FINANCE IS NOT EXCUSED:

WHY FINANCE SHOULD NOT FLOUT BASIC PRINCIPLES OF STATISTICS

David H. Bailey*
Marcos López de Prado^

Forthcoming, Significance (Royal Statistical Society), 2021

First version: April 17, 2021
Current version: July 28, 2021

* Complex Systems Lead (Retired), Lawrence Berkeley National Laboratory, and Research Fellow at University of California, Davis.
^ Professor of Practice, School of Engineering, Cornell University, and Global Head, Quantitative Research & Development, Abu Dhabi Investment Authority. E-mail: ml863@cornell.edu.

FINANCE IS NOT EXCUSED:

WHY FINANCE SHOULD NOT FLOUT BASIC PRINCIPLES OF STATISTICS

ABSTRACT

Several features of financial research make it particularly prone to the occurrence of false discoveries. First, the probability of finding a positive (profitable investment strategy) is very low, due to intense competition. Second, true findings are mostly short-lived, as a result of the non-stationary nature of financial systems. Third, unlike in the natural sciences, it is rarely possible to verify statistical findings through controlled experiments. 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.

Tenure-seeking researchers publish thousands of academic articles that promote dubious investment strategies, without controlling for multiple testing. Some of those articles are written for, funded, or promoted by investment firms with a commercial interest. As a consequence, today's academic finance exhibits some resemblance with medicine's predicament during the 1950-2000 period, when Big Tobacco paid for thousands of studies in support of their bottom line. Unlike finance, medical journals today impose strict controls for multiple testing. Academic finance's denial of its replication crisis risks its branding as a pseudoscience.

Keywords: Multiple testing, selection bias, publication bias, false discovery rate, true positive rate, deflated Sharpe ratio.

JEL Classification: G0, G1, G2, G15, G24, E44.
AMS Classification: 91G10, 91G60, 91G70, 62C, 60E.

1. INTRODUCTION

A common goal of investment funds is to deliver better risk-adjusted performance than the market portfolio, e.g., higher percentage return than the overall market without incurring a greater probability of a financial loss. To devise these investment strategies, firms and analysts typically feed historical market data into computer programs that test a multitude of different combinations of financial instruments, weighting factors, decision points and other parameters, all to identify an "optimal" design. With this "optimal" design in hand, they tout the potential return that an investment based on this design is likely to deliver, based on its simulated performance on historical data (*backtest*). However, in all too many cases, such investments deliver only disappointing performance when actually fielded ([BrightLi2015]).

Three features of financial research make it particularly prone to the occurrence of false discoveries. First, the probability of finding a positive (profitable investment strategy) is very low, due to intense competition. Second, true findings are mostly short-lived, as a result of the non-stationary nature of financial systems. Third, unlike in the natural sciences, it is rarely possible to verify statistical findings through controlled experiments. 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 in true.

A leading reason for investment failures is *backtest overfitting*, namely the usage of historical market data to develop an investment model, fund or strategy, where too many variations are tried, relative to the amount of data available.[1] Backtest overfitting, a form of *selection bias under multiple testing*, has long plagued the field of finance and is now thought to be the leading reason why investments that look great when designed often disappoint when offered to investors. Models, funds and strategies suffering from this type of statistical overfitting typically target the random patters present in the limited *in-sample* test-set on which they are based, and thus often perform erratically when presented with new, truly *out-of-sample data*. The sobering consequence is that a significant portion of the models, funds and strategies employed in the investment world, including many of those marketed to individual investors, may be merely statistical mirages.

The potential for backtest overfitting in the financial field has grown enormously in recent years with the increased utilization of computer programs to search a space of millions or even billions of parameter variations for a given model, fund or strategy, and then to select only the "optimal" choice for publication or market implementation. In this respect, backtest overfitting can be thought of as a financial field's variation of *p*-hacking, namely the deplorable practice, conscious or not, of publishing results of a study based on a subset of the actual data or trials performed, in order to exhibit some desired level of statistical significance [Harvey2017]. In order to control for this effect in the biomedical field, to pick a single example, leading biomedical journals and regulatory bodies increasingly require researchers to report the results from all trial data, so that the likelihood of false positives can be discounted from the reported results. However, in the field of finance, many practitioners do not realize that the very act of performing a computer search for an "optimal" design almost certainly renders the results statistically overfit. Further, textbooks tend to ignore or downplay the challenges posed by multiple testing, and most academic journals fail to require authors to declare the full extent of computer trials involved in a discovery, even

[1] In the following, italics will be used for the first appearance of terms defined in the Glossary.

though the authors may well have performed an extensive computer search for optimal parameters. After the researcher has found a statistical pattern, he can easily build a theoretical explanation around it to rationalize what in reality is nothing more than data snooping.

