Evaluating Trading Strategies

CAMPBELL R. HARVEY AND YAN LIU

CAMPBELL R. HARVEY
is a professor at Duke
University in Durham,
NC, and a fellow at
the National Bureau
of Economic Research
in Cambridge, MA.
cam.harvey@duke.edu

YAN LIU
is an assistant professor at
Texas A&M University in
College Station, TX.
y-liu@mays.tamu.edu

We provide some new tools to evaluate trading strategies. When it is known that many strategies and combinations of strategies have been tried, we need to adjust our evaluation method for these multiple tests. Sharpe ratios and other statistics will be overstated. Our methods are simple to implement and allow for the real-time evaluation of candidate trading strategies.

Consider the following trading strategy detailed in Exhibit 1.[1] Although there is a minor drawdown in the first year, the strategy is consistently profitable through 2014. Indeed, the drawdowns throughout the history are minimal. The strategy even does well during the financial crisis. Overall, this strategy appears very attractive and many investment managers would pursue this strategy.

Our research (see Harvey and Liu [2014a] and Harvey et al. [2014]) offers some tools to evaluate strategies such as the one presented in Exhibit 1. It turns out that simply looking at average profitability, consistency, and size of drawdowns is not sufficient to give a strategy a passing grade.
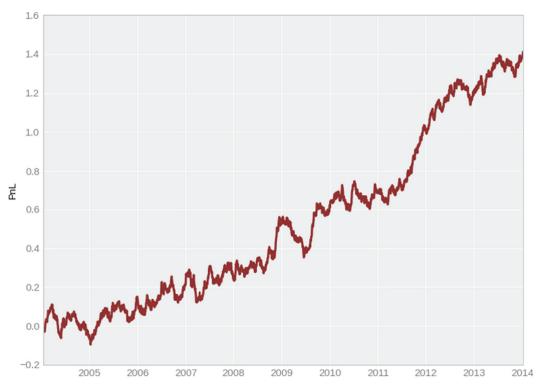
TESTING IN OTHER FIELDS OF SCIENCE

Before presenting our method, it is important to take a step back and determine whether there is anything we can learn in finance from other scientific fields. Although the advent of machine learning is relatively new to investment management, similar situations involving a large number of tests have been around for many years in other sciences. It makes sense that there may be some insights outside of finance that are relevant to finance.

Our first example is the widely heralded discovery of the Higgs Boson in 2012. The particle was first theorized in 1964—the same year as William Sharpe's paper on the capital asset pricing model (CAPM) was published.[2] The first tests of the CAPM were published eight years later[3] and Sharpe was awarded a Nobel Prize in 1990. For Peter Higgs, it was a much longer road. It took years to complete the Large Hadron Collider (LHC) at a cost of about $5 billion.[4] The Higgs Boson was declared "discovered" on July 4, 2012, and Nobel Prizes were awarded in 2013.[5]

So why is this relevant for finance? It has to do with the testing method. Scientists knew that the particle was rare and that it decays very quickly. The idea of the LHC is to have beams of particles collide. Theoretically, you would expect to see the Higgs Boson in one in ten billion collisions within the LHC.[6] The Boson quickly decays and key is measuring the decay signature. Over a quadrillion collisions were conducted and a massive amount of data was collected. The

EXHIBIT 1
A Candidate Trading Strategy

Source: AHL Research.

problem is that each of the so-called decay signatures can also be produced by normal events from known processes.

To declare a discovery, scientists agreed to what appeared to be a very tough standard. The observed occurrences of the candidate particle (Higgs Boson) had to be five standard deviations different from a world where there was no new particle. Five standard deviations is generally considered a tough standard. Yet in finance, we routinely accept discoveries where the *t*-statistic exceeds two—not five. Indeed, there is a hedge fund called Two Sigma.

Particle physics is not alone in having a tougher hurdle to clear. Consider the research done in biogenetics. In genetic association studies, researchers try to link a certain disease to human genes and they do this by testing the causal effect between the disease and a gene. Given that there are more than 20,000 human genes that are expressive, multiple testing is a real issue. To make it even more challenging, a disease is often not caused by

a single gene but the interactions among several genes. Counting all the possibilities, the total number of tests can easily exceed a million. Given this large number of tests, a tougher standard must be applied. With the conventional thresholds, a large percentage of studies that document significant associations are not replicable.[7]

To give an example, a recent study in *Nature* claims to find two genetic linkages for Parkinson's disease.[8] About a half a million genetic sequences are tested for the potential association with the disease. Given this large number of tests, tens of thousands of genetic sequences will appear to affect the disease under conventional standards. We need a tougher standard to lower the possibility of false discoveries. Indeed, the identified gene loci from the tests have *t*-statistics that exceed 5.3.

