



How Many Monkeys Does it Take to Find a Successful Strategy?

By Michael Edesess and Kwok L. Tsui
October 22, 2013

Give a monkey enough darts and she will eventually hit the bulls-eye on a dartboard. We wouldn't dare consider that monkey an expert dart thrower, but investment professionals have been using essentially that same logic to assert that their strategies – often called “smart betas” – will outperform the market. New research exposes the faulty mathematics upon which such claims are based.

Early this month, four academicians — David H. Bailey and Marcos Lopez de Prado of Lawrence Berkeley National Laboratory, Jonathan M. Borwein of the University of Newcastle in Australia and Qiji Jim Zhu of Western Michigan University — posted a [paper](#) on the Social Science Research Network saying deservedly harsh things about backfitting abuses in investment management. “Recent computational advances allow investment managers to search for profitable investment strategies,” the authors wrote. “In many instances, that search involves a pseudo-mathematical argument, which is spuriously validated through a simulation of its historical performance (also called a backtest).”

They feel strongly enough about these abuses to write: “We would like to raise the question of whether mathematicians should continue to tolerate the proliferation of investment products that are misleadingly marketed as mathematically founded.”

Their statements apply, however, beyond conventional backfitting methodology. Several prominent recent claims for investment strategies have been based on the same one-two punch that the authors decry: a pseudo-mathematical argument combined with historical performance that may or may not be repeated in the future.

We will explain the backtesting addressed in their paper. Then we will consider arguments in favor of “smart beta,” which were bolstered by a widely read [article](#) in the July 6 issue of *The Economist*.

The passive superego tussles with the active id

It is a human tendency to assume that past trends can be extrapolated into the future, but it is widely known that this assumption is not valid when it comes to investment returns.

Nevertheless, many claims for particular investment strategies do present past returns. Investors either can't let go of the intuition that past returns must predict future returns, or



they don't know that this has been disproven. For example, almost all mutual fund advertising is based on past returns data.

But many investors' superegos, which know that past investment history is not predictive of future returns, have been overwhelming the ids, which believe past history is predictive. This can be seen in the mounting popularity of passive market-weighted index funds.

To continue selling investment products on the basis of backtests of past history, marketers sometimes try to buttress that history with mathematical-sounding arguments that claim there's a theoretical reason why history should repeat itself. In almost all cases however, those arguments are only pseudo-mathematics.

Failing to test out-of-sample

Bailey et al. first define in-sample (IS) and out-of-sample (OOS) testing. IS refers to the performance of a strategy in the data sample used to design the strategy. OOS performance is for a data sample that is not used in the design of the strategy. It is common to retain a data set for OOS tests – for example, to divide 10 years of data into two five-year periods to use one for the IS design of the strategy and the other for OOS testing.

The authors say they have observed common practices that are somewhat astonishing. “Although there are plenty of academic studies that claim to have identified profitable investment strategies, their reported results are almost unanimously based on IS statistics,” according to the article. They do not cite examples in their article “for obvious reasons” – professional courtesy, one assumes. But one of us (Edesess) interviewed Dr. Lopez de Prado by Skype after reading the article, and he pointed to specific papers as examples. Yes, there are many articles that merely use time-series prediction methods to project from past data into the future without going through OOS testing validation. Some articles using state-of-the-art time-series prediction methods fail to provide OOS validation.

What surprised us even more – but shouldn't have – was Bailey et al.'s statement that “Hedge fund managers may not be aware that most backtests presented to them by researchers and analysts may be useless, and so they unknowingly package into products faulty investment propositions.”

Even knowing how poorly hedge funds have performed on average, and knowing many of them are merely vehicles for levying excessive fees, investors are still naive enough to believe hedge funds would be smarter than this. But Dr. Lopez de Prado should know. He has worked for two leading hedge funds and says many of those managers do not have the background needed to understand the math they use, resulting in incorrect position sizing and strategy selection.



What is wrong with these procedures?