2. DESIGNING INVESTMENT STRATEGIES BY COMPUTER SEARCH

It is important to note that even very simple investment strategies typically have numerous parameters and choices. As an illustration, suppose that an investor believes that there may be monthly patterns in certain sets of stocks that may lead to a profitable strategy, say by purchasing shares on a fixed day of the month, and selling on another fixed date. There are many variations for such a strategy, as illustrated in Figure 1.

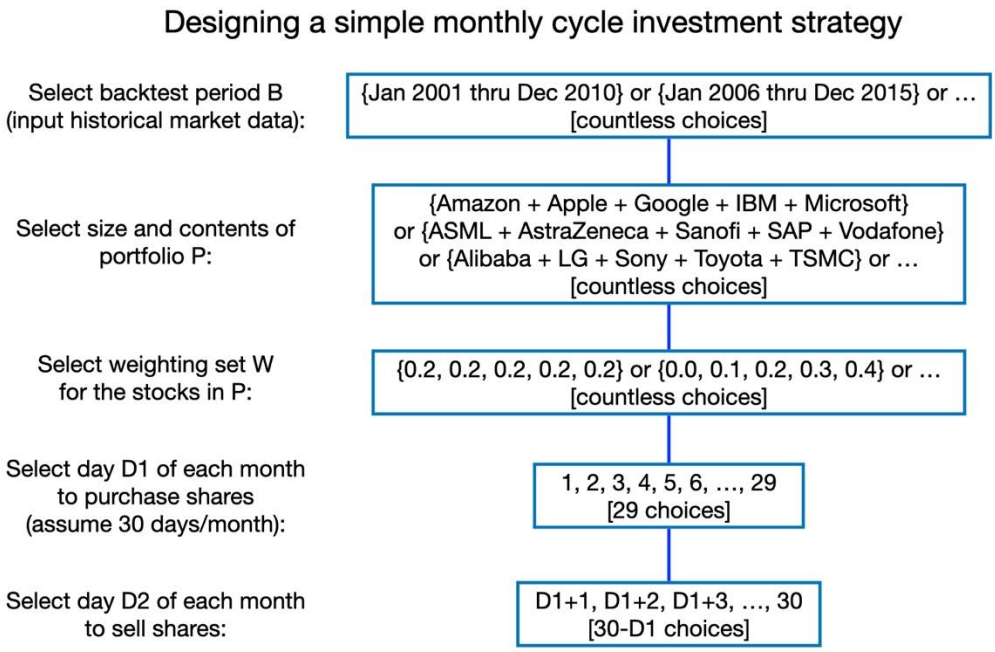

Figure 1 – Illustration of an investment strategy built through trial and error

Note that even with this simple investment strategy (which, by the way, is very unlikely to produce reliable market-beating profits), and even if one fixes the backtest period B, the portfolio P and the weighting set W (each of which has countless choices), there are 435 choices just for the start and end dates of each monthly investment cycle. Admittedly, not all of these choices count as independent trials, but each additional choice raises the probability of a fluke. In any event, it is clear that designing such a strategy by searching via computer over the space of all parameter combinations, in order to design an "optimal" strategy, is virtually certain to produce an overfit backtest.

Given that a real investment strategy might involve many parameters, any one of which may be set to hundreds or more different values, and that it is a relatively simple matter to explore the space of all possible variations by computer, it is clear that backtest overfitting is likely to be the rule, not the exception, in financial strategy development, unless one explicitly guards against it using rigorous statistical tools and a solid economic rationale [LdPLew2019].