There are many more examples such as the search for exoplanets. However, there is a common theme in these examples. A higher threshold is required because the number of tests is large. For the Higgs Boson, there were potentially trillions of tests. For research in bioge-

netics, there are millions of combinations. With multiple tests, there is a chance of a fluke finding.

REVALUATING THE CANDIDATE STRATEGY

Let's return to the candidate trading strategy detailed in Exhibit 1. This strategy has a Sharpe ratio of 0.92. There is a simple formula to translate the Sharpe ratio into a t-statistic:[9]

$$T\text{-statistic} = \text{Sharpe Ratio} \times \sqrt{\text{Number of years}}$$
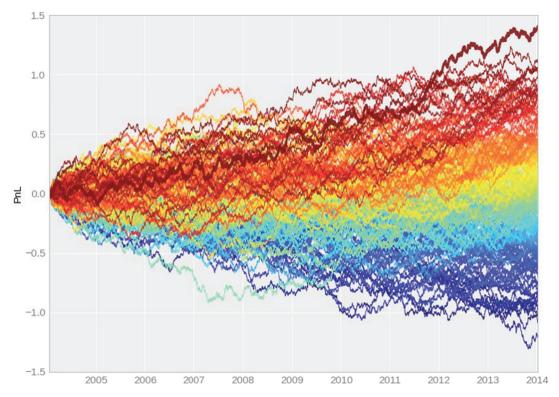
In this case, the t-statistic is 2.91. This means that the observed profitability is about three standard deviations from the null hypothesis of zero profitability. A three-sigma event (assuming a normal distribution) happens only 1% of the time. This means that the chance that our trading strategy is a false discovery is less than 1%.

However, we are making a fundamental mistake with the statistical analysis. The statement about the false discovery percentage is conditional on an independent test. This means there is a single test. That is unlikely to be the case in our trading strategy and it was certainly not the case with the research conducted at the LHC, where there were trillions of tests. With multiple tests, we need to adjust our hurdles for establishing statistical significance. This is the reason why the researchers at LHC used a five-sigma rule. This is the reason why biomedical researchers routinely look for four-sigma events.

Multiple testing is also salient in finance—yet little has been done to adjust the way that we conduct our tests. Exhibit 2 completes the trading strategy example.[10]

Each of the trading strategies in Exhibit 2 was randomly generated at the daily frequency. We assumed an annual volatility of 15% (about the same as the S&P 500) and a mean return of zero. The candidate trading strategy

E X H I B I T 2
200 Randomly Generated Trading Strategies

Source: AHL Research.

highlighted in Exhibit 1 is the best strategy in Exhibit 2 (dark red curve).

To be clear, all of the strategies in Exhibit 2 are based on random numbers—not actual returns. Although the candidate trading strategy in Exhibit 1 seemed very attractive, it was simply a fluke. Yet the usual tools of statistical analysis would have declared this strategy "significant." The techniques we will offer in this article will declare the candidate strategy, with the Sharpe ratio of 0.92, insignificant.

It is crucial to correct for multiple testing. Consider a simple example that has some similarities to this example. Suppose we are interested in predicting Y. We propose a candidate variable X. We run a regression and get a t-statistic of 2.0. Assuming that no one else had tried to predict Y before, this qualifies as an independent test and X would be declared significant at the 5% level. Now let's change the problem. Suppose we still want to predict Y. However, now we have 20 different X variables, X_1, X_2,…,X_{20}. Suppose one of these variables achieves a t-statistic of 2.0. Is it really a true predictor? Probably not. By random chance, when you try so many variables, one might work.

Here is another classic example of multiple tests. Suppose you receive a promotional email from an investment manager promoting a stock. The email asks you to judge the record of recommendations in real time. Only a single stock is recommended and the recommendation is either long or short. You get an email every week for 10 weeks. Each week the manager is correct. The track record is amazing because the probability of such an occurrence is very small ($0.5^{10} = 0.000976$). Conventional statistics would say there is a very small chance (0.00976% this is a false discovery, that is, the manager is no good). You hire the manager.

