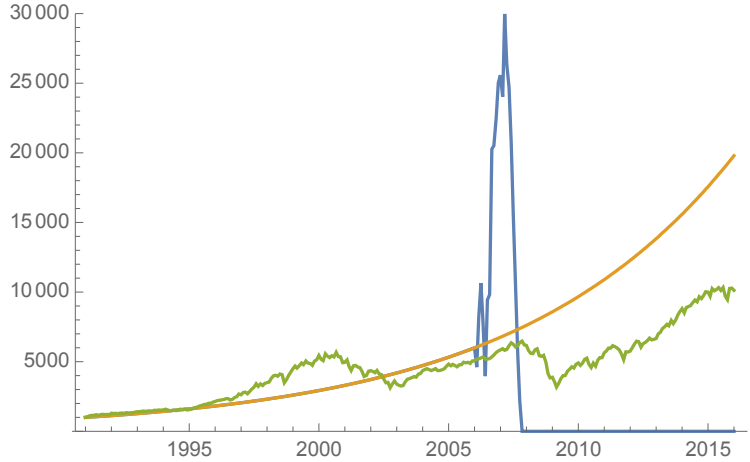
- D. H. Bailey, J. M. Borwein and M. Lopez de Prado, "Stock portfolio design and backtest overfitting," *Journal of Investment Management*, vol. 17 (2017), no. 1, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2739335.

Steady growth profile, APR = 8%



Blue: constructed portfolio; orange: target profile; green: S&P500.

Steady growth profile, APR = 12%



Blue: constructed portfolio; orange: target profile; green: S&P500.

How many actively managed funds have beaten the market averages?

Despite the proliferation of mutual funds, exchange-traded funds and target date funds in recent years, the sad fact is that very few beat the corresponding market averages over a 10-year window:

- Among actively managed "U.S. large value" stock mutual funds, only 8.3% beat the corresponding passive index fund over the 10-year period 2010–2019.
- Among actively managed "U.S. large growth" stock mutual funds, only 8.3% beat the corresponding passive index fund over the 10-year period.
- Among actively managed "world stock" mutual funds, only 26.3% beat the corresponding passive index fund over the 10-year period.

- D. H. Bailey "Mutual fund report card: February 2019," https://mathinvestor.org/2019/02/mutual-fund-report-card-february-2019/.
- Morningstar, "Morningstar active/passive barometer," Feb 2019, https://www.morningstar.com/lp/active-passive-barometer?con=15949&cid=CON_RES0054.

How well have stock market gurus forecasted the market?

Kaissar's analysis of prominent market forecasters (covering 1999 to 2016):

- The forecasters *overestimated* the S&P 500's year-end price by 26.2% on average during the three recession years 2000 to 2002.
- They *underestimated* the index's level by 10.6% for the initial recovery year 2003.
- They *overestimated* the S&P 500's year-end level by a whopping 64.3% in 2008.
- They *underestimated* the index by 10.9% for the first half of 2009.

Kaissar's conclusion: "The forecasts were least useful when they mattered most."

- N. Kaissar, "S&P 500 forecasts: Crystal ball or magic 8?," *Bloomberg News*, 23 December 2016, https://www.bloomberg.com/gadfly/articles/2016-12-23/s-p-500-forecasts-mostly-hit-mark-until-they-matter-most.

Our analysis of market forecasters

In 2012, the CXO Advisory Group ranked 68 forecasters based on their 6,582 forecasts for the S&P 500 index.

In 2018, we extended and advanced this study by classifying forecasts according to time frame and specificity.

