12-2020

Testing Investment Strategies for Superior Predictive Ability

Jack K. Baldwin
Utah State University

Testing Investment Strategies for Superior Predictive Ability

JACK BALDWIN | Utah State University

When different models are tested on one data sample and repeatedly altered in order to be found significant, the results are likely spurious. This is data-snooping – an ever-growing problem in the finance industry likely due to fierce competition and developments in data processing capacity. In academia, although recognized as a deplorable practice, data-snooping is likewise pervasive perhaps as a result of poor incentive structures at both the university and publisher levels. I manifest the problem of data-snooping through multiple academic and industry examples and then summarize Halbert White and Peter Hansen's offered solutions, White's Reality Check and Hansen's Test for Superior Predictive Ability. I demonstrate the application of their tests by examining several passive investment strategies applicable to recent market moves and report my results.

1. INTRODUCTION

Overfit in data analysis occurs when our model predicts the variation in the data but only spuriously. Data-snooping is the generalization of this problem to multiple models. Often, it comes about by performing many statistical tests on one set of data and only reporting those with significant results. Andrew Lo (1994) gives a loose definition of data-snooping – finding patterns in data that do not exist. Although he gets the point across, this definition is not fully correct. Patterns found in the data under examination obviously exist, but they do not carry over into future data. They also have no reason for existing i.e. we cannot explain why they exist. For the purposes of this paper, data-snooping is defined as finding patterns in data which do not continue to exist on live data. It is particularly rampant in financial time-series data because there are numerous studies performed on our single iteration of the stock market.

2. A FEW EXAMPLES

In a 2013 study, Preis et al. (2013) reported that they could time the stock market by tracking Google search words data. They considered 98 different keywords related to the concept of stock markets including terms suggested by Google's related keyword identifier Google Sets. They analyzed 1 to 6 week moving averages for each keyword using both global search term data and U.S. National data. They sum up their trading decision as follows:

Our trading strategy can be decomposed into two strategy components: one in which a decrease in search volume prompts us to buy (or take a long position) and one in which an increase in search volume prompts us to sell (or take a short position).

They tested their models on DJIA data from the 7-year period between January 1, 2004 to February 22, 2011 and reported their best strategy based on the search term, "Debt" in the national data, offered a 23% annual return, compared with a 2.2% annual return for the buy-and-hold benchmark. They conclude that their results suggest search volume data could have been exploited for handsome profits but there's a caveat.

In this work, we provide a quantification of the relationship between changes in search volume and changes in stock market prices. Future work will be needed to provide a thorough explanation of the underlying psychological mechanisms which lead people to search for terms like debt before selling stocks at a lower price.

In other words, they found a significant pattern but acknowledge they don't know why it works. Reporting results with no explanation of why something works is the finest recipe for data-snooping. The researchers considered 98 keywords on both global and national data, 6 different moving averages, and 2 strategies (long or short). Multiply those together and they tested a total of 2,352 potential strategies. With so many possible paths, naturally they should find dozens of spurious patterns. 10 years following their results, Smith (2020) tested their Google search "debt" strategy on the 7-year period following the original study from February 22, 2011 to December 31, 2018 and found the strategy returned 2.81% annually compared to an 8.6% annual return for the buy-and-hold benchmark.

Now for an industry example. In 2017, a company called Equibot (Equibot, 2020) launched AIEQ, the first AI Powered Equity ETF which "harnesses the power of IBM Watson". The fund boasts that their "system mimics a team of 1,000 research analysts working around the clock analyzing millions of data points each day." How has it performed? Not as well as their dreamy claims suggest it should. In the first year after its inception AIEQ underperformed SPY, one of the most popular S&P 500 based ETFs by .4%. In the second year it underperformed by 4.95%. Over the third and final year, AIEQ outperformed SPY by 4.8% cumulatively

Figure 1

underperforming by 2.7% – a gap which has widened even more over the past month. Why the poor performance? If AIEQ's bot is truly analyzing millions of data points daily, the number of useless correlations found is far outpacing those found to be useful. AIEQ is engaging in data-snooping, making decisions based on spurious patterns. Daily trade volume in the fund suggests people are losing interest.