Later you find out the strategy. The manager randomly picks a stock and initially sends out 100,000 emails with 50% saying long and 50% saying short. If the stock goes up in value, the next week's mailing list is trimmed to 50,000 (only sending to the long recommendations). Every week the list is reduced by 50%. By the end of the tenth week, 97 people would have received this *amazing* track record of 10 correct picks in a row.

If these 97 people had realized how the promotion was organized, then getting 10 in a row would be expected. Indeed, you get the 97 people by multiplying $100,000 \times 0.5^{10}$. There is no skill here. It is random.

There are many obvious applications. One that is immediate is in the evaluation of fund managers. With more than 10,000 managers, you expect some to randomly outperform year after year.[11] Indeed, if managers were randomly choosing strategies, you would expect at least 300 of them to have five consecutive years of outperformance.

Our research offers some guidance on handling these multiple-testing problems.

TWO VIEWS OF MULTIPLE TESTING

There are two main approaches to the multiple-testing problem in statistics. They are known as the family-wise error rate (FWER) and the false discovery rate (FDR). The distinction between the two is very intuitive.

In the family-wise error rate, it is unacceptable to make a single false discovery. This is a very severe rule but completely appropriate for certain situations. With the FWER, one false discovery is unacceptable in 100 tests and equally as unacceptable in 1,000,000 tests. In contrast, the false discovery rate views *unacceptable* in terms of a proportion. For example, if one false discovery were unacceptable for 100 tests, then 10 are unacceptable for 1,000 tests. The FDR is much less severe than the FWER.

Which is the more appropriate method? It depends on the application. For instance, the Mars One foundation is planning a one-way manned trip to Mars in 2024 and has plans for many additional landings.[12] It is unacceptable to have any critical part fail during the mission. A critical failure is an example of a false discovery (we thought the part was good but it was not—just as we thought the investment manager was good but she was not).

The best-known FWER test is called the Bonferroni test. It is also the simplest test to implement. Suppose we start with a two-sigma rule for a single (independent) test. This would imply a t-ratio of 2.0. The interpretation is that the chance of the single false discovery is only 5% (remember a single false discovery is unacceptable). Equivalently, we can say that we have 95% confidence that we are not making a false discovery.

Now consider increasing the number of tests to 10. The Bonferroni method adjusts for the multiple tests. Given the chance that one test could randomly show up as significant, the Bonferroni requires the confidence

level to increase. Instead of 5%, you take the 5% and divide by the number of tests, that is, 5%/10 = 0.5%. Again equivalently, you need to be 99.5% confident with 10 tests that you are not making a single false discovery. In terms of the *t*-statistic, the Bonferroni requires a statistic of at least 2.8 for 10 tests. For 1,000 tests, the statistic must exceed 4.1.

However, there are three issues with the Bonferroni test. First, there is the general issue about the FWER error rate versus FDR. Evaluating a trading strategy is not a mission to Mars. Being wrong could cost you your job and money will be lost—but it is unlikely a matter of life and death. However, reasonable people may disagree with this view.

The second issue is related to correlation among the tests. There is a big difference between trying 10 variables that are all highly correlated and 10 variables that are completely uncorrelated. Indeed, at the extreme, if the 10 tests were perfectly correlated, this is equivalent to a single, independent test.

The third issue is that the Bonferroni test omits important information. Since the work of Holm [1979], it has been known that there is information in the individual collection of test statistics and it can be used to sharpen the test.[13] The Bonferroni test ignores all of this information and derives a hurdle rate from the original level of significance divided by the total number of tests.

Let's first tackle the last issue. Holm [1979] provides a way to deal with the information in the test statistics. Again, suppose we have 10 tests. We know that the hurdle for the Bonferroni method would be 0.005 or 0.5%.

The Holm method begins by sorting the tests from the lowest *p*-value (most significant) to the highest (least significant). Let's call the first $k = 1$ and the last $k = 10$. Starting from the first test, the Holm function is evaluated.

$$p_k = \frac{\alpha}{M + 1 - k}$$

where α is the level of significance (0.05) in our case and M is the total number of tests.

Suppose the most significant test in our example has a *p*-value of 0.001. Calculating the Holm function, we get $0.05/(10 + 1 - 1) = 0.005$. The Holm function gives the hurdle (observed *p*-value must be lower than the hurdle). Given the first test has a *p*-value of 0.001, it passes the test. Notice the hurdle for the first test is identical to the Bonferroni. However, in contrast to the Bonferroni, which has a single threshold for all tests, the other tests will have a different hurdle under Holm; for example, the second test would be $0.05/(10 + 1 - 2) = 0.0055$.