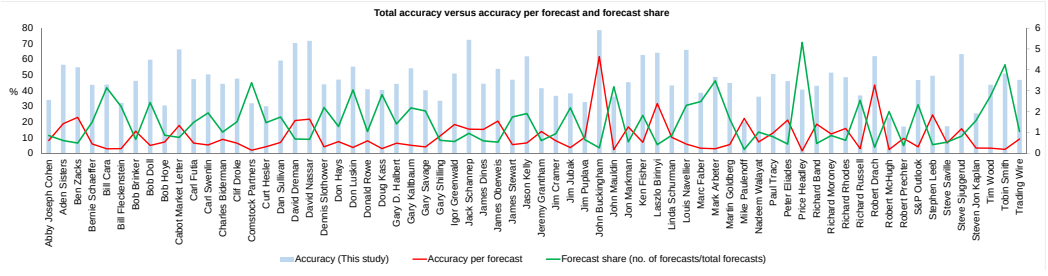
Results:

- Across all forecasts, the average accuracy was 48%.
- Two-thirds of forecasters had an accuracy level below 50%.
- Only about 6% of forecasters had accuracy values between 70% and 79%; the highest accuracy value was still below 80%.

For details:

- D. H. Bailey, J. M. Borwein, A. Salehipour, and M. Lopez de Prado, "Evaluation and ranking of market forecasters", *Journal of Investment Management*, vol. 16, no. 2 (Apr 2018), 47–64, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2944853.

Forecaster accuracy



Study of anomaly indicators

A 2020 study analyzed the statistical reliability of a large set of anomaly indicators (signals in financial market data that may indicate an investment opportunity) from recently published literature.

- The authors concluded that they were not able to statistically replicate most of the findings that had been reported.
- Of the 452 anomaly indicators studied, 65% did not even clear the single test threshold of $t = 1.96$ or greater, when correctly analyzed. With a more stringent criteria that partially compensates for multiple testing, namely $t = 2.78$ at the 5% significance level, the failure rate increases to 82%.

- K. Hou, C. Xue and L. Zhang, "Replicating anomalies," *The Review of Financial Studies*, vol. 33 (May 2020), 2019–2133, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3275496.

The false strategy theorem

- An investment analyst may carry out a large number of simulation trials on historical data, and report only the model, fund or strategy with the maximum Sharpe ratio.
- However, the distribution of the maximum Sharpe ratio is clearly not the same as the distribution of a Sharpe ratio randomly chosen among the trials.
- Instead, the expected value of the maximum Sharpe ratio is greater than the expected value of the Sharpe ratio from a random trial.
- In particular, given an investment strategy with expected Sharpe ratio zero and non-zero variance, the expected value of the maximum Sharpe ratio steadily increases, up from zero, as a function of the number of trials.

Conclusion: No Sharpe ratio criterion is sufficient to rule out a false result. Given enough trials, any preset Sharpe level will eventually be exceeded by random noise.

- M. Lopez de Prado and D. H. Bailey, "The false strategy theorem: A financial application of experimental mathematics," *American Mathematical Monthly*, vol. 128 (2021), no. 9, 825–831, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3221798.

The false strategy theorem

Given a sample of K independent trials, one can deduce an expected maximum Sharpe ratio, namely the hurdle or threshold that the reported Sharpe ratio must exceed before it can be considered a significant finding:

Let $\mathcal{N}$ denote the Gaussian normal distribution; let $Z^{-1}[\cdot]$ denote the inverse of the standard Gaussian cumulative distribution function (CDF); let $E[\cdot]$ and $V[\cdot]$ denote expected value and variance; and let γ is the Euler–Mascheroni constant $= 0.5772156649\ldots$.