The problem with identifying an investment strategy based only on IS data is that a researcher can try one strategy after another until one of them produces good performance based on IS data. One of the strategies will eventually produce a large enough positive result – just as a gaggle of chimpanzees typing on typewriters could eventually produce the works of Shakespeare. The joke line used to describe this procedure is, “If you torture the data hard enough, it will confess to anything.”

The remedy is supposed to be that once researchers find a strategy that works on the IS data, they test it on the OOS data and throw it out if it doesn't work anymore. But this is inadequate. It's quite possible to test one strategy after another on the IS data and on the OOS data. You can beat both data sets until they confess in unison.

One remedy for this, say Bailey et al., is “model sequestration” – which means testing on OOS data that are not in hand yet, such as future data. But the length of sequestration needed to reach a conclusion with confidence would be impracticably long – usually several years, and ideally long enough to encompass a full market cycle. The alternative is to directly estimate the probability that the IS results would not match the OOS results.

How many monkeys does it take to find a successful strategy?

Bailey et al. go through the mathematics of determining how many backtests one would have to do to find an investment strategy that performs well on IS data but not on OOS data. They do this by calculating how many strategies need to be tested until the Sharpe ratio appears to be highly positive, when the true Sharpe ratio is actually zero.

The answer is surprisingly few. For example, if only 10 independent model configurations (algorithmic investment strategies) are tested, under certain assumptions the expected highest Sharpe ratio that will be found is 1.57, despite the fact that the expected Sharpe ratio for OOS data is zero.

Is backtest overfitting a fraud?

This practice of running so many alternative models that they will eventually yield a highly positive Sharpe ratio even when the true Sharpe ratio is zero is called overfitting the data. The problem is that researchers almost never report how many alternative models were tried. Bailey et al. pose the question, “Is backtest overfitting a fraud?”

They show that a deliberate fraud could be perpetrated in this manner if researchers did not report how many strategies were tried before one was found to work. In Edesess's



book *The Big Investment Lie*, he showed that cigarette manufacturers could use this methodology by manufacturing thousands of cigarette brands to prove that one of those brands actually prevents cancer – and then to market it heavily. The Food and Drug Administration (FDA) will not allow this to happen with cigarettes, but the Securities and Exchange Commission (SEC) allows it for mutual funds.

Bailey et al. show that for some stochastic stock market price models, such as one that reverts to the mean, strategies found to outperform the market on historical IS data tend to underperform on future OOS data. They use as an example a random walk that is required to return to its mean after a period of years. Any outperformance discovered by backtesting in an early period, however significant, will reverse itself in the remainder of the whole time period. If, on the other hand, the model is a random walk model, historical outperformance will merely disappear in the future.

The Wobegon Heights miracle

A mainstay of Garrison Keillor’s popular radio program “A Prairie Home Companion” is his “news from Lake Wobegon,” a rambling tale of events transpiring in the previous week in the mythical Minnesota town of Lake Wobegon. Keillor always signs off by saying, “Well that’s the news from Lake Wobegon, where all the women are strong, all the men are good-looking, and all the children are above average.”

It’s a joke, of course – if all the children were above average, the average would rise.

Nevertheless, some purveyors of investment strategies have managed to promote the idea that most or even all randomly generated stock portfolios are above average. Let’s call this the Wobegon Heights effect.

This begins with a pseudo-mathematical argument, which is usually worded as follows: A capitalization-weighted index overweights overvalued stocks and underweights undervalued stocks. Therefore it will underperform, and any other index is better.

This statement is virtually never proven by translation into true mathematical form. When stated rigorously, [it is easily shown to be false](#). Nevertheless it has had a certain marketing appeal that has carried it – and the investment strategies that are sold based on it – surprisingly far.

The same school of thought has argued that “alternative” indices – which have come to be called “smart beta” – will be superior to market-capitalization-weighted indexes. For example, advocates of so-called “fundamental indexing” – which tilts toward value stocks by weighting stocks with lower market-to-book ratios more heavily – have tried to argue



that the strategy will work for theoretical reasons, not just because they expect past performance to persist.