Starting from the first test, we sequentially compare the *p*-values with their hurdles. When we first come across the test such that its *p*-value fails to meet the hurdle, we reject this test and all others with higher *p*-values.

The Holm test captures the information in the distribution of the test statistics. The Holm test is less stringent than the Bonferroni because the hurdles are relaxed after the first test. However, the Holm still fits into the category of the FWER. Next, we explore the other approach.

As mentioned earlier, the false discovery rate approach allows an expected proportional error rate (see Benjamini and Hochberg [1995] and Benjamini and Yekutieli [2001]). As such, it is less stringent than both the Bonferroni and the Holm test. It is also easy to implement. Again, we sort the tests. The BHY formula is

$$p_k = \frac{k \times \alpha}{M \times c(M)}$$

where $c(M)$ is a simple function that is increasing in M and equals 2.93 when $M = 10$.[14] In contrast to the Holm test, we start from the last test (least significant) and evaluate the BHY formula.

For the last test, $k = M = 10$, the BHY hurdle is $0.05/c(10) = 0.05/2.93 = 0.0171$. For the second last test, $k = M - 1 = 9$, the BHY hurdle is $9 \times 0.05/10 \times 2.93 = 0.0154$. Notice that these hurdles are larger and thus more lenient that the Bonferroni implied hurdle (that is, 0.0050).

Starting from the last test, we sequentially compare the *p*-values with their thresholds. When we first come across the test such that its *p*-value falls below its threshold, we declare this test significant and all tests that have a lower *p*-value.

Similar to the Holm test, BHY also relies on the distribution of test statistics. However, in contrast to the Holm test that begins with the most significant test, the BHY approach starts with the least significant

test.[15] There are usually more discoveries with BHY. The reason is that BHY allows for an expected proportion of false discoveries, which is less demanding than the absolute occurrence of false discoveries under the FWER approaches. We believe the BHY approach is the most appropriate for evaluating trading strategies.

FALSE DISCOVERIES AND MISSED DISCOVERIES

So far, we have discussed false discoveries, which are trading strategies that appear to be profitable—but they are not. Multiple testing adjusts the hurdle for significance because some tests will appear significant by chance. The downside of doing this is that some truly significant strategies might be overlooked because they did not pass the more stringent hurdle.

This is the classic tension between Type I errors and Type II errors. The Type I error is the false discovery (investing in an unprofitable trading strategy). The Type II error is missing a truly profitable trading strategy. Inevitably there is a tradeoff between these two errors. In addition, in a multiple testing setting it is not obvious how to jointly optimize these two types of errors.
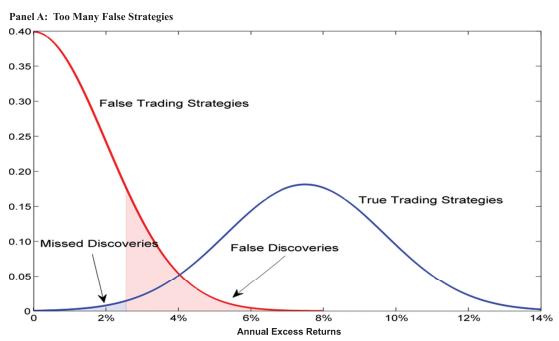
Our view is the following. Making the mistake of using the single test criteria for multiple tests induces a very large number of false discoveries (large amount of Type I error). When we increase the hurdle, we greatly reduce the Type I error at minimal cost to the Type II (missing discoveries). Exhibit 3 illustrates this point.

The first panel denotes the mistake of using single test methods. There are two distributions. The first is the distribution of strategies that don't work. It has an average return of zero. The second is the distribution of truly profitable strategies, which has a mean return greater than zero. Notice that there is a large amount of Type I error (false discoveries). The second panel shows what happens when we increase the threshold. Notice the number of false discoveries is dramatically reduced. However, the increase in missed discoveries is minimal.
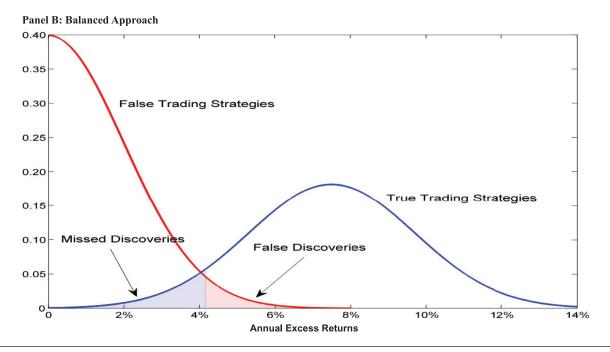
HAIRCUTTING SHARPE RATIOS