False strategy theorem

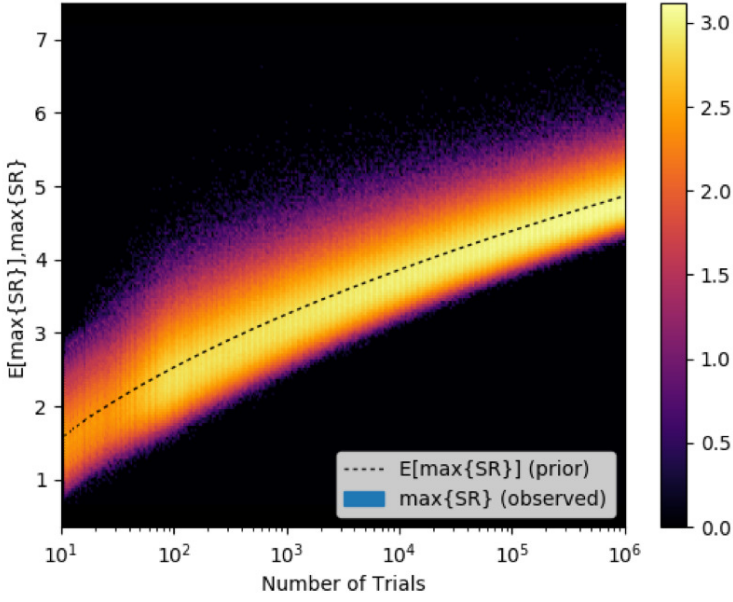
Given a sample of estimated performance statistics $\left\{\widehat{SR}_k\right\}$, $k = 1, \cdots, K$, with independent and identically distributed Gaussian distribution, i.e., $\left\{\widehat{SR}_k\right\} \sim \mathcal{N}\left[0, V\left[\left\{\widehat{SR}_k\right\}\right]\right]$, then

$$E\left[\max_k\left\{\widehat{SR}_k\right\}\right]\left(V\left[\left\{\widehat{SR}_k\right\}\right]\right)^{-1/2} \approx (1-\gamma)Z^{-1}\left[1-\frac{1}{K}\right] + \gamma Z^{-1}\left[1-\frac{1}{Ke}\right]$$

The false strategy theorem: Monte Carlo analysis

1. We generate a random array of size $(S \times K)$, where S is the number of Monte Carlo experiments, and K is number of trials among which the highest Sharpe ratio will be selected. The values in this random array are drawn from a standard normal distribution centered at zero.
2. The rows in this array are centered and scaled to match zero mean and $V\left[\left\{\widehat{SR}_k\right\}\right]$ variance.
3. The maximum value across each row, $\max_k\left\{\widehat{SR}_k\right\}$, is computed, resulting in a number S of such maxima.
4. We compute the empirical average value across the S maxima, namely $\widehat{E}\left[\left\{\widehat{SR}_k\right\}\right]$.
5. We then compute the estimation error, in relative terms, to the analytical prediction made by the theorem as $\epsilon = \widehat{E}\left[\left\{\widehat{SR}_k\right\}\right] / E\left[\left\{\widehat{SR}_k\right\}\right] - 1$.
6. We repeat the previous steps R times, resulting in a set of estimation errors $\{\epsilon_r\}$, $r = 1, 2, \ldots, R$, and allowing us to compute the mean and standard deviation of the estimation errors associated with K trials.

The false strategy theorem: Monte Carlo plot



Why the silence?

Historically scientists have exposed pseudoscience: astrology in the 1800s; young-earth creationism in the 1900s; global warming denial and anti-vaxx propaganda in the 2000s.

Yet present-day financial researchers have remained disappointingly silent with regards to those in the finance community who, knowingly or not:

- Fail to disclose the number of models that were tried in their study or fund design.
- Withhold key details, such as the extent of computer searches used in the analysis.
- Use pseudomathematical jargon (e.g., "Fibonacci ratios," "cycles," "waves," "golden ratios," "parabolic SARs," "pivot points," "rising wedges", "shoulders").
- Use or recommend discredited statistical methods (e.g., "technical analysis").
- Use or recommend models or strategies that are no longer provably effective.
- Make public claims and forecasts not backed up by solid scientific facts.
- Resist rigorous analysis and assessments of long-term performance.

As we wrote in a recent paper:

Our silence is consent, making us accomplices in these abuses.

Mathematicians Against Fraudulent Financial and Investment Advice (MAFFIA)

Visit our MAFFIA websites at:
https://www.maffia.org
https://www.mathinvestor.org

This talk is available at: http://www.davidhbailey.com/dhbtalks/dhb-risk-2022.pdf