3. SOME FINANCIAL BACKGROUND

Investments are typically evaluated by the *Sharpe ratio*, which is essentially the ratio between expected returns in excess of the risk-free rate (or alternative relevant benchmark) and the standard deviation of these returns [Sharpe1994]. To make Sharpe ratios comparable across investments with different sampling frequency, the ratio is often "annualized," by multiplying it with the square root of the number of observations in a year. However, annualized Sharpe ratios should not be thought of as *t*-values for testing the significance of the sample mean, since they do not take into account the number of observations. To correct for this problem, the present authors proposed the Probabilistic Sharpe Ratio [BaiLdP2012], which allows one to test the significance of the Sharpe ratio under general conditions of stationarity and ergodicity.

Another useful tool is the false strategy theorem (see sidebar) [BaiLdP2014]. In practical terms, the false strategy theorem tells us that the optimal outcome of an unknown number of historical market data simulations is right-unbounded. In other words, with enough trials there is no Sharpe ratio threshold sufficiently large to reject the hypothesis that a strategy is false. The rule of thumb of halving the backtest's Sharpe ratio, popular among many investment professionals, has no scientific basis. Again, given the ease with which one can use a computer to explore literally thousands, millions, or even billions of variations of a given strategy and only select the "optimal" variation, it follows that it is very easy to find impressive-looking strategy variations that are nothing more than false positives.

4. THE FALSE STRATEGY THEOREM

As mentioned above, 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. But 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 nonzero variance, the expected value of the maximum Sharpe ratio steadily increases, up from zero, as a function of the number of trials. One can thus 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. This result is known as the false strategy theorem [BaiBorLdPZhu2014].

The false strategy theorem: Given a sample of estimated performance statistics $\{S_k\}, k = 1, \ldots, K$, each independently following a zero-mean, unit-variance Gaussian distribution, then

$$\mathrm{E}[\max_k\{S_k\}] \approx (1-\gamma)\mathrm{Z}^{-1}\left[1-\frac{1}{K}\right] + \gamma\mathrm{Z}^{-1}\left[1-\frac{1}{Ke}\right],$$

where $\mathrm{E}[.]$ denotes expected value, $\mathrm{Z}^{-1}[.]$ denotes the inverse of the standard Gaussian cumulative distribution function, e is Euler's number, and γ is the Euler-Mascheroni constant (approx. 0.5772156649…).

The present authors combined the ideas behind the Probabilistic Sharpe Ratio and the False Strategy Theorem to derive a formula for deflating the Sharpe ratio. The Deflated Sharpe Ratio is the probability that an observed Sharpe ratio was drawn from a distribution with positive mean,

after controlling for sample length, skewness, kurtosis, and the number of strategy variations explored. Let us suppose, for purposes of illustration, that a researcher is constructing a financial model or strategy based on the daily closing values of the FTSE 100 index. An observed annualized Sharpe ratio of 1, where the backtest length is 10 years of daily returns drawn, may appear to be strong evidence of a true discovery. However, if the researcher conducted three or more independent trials, our confidence that the finding is statistically significant is below the standard 95% cutoff (see [BaiLdP2014] for details on these calculations). Figure 2 shows the Deflated Sharpe Ratios for strategies with observed annualized Sharpe ratios of 0.5, 1, and 1.5, as a function of the number of trials. In practice, investment strategies' returns often exhibit positive autocorrelation, negative skewness, and fat tails, which further depress the Deflated Sharpe Ratio. The implication is that, in most cases, as few as three independent trials suffice to produce an investment strategy that is likely false.