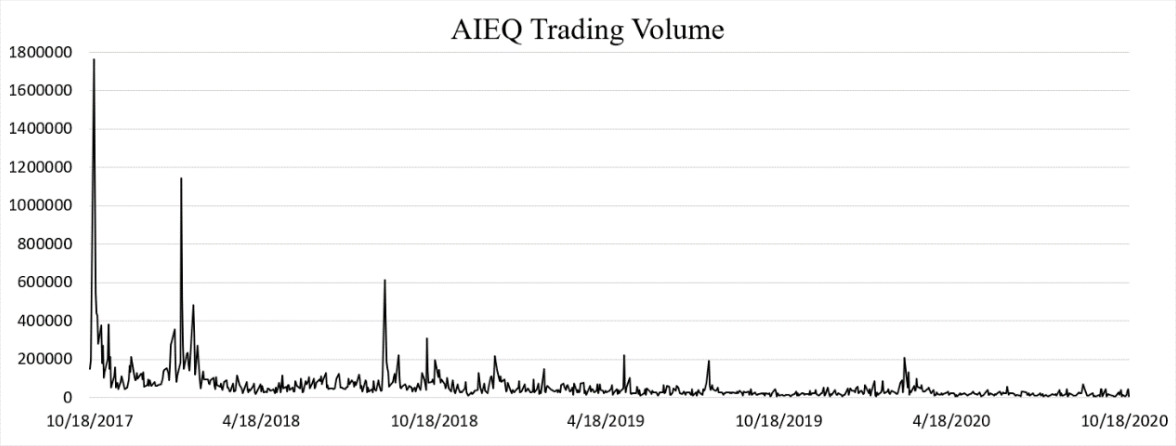

Figure 2

Smith (2020) conjectures if we suppose 1 in 1000 patterns is useful and we use reliable statistical tests which will both correctly identify useful patterns 95% of the time but also incorrectly classify useless patterns as useful 5% of the time (95% confidence interval), there is only a 20% chance that a pattern we discover is useful. The assumption that 1 in 1000 patterns is useful may also be presumptuous. He explains,

We do not know precisely how many useless patterns are out there waiting to be discovered, but we do know that with big data and powerful computers, it is a very large number that is getting larger every day, which means that the probability that a randomly discovered pattern is useful is getting ever closer to 0.

3. HUMAN BIAS

The problem of data-snooping is naturally exacerbated by human bias. If Utah State University's Maverik Stadium were filled to capacity, more than 25,000 football fans could be heard across Utah's Cache Valley. Assume we replace those 25,000 fans with 25,000

quantitative traders and ask them all to flip a fair coin. Those who flip a tail are shuffled out of the stands while those who flip a head get to flip again. We repeat this process – weeding out the tail flippers – until only 24 or so quants are left. How many flips did it take to narrow down our group of experts from 25,000 to 24? – Ten. (25,000/2^10 = 24.414). Only 24 of our original 25,000 are able to consecutively flip a head 10 times. These geniuses, we conclude, are excellent coin flippers.

Human bias naturally draws our attention to the unique and seemingly impossible. Thus, our surprise when someone flips heads 10 times in a row, the probability of which is only 1 out of 1,024 (1/(.5^10)). Andrew Lo (1994) demonstrated this paradox through his own simulation. Assuming there is a 50% chance of beating the stock market each year, Lo calculated the possibility that a portfolio manager would outperform the market in 11 out of 13 years to be about 9 in 1,000 or 0.9%. In that light, those who beat the market in 11 out of 13 years seem extraordinary.

However, if taken from a pool of 500 managers, Lo found the probability of the best manager outperforming the market in 11 out of 13 years to be 57.7% – far less impressive. Looking backwards our attention is drawn to unique events and extraordinary outliers or that which we perceive to be extraordinary as human bias blinds us from perceiving the pool of ordinary events from which outliers are drawn. If we step back and consider the ocean of normal everyday occurrences from which oddities emerge, statistical inference shows they should be expected. Likewise, although our 0.098% chance of flipping 10 consecutive heads seems impossible. Out of a pool of 25,000 people, we are almost certain to find several who do.

4. A PROBLEM OF REPLICABILITY

Over the last two decades the field of psychology has been under scrutiny for the irreplicability of dozens of landmark experiments (Bohannon, 2015). Stanley et al. (2018) argue that this is due to the low statistical power and high heterogeneity inherent to the data used in these experiments. Some prestigious psychological journals, however, are beginning to ban the use of p-values in published papers (Woolston, 2015) suggesting that the problem is in the research not the data. Perhaps what ails the field of psychology is a severe case of data-snooping bias.