Harvey and Liu [2014a] provide a method for adjusting Sharpe ratios to take into account multiple testing. Sharpe ratios based on historical back tests are often inflated because of multiple testing. Researchers

E X H I B I T   3
False Trading Strategies, True Trading Strategies

Panel A: Too Many False Strategies

E X H I B I T 3 (*continued*)

Panel B: Balanced Approach

explore many strategies and often choose to present the one with the largest Sharpe ratio. But the Sharpe ratio for this strategy no longer represents its true expected profitability. With a large number of tests, it is very likely that the selected strategy will appear to be highly profitable just by chance. To take this into account, we need to haircut the reported Sharpe ratio. In addition, the haircut needs to be larger if there are more tests tried.

Take the candidate strategy in Exhibit 1 as an example. It has a Sharpe ratio of 0.92 and a corresponding *t*-statistic of 2.91. The *p*-value is 0.4%; hence, if there were only one test, the strategy would look very attractive because there is only a 0.4% chance it is a fluke. However, with 200 tests tried, the story is completely different. Using the Bonferroni multiple-testing method, we need to adjust the *p*-value cutoff to $0.05/200 = 0.00025$. Hence, we would need to observe a *t*-statistic of at least 3.66 to declare the strategy a true discovery with 95% confidence. The observed *t*-statistic, 2.92 is well below 3.66—hence, we would pass on this strategy.

There is an equivalent way of looking at the Bonferroni test. To declare a strategy true, its *p*-value must be less than some predetermined threshold such as 5%

(or 95% confidence that the identified strategy is not false):

$$p\text{-value of test} < \text{threshold}$$

Bonferroni divides the threshold (0.05) by the number of tests, our case 200:

$$p\text{-value of test} < 0.05/200.$$

Equivalently, we could multiply the *p*-value of the individual test by 200 and check each test to identify which ones are less than 0.05, i.e.

$$(p\text{-value of test}) \times 200 < 0.05$$

In our case, the original *p*-value is 0.004 and when multiply by 200 the adjusted *p*-value is 0.80 and the corresponding *t*-statistic is 0.25. This high *p*-value is significantly greater than the threshold, 0.05. Our method asks how large the Sharpe ratio should be in order to generate a *t*-statistic of 0.25. The answer is 0.08. Therefore, knowing that 200 tests have been tried and under Bonferroni's test, we successfully declare the candidate strategy with the original Sharpe ratio of 0.92 as insignificant—the Sharpe ratio that adjusts for multiple tests

is only 0.08. The corresponding haircut is large, 91% (= (0.92 − 0.08)/0.92).

Turning to the other two approaches, the Holm test makes the same adjustment as Bonferroni since the *t*-statistic for the candidate strategy is the smallest among the 200 strategies. Not surprisingly, BHY also strongly rejects the candidate strategy.

The fact that each of the multiple-testing methods rejects the candidate strategy is a good outcome because we know all of these 200 strategies are just random numbers. A proper test also depends on the correlation among test statistics, as we discussed previously. This is not an issue in the 200 strategies because we did not impose any correlation structure on the random variables. Harvey and Liu [2014b] explicitly take the correlation among tests into account and provide multiple-testing-adjusted Sharpe ratios using a variety of methods.

AN EXAMPLE WITH STANDARD AND POOR'S CAPITAL IQ

To see how our method works on a real data set of strategy returns, we use the S&P Capital IQ database. It includes detailed information on the time-series of 484 strategies for the U.S. equity market. Additionally, these strategies are catalogued into eight groups based on the types of risks to which they are exposed. We choose the most profitable strategy from each of the three categories: price momentum, analyst expectations, and capital efficiency. These trading strategies are before costs and, as such, the Sharpe ratios will be overstated.