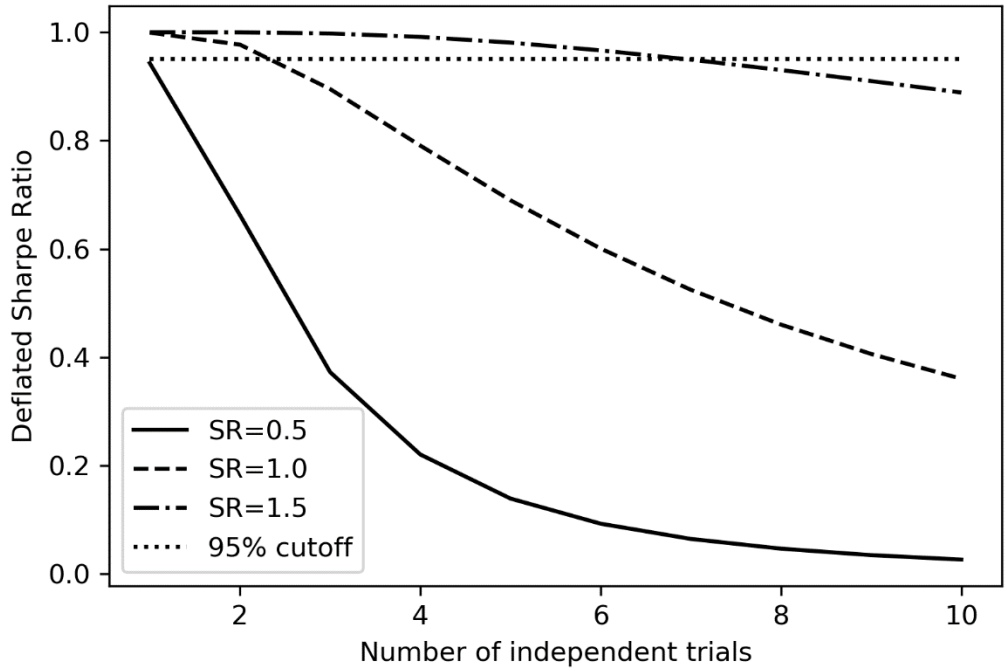

Figure 2 – Deflated Sharpe ratios as a function of the number of trials, based on backtests of 10 years of IID normal daily returns

5. OVERFITTING IN THE DESIGN OF INVESTMENT FUNDS

These may be intriguing theoretical results. But what happens in practice? Consider the problem of designing an investment fund to meet some desired performance profile. One increasingly popular investment product is the *exchange-traded fund* (ETF), namely a *mutual fund* that may be freely traded during the day like an individual stock or bond. Just in the U.S. alone, there is currently over US$5 trillion in ETFs. Hundreds of new ETFs are minted each year, many of them following some custom-designed *index* (i.e., a custom-designed set of stocks and weights). In a 2012 study, researchers found that the median time between the definition of a new index and the inception of a new ETF based on that index dropped from almost three years in 2000 to only 77 days in 2011. As a result, "most indexes have little live performance history for investors to assess in the context of a new ETF investment" [DickPadHamm2012].

One might ask, how do these newly-minted index ETFs perform? 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., in backtests). This strong performance contrasts with average annual excess returns of approx. 0% out-of-sample (see Figure 3) [BrightLi2015]. Such disappointing behavior is entirely consistent with a design process that involves extensive computer exploration of index parameters and selecting only the "optimal" parameters for an *index fund* subsequently fielded in the financial markets.