Consider a humble assistant psychology professor seeking tenure. In order to save his

salary and position with the university he must add valuable research and commentary to the already vast library of psychological studies. How does he do it? By p-hacking. He performs many statistical tests on his data and makes a publishable paper from the few tests that return significant p-values. Going through that process enough times solidifies his position as a contributing scholar but only adds to the mounting pile of data-snooped literature.

Now consider a quantitative trader who also aims to advance her career by finding trading strategies which yield above market returns. She too will use statistical analysis to test hundreds of models on financial data and, like the professor, may only report those which show significant results. Unlike the professor, however, her engagement in data-snooping likely ends in her termination unless she gets lucky or truly stumbles across a strategy no one else has tried.

The professor and trader are dealing with the problem of replicability. Few of their working models are easily reproducible on any other data set than the one used in their original analysis – evidence of poor models. Perhaps then, the key to effectively testing the success of a model is simply to run it on out-of-sample data. Doing so can make data-snooping more difficult but is still far from a complete solution.

Assuming a confidence interval of 95%, we have a 5% chance of classifying an insignificant model as significant. In other words, for every 20 models we test, we should expect one to falsely return significant results. If we then test those models on out-of-sample data, we should again expect 5% of them to return significant results purely by chance. Therefore 1 in every 400 models will spuriously pass both the in-sample and out-of-sample test (1/20 * 1/20). Remember that 'model' includes every minute alteration to any model tested. Processing power available today makes it very easy to stretch our model count into the thousands (Smith 2020). Recall the vast number of Google search strategies tried by Preis et al (2013).

5. WHITES REALITY CHECK

There exist more robust solutions than out-of-sample data to address the problem of data-snooping as well as adjust end results based on the size of our estimator pool. Of classical models, White's (2000) Reality Check (WRC) and the frameworks which build upon it are the most influential. I will utilize Sheppard's (2014) outline of Hansen's (2005) Test for Superior Predictive Ability (SPA) which is a simple expansion to WRC.

WRC uses a similar hypothesis testing structure as Diebold and Mariano (1995) and West (1996) for determining predictive ability. Their method examines whether a model has equal predictive ability as a benchmark by computing the difference of the two loss functions. Mathematically, the loss differentials at time t are expressed as

$$\delta_{k,t} = L\left(y_{t-h}, \hat{y}_{t-h,BM|t}\right) - L\left(y_{t-h}, \hat{y}_{t-h,k|t}\right)$$

for models $k = 1, \ldots, m$, where $\hat{y}_{t-h,BM|t}$ are the predictions from the benchmark model. WRC then puts the loss differentials at time t into a vector δ_t.

$$\delta_t = \begin{bmatrix} L\left(y_{t+h}, \hat{y}_{t+h,BM|t}\right) - L\left(y_{t+h}, \hat{y}_{t+h,1|t}\right) \\ L\left(y_{t+h}, \hat{y}_{t+h,BM|t}\right) - L\left(y_{t+h}, \hat{y}_{t+h,2|t}\right) \\ \vdots \\ L\left(y_{t+h}, \hat{y}_{t+h,BM|t}\right) - L\left(y_{t+h}, \hat{y}_{t+h,m|t}\right) \end{bmatrix}$$

with $\mu = E(\delta_t)$. The null-hypothesis for WRC suggests that the loss differentials from the models will be less than or equal to the losses from the benchmark. Alternatively, the model losses are greater than those of the benchmark.

$$H_0: \mu \leq 0 \qquad H_1: \mu > 0$$

Hansen (2005), a previous student of White, shows that WRC is sensitive to manipulation with the inclusion of poor and irrelevant alternative models and offers his Test for Superior Predictive Ability (SPA) as an improvement to WRC. Hansen's model standardizes the loss differentials of the models and uses a sample-dependent null distribution to omit very bad models. This omission corrects the model sensitivity problem of WRC.