The top performers in the three categories generate Sharpe ratios of 0.83, 0.37, and 0.67, respectively. The corresponding *t*-statistics are 3.93, 1.14, and 3.17 and their *p*-values (under independent testing) are 0.00008, 0.2543, and 0.0015.[16] We use the BHY method—our recommended method—to adjust the three *p*-values based on the *p*-values for the 484 strategies (we assume the total number of tried strategies is 484, that is, there are no missing tests). The three BHY-adjusted *p*-values are 0.0134, 0.9995, and 0.1093 and their associated *t*-statistics are 2.47, 0.00 and 1.60. The adjusted Sharpe ratios are 0.52, 0.00 and 0.34, respectively. Therefore, by applying the BHY method, we haircut the Sharpe ratios of the three top performers by 37% (=(0.83 − 0.52)/0.83), 100% (=(0.42 − 0)/0.42), and 49% (=(0.67 − 0.34)/0.67).[17]

IN-SAMPLE AND OUT-OF-SAMPLE

Until now, we evaluate trading strategies from an in-sample (IS) testing perspective, that is, we use all the information in the history of returns to make a judgment. Alternatively, one can divide the history into two subsamples—one in-sample period and the other out-of-sample (OOS) period—and use OOS observations to evaluate decisions made based on the IS period.

There are a number of immediate issues. First, often the OOS period is not really out-of-sample because the researcher knows what has happened in that period. Second, in dicing up the data, we run into the possibility that, with fewer observations in the in-sample period, we might not have enough power to identify true strategies. That is, some profitable trading strategies do not make it to the OOS stage. Finally, with few observations in the OOS period, some true strategies from the IS period may not pass the test in the OOS period and be mistakenly discarded.

Indeed, for the three strategies in the Capital IQ data, if we use the recent five years as the OOS period for the OOS approach, the OOS Sharpe ratios are 0.64, −0.30, and 0.18, respectively. We see that the third strategy has a small Sharpe ratio and is insignificant (*p*-value = 0.53) for this five-year OOS period, although it is borderline significant for the full sample (*p*-value = 0.11), even after multiple-testing adjustment. The problem is that with only 60 monthly observations in the OOS period, a true strategy will have a good chance to fail the OOS test.

Recent research by López de Prado and his coauthors pursues the out-of-sample route and develops a concept called the probability of backtest overfitting (PBO) to gauge the extent of backtest overfitting (see Bailey et al. [2013a, b] and López de Prado [2013]). In particular, the PBO measures how likely it is for a superior strategy that is fit IS to underperform in the OOS period. It succinctly captures the degree of backtest overfitting from a probabilistic perspective and should be useful in a variety of situations.

To see the differences between the IS and OOS approach, we again take the 200 strategy returns in Exhibit 2 as an example. One way to do OOS testing is to divide the entire sample in half and evaluate the performances of these 200 strategies based on the first half of the sample (IS), that is, the first five years. The evaluation is then put into further scrutiny based on the

second half of the sample (OOS). The idea is that strategies that appear to be significant for the in-sample period but are actually not true will likely to perform poorly for the out-of-sample period. Our IS sample approach, on the other hand, uses all ten years' information and makes the decision at the end of the sample. Using the method developed by López de Prado and his coauthors, we can calculate PBO to be 0.45.[18] Therefore, there is high chance (that is, a probability of 0.45) for the IS best performer to have a below-median performance in the OOS. This is consistent with our result that based on the entire sample, the best performer is insignificant if we take multiple testing into account. However, unlike the PBO approach that evaluates a particular strategy selection procedure, our method determines a haircut Sharpe ratio for each of the strategies.

In principle, we believe there are merits in both the PBO as well as the multiple-testing approaches. A successful merger of these approaches could potentially yield more powerful tools to help asset managers successfully evaluate trading strategies.

TRADING STRATEGIES AND FINANCIAL PRODUCTS

The multiple-testing problem greatly confounds the identification of truly profitable trading strategies and the same problems apply to a variety of sciences. Indeed, there is an influential paper in medicine by Ioannidis [2005] called "Why Most Published Research Findings Are False." Harvey et al. [2014] look at 315 different financial factors and conclude that most are likely false after you apply the insights from multiple testing.

In medicine, the first researcher to publish a new finding is subject to what they call the winner's curse. Given the multiple tests, subsequent papers are likely to find a lesser effect or no effect (which would mean the research paper would have to be retracted). Similar effects are evident in finance where Schwert [2003] and McLean and Pontiff [2014] find that the impact of famous finance anomalies is greatly diminished out-of-sample—or never existed in the first place.