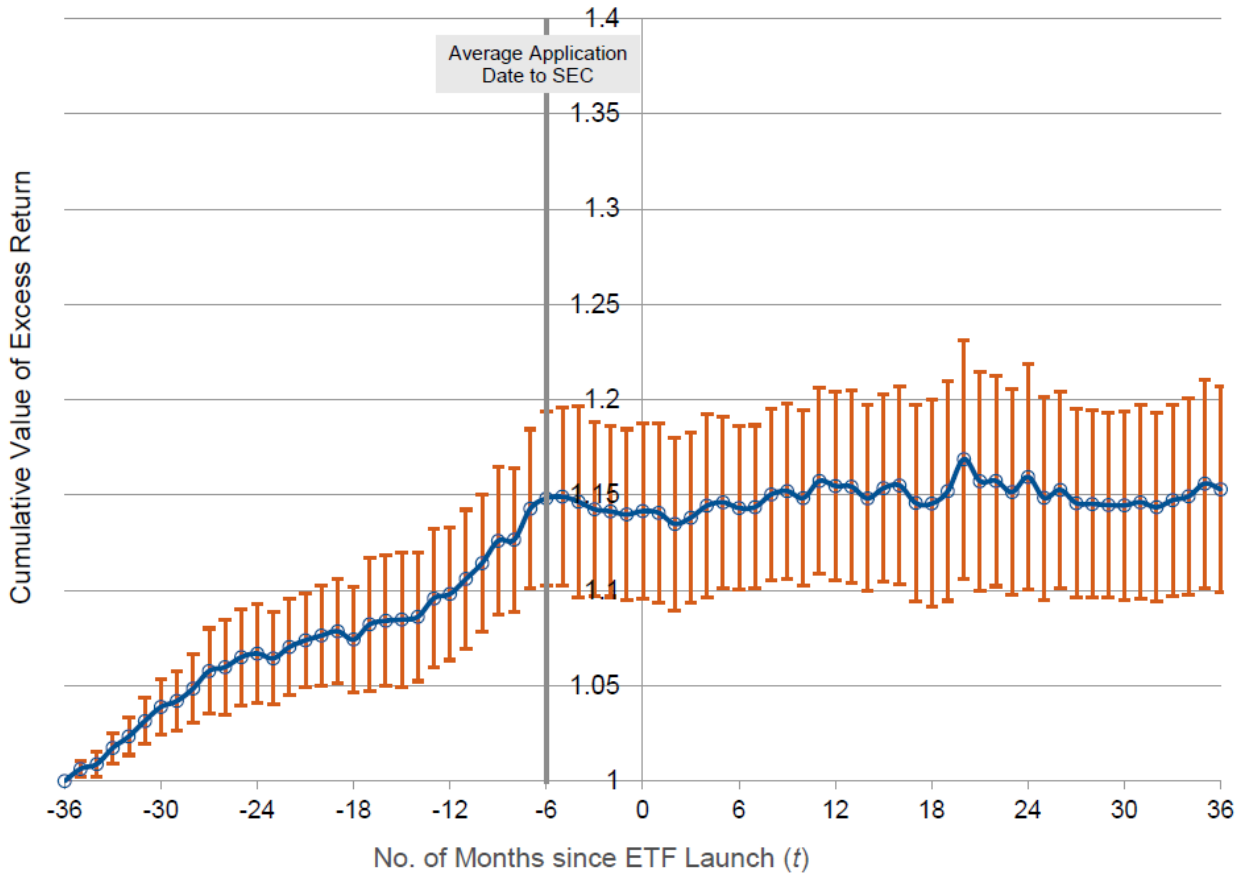

Figure 3 – Backtested performance vs. performance out-of-sample.
Reproduced from [BrightLi2015], Figure 3, with permission.

6. MUTUAL FUNDS, FORECASTERS AND ANOMALIES

In the past few years, it has become clear to many individual investors that few mutual funds or other financial investments can consistently generate gains above the overall market averages. For example, a 2019 report found that among *actively managed funds* (i.e., funds whose stocks are actively selected, bought and sold by experts at a financial firm) in the "U.S. large value" category, only 8.3% beat the comparable *passive index fund* (i.e., a fund where no attempt is made to manage, excerpt to follow a relevant broad-market index) over a 10-year period. Among "world stock" actively managed funds, only 26.3% beat the comparable passive world stock index fund over a 10-year period [JohnMcCull2019)]. In other words, very few actively managed funds have beaten the overall market averages over the long haul.

The issue of selection bias reaches far beyond the realm of quantitative investing. Prominent market forecasters often promote in the media their success at predicting some events, while hoping that the audience has forgotten an equal or greater number of false calls. In 2016, Nir Kaissar analyzed a set of predictions by professional market forecasters over a 17-year period from 1999 through 2016 [Kaissar2016]. He found that, although there was a reasonably high correlation between the average forecast and the year-end price of the S&P 500 index for the given year, these predictions were surprisingly unreliable during major shifts in the market. For example, Kaissar found that the strategists overestimated the S&P 500's year-end price by 26.2% on average during the three recession years 2000 through 2002, yet they underestimated the index level by 10.6% for the initial recovery year 2003. A similar phenomenon was seen in 2008, when strategists overestimated the S&P 500's year-end level by a whopping 64.3% in 2008, but then underestimated the index by 10.9% for the first half of 2009. In other words, as Kaissar lamented, "the forecasts were least useful when they mattered most."