The tilt toward value and small-cap

A large quantity of historical data has shown that the past performance of “value” (low market-to-book ratio) and small-capitalization stocks has been superior over long periods of time to that of the market as a whole. Some speculate that value and small-cap stocks may have risk characteristics that are not fully revealed by conventional measures. (We believe it may also be due to the fact that finance researchers almost always use holding-period returns in their regressions instead of continuously-compounded returns – i.e., log-returns – which would correct for the skewness in returns distributions. The two alternatives produce [very different results](#).)

Pseudo-mathematical arguments and historical data have been used to argue for investing portfolios in value and small-cap stocks. These arguments are intended to overcome the objection that the value and small-cap effect may reverse itself in the future, if mean-reversion takes effect or if the popularity of these asset classes becomes so great (or already has become so great) that prices are driven up.

If it could be argued, however, that there’s a mathematical reason why the outperformance should continue, then the skepticism might be overcome.

The Cass Business School study

Unfortunately for those of us seeking to preserve mathematical rigor in the investment industry, a study conducted by the Cass Business School at City University London seems to confirm the Wobegon Heights effect. Three Cass researchers — Andrew Clare, Nick Motson and Steve Thomas — produced a [two-part](#) study in which they performed a simulation and found that almost all randomly-generated portfolios beat the market average.

This is impossible of course, just as the Lake Wobegon effect – all the children are above average – is impossible. So we got a copy of the study to find out how they did it.

As suspected, the problem lies in their method of “randomly” generating portfolios. This method is clearly explained in a [video](#) by Motson. The methodology they use is one that will result in the average portfolio being an equally weighted portfolio rather than a cap-weighted portfolio.

Edesess communicated with Motson in several emails, in which Motson said that they also tried a second random-generation methodology and obtained the same result. That



[second methodology](#), called “sampling uniformly from the unit simplex” (a way of generating random stock weightings in which all sets of weightings are equally likely) also gives rise to randomly generated portfolios that are, on average, equally weighted.

A naïve investment professional might conclude that Motson and his colleagues diligently researched appropriate methods to randomly generate portfolios. Yet they could have used other, equally valid random generation methods – which they themselves identify – to create portfolios that are cap-weighted. Those methods produce entirely different results. Had they based their conclusions on those random portfolio generation methods, they would have shown that the Wobegon Heights effect disappears.

The result is nothing more than the argument from history redux

An equal-weighted portfolio must have more weight in smaller-cap stocks than a market-cap-weighted portfolio. Since we already know that small-cap stocks outperformed the cap-weighted total market over the period that the Cass researchers studied (without adjusting for risk), it is no surprise that their randomly generated portfolios outperformed the cap-weighted market index.

So the Cass results add nothing to the knowledge we already have. They provide no theoretical, mathematical or computational result capable of adding heft to the bare fact that over some historical time period, small-cap stocks outperformed the broad market average.

Nonetheless, the Cass research findings have been interpreted by many as confirmation that non-market-cap-weighted portfolios (which, in “smart beta” indices, are almost invariably portfolios tilted toward small-cap and value stocks) will continue to outperform.

An adage attributed to Mark Twain says that a lie can travel halfway around the world while the truth is getting its boots on. The one-two punch of a pseudo-mathematical argument and historical data mining has proven particularly adept at globe-circling.



Michael Edesess, a mathematician and economist, is a visiting fellow with the Centre for Systems Informatics Engineering at City University of Hong Kong, a partner and chief investment officer of Denver-based [Fair Advisors](#), and a project consultant at the Fung Global Institute. In 2007, he authored a book about the investment services industry titled [The Big Investment Lie](#), published by Berrett-Koehler. His new book, [The Three Simple Rules of Investing](#), co-authored with Kwok L. Tsui, Carol Fabbri, and George Peacock, will be published by Berrett-Koehler in spring 2014.

Kwok L. Tsui is a distinguished statistician and Head of the Systems Engineering and Engineering Management department and Chair Professor of Industrial Engineering at City University of Hong Kong.

www.advisorperspectives.com

For a free subscription to the Advisor Perspectives newsletter, visit:
<http://www.advisorperspectives.com/subscribers/subscribe.php>