So where does this leave us? First, there is no reason to think that there is any difference between physical sciences and finance. Most of the empirical research in finance, whether published in academic journals or put into production as an active trading strategy by an investment manager, is likely false. Second, this implies

that half the financial products (promising outperformance) that companies are selling to clients are false.

To be clear, we are not accusing asset managers of knowingly selling false products. We are pointing out that the statistical tools being employed to evaluate these trading strategies are inappropriate. This critique also applies to much of the academic empirical literature in finance—including many papers by one of the authors of this article (Harvey).

It is also clear that investment managers want to promote products that are most likely to outperform in the future. That is, there is a strong incentive to get the testing right. No one wants to disappoint a client and no one wants to lose a bonus—or a job. Employing the statistical tools of multiple testing in the evaluation of trading strategies reduces the number of false discoveries.

LIMITATIONS AND CONCLUSIONS

Our work has two important limitations. First, for a number of applications the Sharpe ratio is not appropriate because the distribution of the strategy returns is not normal. For example, two trading strategies might have identical Sharpe ratios but one of them might be preferred because it has less severe downside risk.

Second, our work focuses on individual strategies. In actual practice, the investment manager needs to examine how the proposed strategy interacts with the current collection of strategies. For example, a strategy with a lower Sharpe ratio might be preferred because the strategy is relatively uncorrelated with current strategies. The denominator in the Sharpe ratio is simply the strategy volatility and does not measure the contribution of the strategy to the portfolio volatility. The strategy portfolio problem, that is, adding a new strategy to a portfolio of existing strategies is the topic of Harvey and Liu [2014c].

In summary, the message of our research is simple. Researchers in finance, whether practitioners or academics, need to realize that they will find seemingly successful trading strategies by chance. We can no longer use the traditional tools of statistical analysis that assume that no one has looked at the data before and there is only a single strategy tried. A multiple-testing framework offers help in reducing the number of false strategies adapted by firms. Two sigma is no longer an appropriate benchmark for evaluating trading strategies.

ENDNOTES

[1]AHL Research [2014].

[2]Sharpe [1964] for the CAPM. Higgs [1964] for the Higgs Boson.

[3]See Black et al. [1972] and Fama and MacBeth [1973].

[4]A 2009 brochure put the cost of the machine at about $4 billion and this does not include all other costs. See http://cds.cern.ch/record/1165534/files/CERN-Brochure-2009-003-Eng.pdf retrieved July 10, 2014.

[5]CMS [2012] and ATLAS [2012].

[6]See Baglio and Djouadi [2011].

[7]See Hardy [2002].

[8]See Simon-Sanchez et al. [2009].

[9]When returns are realized at higher frequencies, Sharpe ratios and the corresponding t-statistics can be calculated in a straightforward way. Assuming that there are N return realizations in a year and the mean and standard deviation of returns at the higher frequency is μ and σ, the annualized Sharpe ratio can be calculated as $(\mu \times N)/(\sigma \times \sqrt{N}) = (\mu/\sigma) \times \sqrt{N}$. The corresponding t-statistic is $(\mu/\sigma) \times \sqrt{N \times \text{Number of Years}}$. For example, for monthly returns, the annualized Sharpe ratio and the corresponding t-statistic are $(\mu \times \sigma) \times \sqrt{12}$ and $(\mu/\sigma) \times \sqrt{12 \times \text{Number of years}}$, respectively, where μ and σ are the monthly mean and standard deviation for returns. Similarly, assuming μ and σ are the daily mean and standard deviation for returns and there are 252 trading days in a year, the annualized Sharpe ratio and the corresponding t-statistics are $(\mu/\sigma) \times \sqrt{252}$ and $(\mu/\sigma) \times \sqrt{252 \times \text{Number of years}}$.

[10]AHL Research [2014].

[11]See Barras et al. [2010].

[12]See http://www.mars-one.com/mission/roadmap retrieved July 10, 2014.

[13]See Schweder and Spjotvoll [1982].

[14]More specifically, $c(M) = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{M} = \sum_{i=1}^{M} \frac{1}{i}$ and approximately equals $\log(M)$ when M is large.

[15]For the p-value thresholds, whether or not BHY is more lenient than Holm depends on the specific distribution of p-values, especially when the number of tests M is small. When M is large, BHY implied hurdles are usually much larger than Holm.