As defined by Sheppard (2014), the test-statistic of the SPA is the maximum standardized loss differential over the analyzed time period and represents the best performing model. It is written as

$$T^{SPA} = \max_{j=1,\ldots,m} \left(\frac{\overline{\delta}_j}{\sqrt{\widehat{\omega}_j^2 / T}} \right)$$

Where T is the sum total of time periods t in the sample and $\widehat{\omega}_t^2$ is the estimate of the long-run variance of $\overline{\delta}_j$ and T. $\widehat{\omega}_t^2$ is calculated as

$$\widehat{\omega}_t^2 = \gamma\hat{j}_{j,0} + 2\sum_{i=1}^{T-1} \kappa_i \gamma_{j,i}$$

where $\hat{\gamma}_{j,0}$ is the variance of $\overline{\delta}_j$ and $\gamma_{j,i}$ is the auto-covariance of $\delta_{j,t}$. κ_i weights the auto-covariances using window length W for a stationary bootstrap which is key to WRC.

$$\kappa_i = \frac{T-i}{T}\left(1-\frac{1}{W}\right)^i + \frac{i}{T}\left(1-\frac{1}{W}\right)^{T-i}$$

In theory WRC assumes a known sample distribution and utilizes a pure Monte Carlo simulation to generate draws from that distribution. In practice, however, White applies the stationary bootstrap of Politis and Romano (1994) to the loss differentials in which data are resampled in random blocks of length (W), and geometrically distributed.

Bootstrapping is an effective sample simulation method and was prior to WRC. WRC's innovation is to use bootstrapping to develop the sampling distribution for the best models tested from a body of models under the assumption that none of the models are expected to perform better than a given benchmark. Put broadly, WRC is the meta-analysis of a statistical analysis (Aaronson, 2007). In the case of the psychology professor and trader, a traditional bootstrap could be used to simulate thousands of additional data paths on which they could test their models (out-of-sample data). WRC, on the other hand, simulates their entire testing process thousands of times therefore illuminating the data-snooping they are engaging in.

The bootstrap algorithm constructs a new data set by drawing with replacement from the existing loss differential vectors. Drawing from the differentials rather than from the original data helps increase stationarity. Drawing begins at a randomly selected point in the data. A probability value $1/W$ is used to determine the jump to a new index where W is the length of the data window i.e. how much data is pulled. A number between 1 and 0 is randomly generated for each draw. If the generated number is greater than $1/W$, the next sequential data point is also drawn. When the generated number is less than $1/W$, a new starting point is randomly assigned. Notice that with each sequentially drawn index the probability of a jump increases. This process continues until a new bootstrapped sample $[\delta_{b,t}^{\star}]$ is constructed with the same length as the original and a corresponding test statistic $T_{s,b}^{\star SPA}$ is calculated for each iteration of the bootstrap.

$$T_{s,b}^{\star SPA} = \max\left(\frac{T^{-1}\sum_{t=R+1}^{T}\delta_{j,b,t}^{\star} - I_j^s\overline{\delta}_j}{\sqrt{\widehat{\omega}_j^2/T}}\right)$$

$$s = u, c, l$$

$T_{s,b}^{\star SPA}$ is the maximum standardized differential of each bootstrapped sample. The variable s corresponds to three indicators u, c, and l. I_j^s represents the indicator functions.

$$I_j^u = 1, \quad I_j^c = \frac{\overline{\delta}_j}{\sqrt{\widehat{\omega}_j^2/T}} > -\sqrt{2\ln\big(\ln(T)\big)}, \quad I_j^l = \overline{\delta}_j > 0$$

After the stationary bootstrap is repeated many times, each indicator is used to compute three different p-values: p-value$_c$, p-value$_u$, and p-value$_c$.

$$p - \text{value}_s = \frac{1}{B}\sum_{b=1}^{b} I\left[T_{s,b}^{\star}SPA > T^{SPA}\right]$$

The p-values are calculated as the percentage of the bootstrapped t-statistics which are greater than the original sample t-statistic, shown by$T_{s,b}^{\star SPA} > T^{SPA}$. The upper p-value, p-value$_u$ represents the value obtained under the assumption that all the models are as good as the benchmark and is obtained by recentering each bootstrapped loss differential around the mean $\bar{\delta}_j$. If a model is rejectably bad, it may be best to exclude it from the test. The lower p-value, p-value$_l$ only recenters the bootstrapped loss differentials of those which outperform the benchmark. The consistent p-value, p-value$_c$ provides a value which represents the true p-value of the test and is the one I report.

6. INVESTMENT STRATEGIES

In early 2020 the public was inundated by news regarding a new acute respiratory virus which began spreading in Wuhan China in December of 2019. With the spread of the virus and the declaration of a pandemic by the World Health Organization came panic selling in financial markets across the world. By March 23rd the S&P 500 index had fallen 34% from its February 19th high. Prices took little time to rebound, however, as the Federal Reserve and Federal Government unveiled the largest economic stimulus package ever seen. By August 18, the S&P

500 index had achieved a new all-time record high marking the fastest bear market recovery in history (Wursthorn, 2020).