In 2018, the present authors and two colleagues published an in-depth analysis of 68 market forecasters, including many who employ *technical analysis*, a relatively unsophisticated form of historical data analysis [BaiBorLdP2018]. Expanding on an earlier study, we analyzed forecasts based on two key factors: the time frame of the forecast and the importance and specificity of the forecast. Our study found that the average accuracy score of these forecasts was 48%, not significantly different than chance. Although a handful of forecasters did well, there was no statistically significant evidence of overall forecasting skill in the set studied.

In another recent study, Kewei Hou, Chen Xue and Lu Zhang published an in-depth analysis on the statistical reliability of 452 *anomaly indicators* in finance (signals in financial market data that may indicate an investment opportunity), taken from a large set of published papers in the academic finance field [HouXueZhang2020)]. After removing the smallest companies, for which data quality is more questionable, these authors soberly concluded that they were not able to statistically replicate most of these anomaly indicator findings. Out of the 452 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%.

Why the poor performance in these studies? In some cases, the discovered phenomenon may be arbitraged away following its publication [McLeanPontiff2015], however a more likely explanation is that the published phenomenon was a false discovery to begin with, as a result of widespread selection bias under multiple testing.

7. CONCLUSION

Some investors may not be too surprised at the fact that most mutual funds, forecasts and anomaly indicators do not perform much better than random chance, since such an outcome is a straightforward implication of the *efficient market hypothesis*, namely the notion that since modern financial markets efficiently incorporate all available information into prices, unsophisticated investment approaches are unlikely to beat the market averages [Fama1970].

However, markets are not efficient by design. Instead, market efficiency is a byproduct of market competition, thus some firms must be able to extract a profit. But what firms? Among the best performing funds in history, those founded by mathematicians and natural scientists are disproportionally represented. These funds employ sophisticated and rigorous statistical techniques, using state-of-the-art computing equipment operating on numerous massive datasets. These scientists either manage their own assets and thus do not accept outside investors, or their fundraising relies on their track record, not on academic marketing of supposed economic factors, hence they have no incentive to engage in selection bias.

~~Some in the finance field have questioned the existence of a replication crisis. They have argued that concerns with backtest overfitting are overblown, or that certain investment styles (e.g., "factor investing") are not as susceptible as others to overfitting. We do not concur. Rather, the preponderance of poor out-of-sample performance points to a pervasive problem in the field.~~ As Campbell R. Harvey, past President of the American Finance Association lamented in his 2017 presidential address, "our standard testing methods are often ill equipped to answer the questions that we pose," and that there is a serious danger of "stumbling down the same path as some other fields" [Harvey2017].

Researchers publish thousands of academic articles that promote dubious investment strategies, without controlling for multiple testing. In some cases, such methodological error could be attributed to negligence, but in other cases it responds to conflicts of interest. First, tenure-seeking authors have a strong incentive to engage in uncontrolled multiple testing. Second, asset managers commercialize some of the ideas promoted in academic papers. Third, when those products fail to perform, the same asset managers have an incentive to selectively publish evidence supportive of those false discoveries, in the hope that investors will continue to pay fees for a little longer. As a consequence, today's academic finance exhibits some resemblance with medicine's predicament during the 1950-2000 period, when Big Tobacco paid for hundreds of studies in support of their bottom line. Unlike finance, medical journals today impose strict controls for multiple testing. Academic finance's denial of its replication crisis risks its branding as a pseudoscience.

Statisticians pursuing a career in finance should choose their employers carefully. When offered a job, ask yourself, does this firm acknowledge the existence of a replication crisis in finance, or is it still in denial? Does it control for, and report, all trials involved in a discovery? Does it conduct research through backtesting, or does it attempt to refute a well-grounded theory? Does it employ only the most rigorous, objective statistical methodologies, rather than bowing to marketing or other commercial considerations? Considering the answers to those questions, ask yourself whether you want to be part of the solution, or part of the problem.

8. REFERENCES

[BaiBorLdP2018] David H. Bailey, Jonathan M. Borwein and Marcos Lopez de Prado, "Evaluation and ranking of market forecasters," *Journal of Investment Management*, vol. 16, no. 2 (Apr 2018), 47-64.

[BaiBorLdP2017] David H. Bailey, Jonathan M. Borwein and Marcos Lopez de Prado, "Stock portfolio design and backtest overfitting," *Journal of Investment Management*, vol.15 (2017), 75-87.