[16]We have 269 monthly observations for the strategies in the price momentum and capital efficiency groups and 113 monthly observations for the strategies in the analyst expectations group. Therefore, the t-statistics are calculated as $0.83 \times \sqrt{(269/12)} = 3.93$, $0.37 \times \sqrt{(113/12)} = 1.14$ and $0.67 \times \sqrt{(269/12)} = 3.17$.

[17]Applying the Bonferroni test, the three p-values are adjusted to be 0.0387, 1.0 and 0.7260. The corresponding adjusted Sharpe ratios are 0.44, 0, 0.07 and the haircuts are 47%, 100%, and 90%. These haircuts are larger than under the BHY approach.

[18]See AHL Research [2014]. The 0.45 is based on 16 partitions of the data.

REFERENCES

AHL Research. "Strategy Selection." AHL internal research paper, London, 2014.

ATLAS collaboration. "Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC." *Physics Letters B*, Vol. 716, No. 1 (2012), pp. 1-29.

Bailey, D., J. Borwein, M. López de Prado, and Q.J. Zhu. "Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample." Working paper, Lawrence Berkeley National Laboratory, 2013a.

——. "The Probability of Backtest Overfitting." Working paper, Lawrence Berkeley National Laboratory, 2013b.

Barras, L., O. Scaillet, and R. Wermers. "False Discoveries in Mutual Fund Performance: Measuring Luck in Estimated Alphas." *Journal of Finance,* No. 65 (2010), pp. 179-216.

Baglio, J., and A. Djouadi. "Higgs Production at the IHC." *Journal of High Energy Physics,* Vol. 1103, No. 3 (2011), p. 55.

Benjamini, Y., and Y. Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society,* Series B 57 (1995), pp. 289-300.

Benjamini, Y., and D. Yekutieli. "The Control of the False Discovery Rate in Multiple Testing Under Dependency." *Annals of Statistics,* No. 29 (2001), pp. 1165-1188.

Black, F., M.C. Jensen, and M. Scholes. "The Capital Asset Pricing Model: Some Empirical Tests." In *Studies in the Theory of Capital Markets,* edited by M. Jensen, pp. 79-121. New York: Praeger, 1972.

CMS collaboration. "Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC." *Physics Letters B,* Vol. 716, No. 1 (2012), pp. 30-61.

Fama, E.F., and J.D. MacBeth. "Risk, Return, and Equilibrium: Empirical Tests." *Journal of Political Economy,* No. 81 (1973), pp. 607-636.

López de Prado, M. "What to Look for in a Backtest." Working paper, Lawrence Berkeley National Laboratory, 2013.

McLean, R.D., and J. Pontiff. "Does Academic Research Destroy Stock Return Predictability?" Working paper, University of Alberta, 2014.

Ioannidis, J.P. "Why Most Published Research Findings Are False." *PLoS Medicine,* No. 2, e124 (2005), pp. 694-701.

Hardy, J. "The Real Problem in Association Studies." *American Journal of Medical Genetics,* Vol. 114, No. 2 (2002), p. 253.

Harvey, C.R., and Y. Liu. "Backtesting." Working paper, Duke University, 2014a. Available at http://papers.ssrn.com/abstract=2345489.

———. "Multiple Testing in Economics." Working paper, Duke University, 2014b. Available at http://papers.ssrn.com/abstract=2358214.

———. "Incremental Factors." Working paper, Duke University, 2014c.

Harvey, C.R., Y. Liu, and H. Zhu. "… and the Cross-section of Expected Returns." Working paper, Duke University, 2014. Available at http://papers.ssrn.com/abstract=2249314.

Higgs, P. "Broken Symmetries and the Masses of Gauge Bosons." *Physical Review Letters,* Vol. 13, No. 16 (1964), pp. 508-509.

Schweder, T., and E. Spjotvoll. "Plots of *P*-values to Evaluate Many Tests Simultaneously." *Biometrika* No. 69 (1982), pp. 439-502.

Schwert, G.W. "Anomalies and Market Efficiency." In *The Handbook of the Economics of Finance,* edited by G. Constantinides, M. Harris, and R.M. Stulz, 1 (2003), pp. 937-972.

Simon-Sanchez, J., C. Schulte, and T. Gasser, "Genome-wide Association Study Reveals Genetic Risk Underlying Parkinson's Disease." *Nature Genetics,* No. 41 (2009), pp. 1308-1312.

To order reprints of this article, please contact Dewey Palmieri at dpalmieri@iijournals.com or 212-224-3675.