Figure 3

Even as prices kept falling speculation regarding the "shape of recovery" began to emerge (Marte, 2020) as did talk of "buying the dip". After-all, investors who are lucky or smart enough to time the bottom of economic downturns can cash in handsomely on the upswing. Google Trends data show that the search term "buy the dip" reached its highest point ever in March 2020.

"Buying the dip" is a sensible strategy. We can easily imagine impressive potential returns if we forgo investing when equities are expensive and then dump money into the market when assets are cheap. Contrary to buying the dip, most working US adults passively invest a portion of their salary into the market every two weeks or month in a process called dollar cost averaging. What returns could be expected if, rather than dollar cost averaging, an investor built

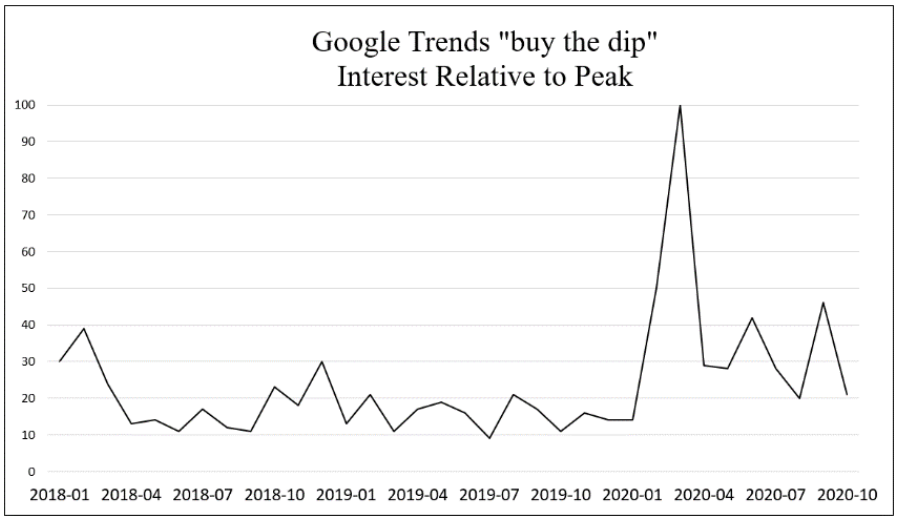

Figure 4

up their cash reserves and invested only when the market fell below a certain threshold?

In my study I construct several dip buying strategies and compare them to a dollar cost average (DCA) benchmark. They are performed on daily adjusted closing price data of the S&P 500 index taken from CRSP. The models are tested on two different time frames to explore the mid and long term expected returns: first, January 1, 2000 to October 23, 2020 and second, January 1, 1961 to October 23, 2020. The strategies and the benchmark are passive in that once cash is invested, it stays invested. The DCA rule invests $5 every day. The dip buying strategies invest only when the price of the S&P 500 falls a certain threshold below its 52-week high. I test both a 15% and 30% threshold. The buy decision rule at time t can be written as

$$\text{Buy when } P_t^{\star SP500} < (1 - Th) P_{52high}^{\star SP500}$$

Where $P_t^{\star SP500}$ is the price of the S&P 500 at time t, Th is either the 15% or 30% threshold, and $P_{52high}^{\star SP500}$ is the 52 week high price of the S&P 500. Daily returns for each model are calculated as

$$R_t = \left(\frac{V_t - V_{t-1}}{V_{t-1}}\right)$$

Where V_t is the cumulated value of all investments up to time period t and is mathematically shown as

$$V_t = R_t^{\star SP500}(V_{t-1} + I_t)$$

with $R_t^{\star SP500}$ representing the daily return of the S&P 500 index at time period t and I_t equal to the dollar investment amount at time period t.

For the Benchmark, I = 5 for each day period. For the dip buying strategies, I = 0 unless a buy decision rule is reached in which case I is equal to the sum of all forgone $5 daily investments since the last dip Investment. That is to say I = $5 multiplied by the total days passed since I was greater than 0. The dip buying strategy is constrained to buy only once every 30 days and any non-invested cash at the end of the period is added to the last day of the period to ensure that within each time period the dollar amounts invested in each model are equal. Because returns account for daily amounts invested, the dip buy models can suffer from extremely high variance, decreasing the integrity of calculations. To combat this, all model

values begin at $5,000 which decreases the effect individual investments have on the over-all variance of the returns.