[BaiLdP2012] David H. Bailey and Marcos Lopez de Prado, "The Sharpe ratio efficient frontier," *Journal of Risk,* vol. 15, no. 2 (2012), 3-44.

[BaiLdP2014] David H. Bailey and Marcos Lopez de Prado, "The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting and non-normality," *Journal of Portfolio Management,* vol. 40, no. 5 (2014), 94-107.

[BaiBorLdPZhu2014] David H. Bailey, Jonathan M. Borwein, Marcos Lopez de Prado and Qiji Jim 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.

[BrightLi2015] Chris Brightman, Feifei Li, Xi Liu, "Chasing performance with ETFs," *Fundamentals,* Research Affiliates, November 2015.

[DickPadHamm2012] Joel M. Dickson, Sachin Padmawar and Sarah Hammer, "Joined at the hip: ETF and index development," July 2012, https://www.vanguardcanada.ca/documents/joined-at-the-hip.pdf.

[Fama1970] Eugene Fama, "Efficient capital markets: A review of theory and empirical work," *Journal of Finance,* vol. 25 (1970), 383-417.

[Harvey2017] Campbell R. Harvey, "Presidential address: The scientific outlook in financial economics," Duke I&E Research Paper No. 2017-05, https://ssrn.com/abstract=2893930.

[HouXueZhang2020] Kewei Hou, Chen Xue and Lu Zhang, "Replicating anomalies," *The Review of Financial Studies,* vol. 33 (May 2020), 2019-2133, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3275496.

[Kaissar2016] Nir Kaissar, "S&P 500 forecasts: Crystal ball or magic 8?," Bloomberg News, 23 Dec 2016, https://www.bloomberg.com/opinion/articles/2016-12-23/s-p-500-forecasts-mostly-hit-mark-until-they-matter-most.

[JohnMcCull2019] Ben Johnson and Adam McCullough, "Morningstar's active/passive barometer," February 2019, https://www.morningstar.com/lp/active-passive-barometer.

[LdPLew2019] Marcos Lopez de Prado and Michael Lewis, "Detection of false investment strategies using unsupervised learning methods," *Quantitative Finance,* vol. 19, no. 9 (2019), 1555-1565.

[McLeanPontiff2015] R. David McLean and Jeffrey Pontiff, "Does Academic Research Destroy Return Predictability?" *Journal of Finance,* vol. 71, no. 1 (2015), 5-32.

[Sharpe1994] William F. Sharpe, "The Sharpe ratio," *Journal of Portfolio Management,* vol. 21 (1994), 49-58.

9. GLOSSARY

Actively managed fund: A mutual fund that is actively managed (stocks or bonds selected, bought and sold) by experts at an investment firm.

Anomaly indicator: A signal in financial market data that may indicate a notable change in direction or an investment opportunity.

Backtest overfitting: The usage of historical market data to develop an investment model, fund or strategy, where too many variations are tried, relative to the amount of data available.

Efficient market hypothesis: The notion that modern financial markets efficiently incorporate all available data into prices, so that simple approaches are unlikely to beat market averages.

Exchange-traded fund (ETF): A mutual fund whose shares may be freely traded during the trading day like shares of an individual stock or bond.

Index: A set of stocks or bonds, together with corresponding weights, typically defined in an objective way by some fixed definition or governing committee; examples: the S&P 500 (U.S. stocks) and the FTSE 100 (European stocks).

Index fund or *passive index fund*: A mutual fund tied to a defined index, where no attempt is made to manage the fund except to follow the index as closely as possible.

In-sample data: Historical data used as input to the design of a model, fund or strategy.

Mutual fund: An investment fund, typically consisting of a certain set of stocks or bonds selected according to some strategy, index or risk level; may be "active" or "passive."

Out-of-sample data: New input data used to test a model, strategy or fund.

Selection bias under multiple testing: Statistical bias that occurs when a researcher conducts multiple tests or analyses, but only reports the test with the best outcome.

Sharpe ratio: The ratio between expected returns in excess of the risk-free rate or alternative relevant benchmark and the standard deviation of these returns.

Technical analysis: A relatively unsophisticated form of historical market data analysis, often involving charts and graphs, that typically ignores statistical problems such as overfitting.