Before doing the SPA, I ran one-tailed two-sample t-tests between the dip buy models and the benchmark in each time period resulting in 4 different hypothesis outputs. The null hypothesis is that the mean returns of the dip buying models are less than or equal to the mean returns of the DCA. The alternative hypothesis is that the mean returns of the dip buy models are greater than the mean returns of the benchmarks in each time period.

$$H_0: \mu_{m|t} - \mu_{bm|t} \leq 0 \qquad H_1: \mu_{\mathrm{m|t}} - \mu_{\mathrm{bm|t}} > 0$$

Where $\mu_{\mathrm{m|t}}$ are the mean returns of the models and $\mu_{bm|t}$ are the mean returns of the benchmark.

The results are summarized in Table 1 and 2. Table 1 shows the one-tailed t-test results for the 15% and 30% dip compared to the DCA benchmark results carried out during the shorter time period data. Although the means suggest the dip buying strategies may perform slightly better, we fail to reject the null hypothesis in both instances. The 15% dip vs the DCA returned a p-value of .47 and the 30% dip vs the DCA returned a p-value of .42.

Table 2 summarizes the result from the longer period data. Here too the dip models are far from significant. With p-values of .48 and .40 respectively we obviously can make no claim that the 15% dip and 30% dip outperformed the DCA during the long-term period. We therefore fail to reject the null hypotheses.

Table 1

One-tailed t Test on Dip Buy Models: Jan 1, 2000 - Oct 23, 2020

	15% Dip	*DCA*	*30% Dip*	*DCA*
Mean	0.00062	0.00060	0.00066	0.00060
Variance	0.00024	0.00016	0.00035	0.00016
Observations	5236	5236	5236	5236
Pooled Variance	0.00020		0.00025	
Hypothesized Mean Diff	0		0	
df	10470		10470	
t Stat	0.08196		0.19415	
P(T<=t) one-tail	0.46734		0.42303	
t Critical one-tail	1.64500		1.64500	

Table 2

One-tailed t Test on Dip Buy Models: Jan 1, 1961 - Oct 23, 2020

	15% Dip	*DCA*	*30% Dip*	*DCA*
Mean	0.00045	0.00044	0.00048	0.00044
Variance	0.00013	0.00011	0.00032	0.00011
Observations	15056	15056	15056	15056
Pooled Variance	0.00012		0.00021	
Hypothesized Mean Diff	0		0	
df	30110		30110	
t Stat	0.05582		0.24463	
P(T<=t) one-tail	0.47774		0.40337	
t Critical one-tail	1.64490		1.64490	

I next looked at the cumulative returns of the different strategies. Table 3 shows the historical returns from following the different strategies. These results are even less promising for the dip buying strategies. In both the long and short horizon data, it was the benchmark with the highest cumulative returns. Table 4 shows the model cumulative returns in terms of percentage outperformance of the benchmark.

Table 3

Cumulative Returns of Dip Buy Models and Benchmark

	Jan 1, 2000 - Oct 23, 2020	Jan 1, 1961 - Oct 23, 2020
DCA	$75,445.18	$1,639,200.04
15% Dip	$71,018.68	$1,572,244.58
30% Dip	$72,104.43	$1,412,509.54

Table 4

Outperformance of Dip Buy Models compared to Benchmark

	Jan 1, 2000 - Oct 23, 2020	Jan 1, 1961 - Oct 23, 2020
DCA	0.00%	0.00%
15% Dip	-5.87%	-4.08%
30% Dip	-4.43%	-13.83%

The naïve back test shows that the benchmark outperformed the dip buy models by more than 4% in every case – nearly 14% for the 30% dip over the long-term period. Again, these

results are not statistically significant, but they do make a sad case for the long-term dip buying strategy. We can at least historically conclude that rather than hording cash to invest at draw downs, investors would be better off investing in the market as soon as they can. These results align with the findings of Constantinides (1976) and Knight and Mandell (1992) who concluded that when presented with a lump sum of cash, it is optimal to invest the entirety of the sum immediately rather than slowly over time. For most people, of course, retirement must be built paycheck by paycheck.

In an effort to find a passive model which could outperform the benchmark I purposely engaged in data-snooping by tweaking and modifying strategies until they resulted in positive cumulative returns. I altered the time period of the returns, the percentage drawdowns of the strategies, and the strategies themselves. Over-all I probably tested 20 or so different strategies only 2 of which outperformed the benchmark. Rather than buying at dips these two passive strategies took a slightly momentum-based approach. Momentum model 1 bought when the price of the S&P 500 at time t was greater than the highest price from the previous 60 days. The second model bought when the price of the S&P 500 at time t was 1% greater than the highest price from the previous 60 days. That is

$$\text{Momentum model 1: Buy when } P_t^{\star SP500} > P_{60high}^{\star SP500}$$

$$\text{Momentum model 2: Buy when } P_t^{\star SP500} > (1.01\%)P_{60high}^{\star SP500}$$

Where $P_{60high}^{\star SP500}$ is the highest price of the S&P 500 over the past 60 days. These had the greatest success on the time period from Jan 1, 2000 to Oct 23, 2020. Table 5 contains the cumulative returns and t-test results.

The first momentum-based model produced $75,709.73 in cumulative dollar return, outpacing the benchmark by a mere .67%. Model 2 produced $80,127.89 in cumulative dollar returns outpacing the benchmark by 6.55%. The one-tailed p-value on Model 2 is .40, far from significant but closer than most of the dip buy models. In association with the positive return, this model seems to be heading the right direction. Now we're beginning to see how trying new rules and altering aspects of the models will eventually produce a strategy which outpaces the benchmark well enough to consider it significant.

Table 5

Cumulative Returns and Outperformance of Momentum Models to Benchmark

	Cumulative Return	Outperformance
DCA	$75,204.94	0.00%
Over 60 day high	$75,709.73	0.67%
1% Over 60 day high	$80,127.89	6.55%

Table 6

One-tailed t Test on Momentum Models: Jan 1, 2000 - Oct 23, 2020

	Over 60-day	*DCA*	*1% Over 60 day*	*DCA*
Mean	0.00062	0.00060	0.00067	0.00060
Variance	0.00020	0.00016	0.00033	0.00016
Observations	5236	5236	5236	5236
Pooled Variance	0.00018		0.00024	
Hypothesized Mean Diff	0		0	
df	10470		10470	
t Stat	0.08266		0.24691	
P(T<=t) one-tail	0.46706		0.40249	
t Critical one-tail	1.64500		1.64500	

Lastly, I used Sheppard's python library, ARCH (2020), which contains code for running the SPA on the model return losses. Below is a table summarizing the constant p-values generated from the SPA compared to the t-test p-values.

Table 7

One-tailed P-values compared to SPA P-values

	T-Test	SPA
Jan 1, 2000 - Oct 23, 2020		
15% Dip	0.46734	0.515
30% Dip	0.42303	
Jan 1, 2000 - Oct 23, 2020		
15% Dip	0.46734	
30% Dip	0.42303	0.612
Above 60-day High	0.46706	
1% Above 60-day High	0.40249	
Jan 1, 1961 - Oct 23, 2020		
15% Dip	0.47774	0.42
30% Dip	0.40337	

When running the SPA, model losses generated on the same data should be tested jointly. I performed the SPA on the dip-buy strategies for the shorter time period and then repeated the test with the addition of the momentum strategies. As you can see, the SPA p-value is penalized with the addition of the momentum strategies. The SPA p-value increases by nearly .1 when including the momentum strategies. In White's (2001) original paper he explains that "The difference between the naïve p-value and that of the Reality Check gives a direct estimate of the data-mining bias." Even with accounting for only a portion of the total models tested, the vast majority of the results suggest an expected positive data-mining bias as the SPA p-values are greater than the naïve p-values. The only result contrary to this is the case of the long period 15% dip strategy which suggests a negative data-mining bias. Its SPA p-value of .42 is less than the naïve p-value of .40. This anomaly would likely not exist if I reported more of the unsuccessful strategies tested on the long period data.

7. CONCLUSION

Data-snooping is a pervasive problem in both the finance industry and academia. Of classical statistical models used to correct for it, White's Reality Check and Hansen's Test of Superior Predictive Ability are paramount. Before implementing Hansen's SPA, I used two-sample t-tests to examine several potential passive investment strategies all of which were statistically no better than the benchmark and historically worse. I then unsuccessfully engaged in data-snooping as I searched for a model which would outperform the benchmark. This halfhearted data mining attempt was not the point of my study but adding the best performing models to the final SPA test proved interesting. The SPA results confirmed the t-test results with even larger p-values. In all cases but one, the passive investment models which I tested resulted in higher p-values from the SPA than from the two-sample t-tests. Including more models in the SPA resulted in a higher p-value. I failed to reject the null that investing large sums at both 30% and 15% drawdowns either underperforms or performs equally to the benchmark of investing small portion every day. Put simply, buying the dip as a passive long-term strategy is statistically either worse or equal to dollar cost averaging and is definitely worse historically.

8. CODE

Here is the Python code used to create a data frame of ten coin-flip outcomes performed 25,000 times. This code as well as the code and data used in the SPA test are available at https://github.com/jkbaldwin/datasnooping.git.

Coin Flip Simulation:

```
import random, pandas as pd
def coinFlip(participants):
   flipList = [0] * participants
   for i in range(participants):
      flipNumber = 10
      for j in range(flipNumber):
         flip = random.randint(0, 1)
         if (flip == 1):
            flipList[i] += 1
   return flipList
df = pd.DataFrame(coinFlip(25000));
df.plot.hist()
```

9. REFERENCES

ARCH (2020) available at: https://pypi.org/project/arch/ (accessed 28, October, 2020)

Aronson, David R. Evidence-Based Technical Analysis: Applying the Scientific Method and Statistical Inference to Trading Signals. John Wiley & Sons, 2007.

Bohannon J, "Reproducibility. Many psychology papers fail replication test", Science, vol 349, 2015, pp 910–911.

Diebold F, and Mariano R, 1995, ''Comparing Predictive Accuracy,'' Journal of Business and Economic Statistics, 13, 253-265.

Equbot (2019) Available at: https://equbot.com (accessed 29 September 2020).

Fricker R, Burke K, Han X, Woodall W, "Assessing the Statistical Analyses Used in Basic and Applied Social Psychology After Their p-Value Ban", The American Statistician, vol 73, 2019, pp 374-384.

George M. Constantinides, "A Note on the Suboptimality of Dollar-Cost Averaging as an Investment Policy", The Journal of Financial and Quantitative Analysis, vol 14, 1979, pp 443-450

Hansen, Peter R, and Asger Lunde, "A Forecast Comparison of Volatility Models: Does Anything Beat a Garch (1, 1)?" Journal of Applied Econometrics, vol 20, 2005, pp 873–889.

Hansen, Peter Reinhard, "A Test for Superior Predictive Ability." Journal of Business & Economic Statistics, vol 23, 2005, pp 365–80.

Knight John R, Mandell Lewis, "Nobody gains from dollar cost averaging analytical, numerical and empirical results", Financial Services Review, vol 2, 1992–1993, pp 51-61

Marte, Jonnelle. "Coronavirus Shifts U.S. Recession Debate from 'If' to 'What Shape'?" Reuters, Thomson Reuters, 11 Mar. 2020, www.reuters.com/article/us-health-coronavirus-usa-recession-idUSKBN20Y33B.

Politis D, Romano J, "The Stationary Bootstrap," Journal of the American Statistical Association, vol 89, no 428, 1994, pp 1303-1313

Preis T, Moat HS, and Stanley HE, "Quantifying trading behavior in financial markets using Google trends." Scientific Reports vol 3, 2013, p 1684.

Sheppard, Kevin, “Technical Trading: Fools Gold? Forecast Evaluation with Many Forecasts: The Econometrics of Predictability”, 2014, available at: https://www.kevinsheppard.com/files/teaching/mfe/advancedeconometrics/AFE_part_1.pdf U.S.

Smith, Gary, “Data mining fool’s gold,” Journal of Information Technology, vol 35, no 3, 2020, pp 182-194

Stanley TD, Doucouliagos H, Carter EC, “What Meta-Analyses Reveal About the Replicability of Psychological Research,” American Psychological Association, vol 144, no 12, 2018, pp 1325–1346

West K, 1994, ‘‘Asymptotic Inference About Predictive Ability,’’ University of Wisconsin Department of Economics Discussion Paper, 1996, ‘‘Asymptotic Inference About Predictive Ability,’’ Econometrica, 64, 1067-1084.

Woolston, Chris. “Psychology Journal Bans P Values.” Nature News, Nature Publishing Group, 26 Feb. 2015, www.nature.com/news/psychology-journal-bans-p-values-1.17001. (accessed 20 October 2020)

Wursthorn, Michael. “S&P 500 Sets First Record Since February, Erasing Its Coronavirus Plunge.” The Wall Street Journal, Dow Jones & Company, 18 Aug. 2020, www.wsj.com/articles/the-s-p-500-sets-first-record-since-february-erasing-its-coronavirus-plunge-11597781130.