The Technical Analysis Challenge

Patrick Burns*

7th October 2003

Abstract

We report on a study of the ability of analysts to distinguish an actual price series of an equity from random alternatives. Virtually all of the statistical tests on the results support the hypothesis that no skill was exhibited in selecting the correct response. Many of the analysts were extremely over-confident about their ability to select correct answers. The one area where it seems skill might have been exhibited is in the selection of correct answers that happened to be far from the random choices.

1 Introduction

A multiple choice test to investigate the efficacy of technical analysis was sponsored by Burns Statistics. One hundred price series were given, each containing 500 daily closing prices. Four possible extensions were presented for each series—one was the actual continuation of the series, the remaining three were randomly generated. Figure 1 shows an example of one series with its extensions. Submissions were accepted from 6 September 2003 through 4 October 2003. Both the test and its results are available on the Burns Statistics website.

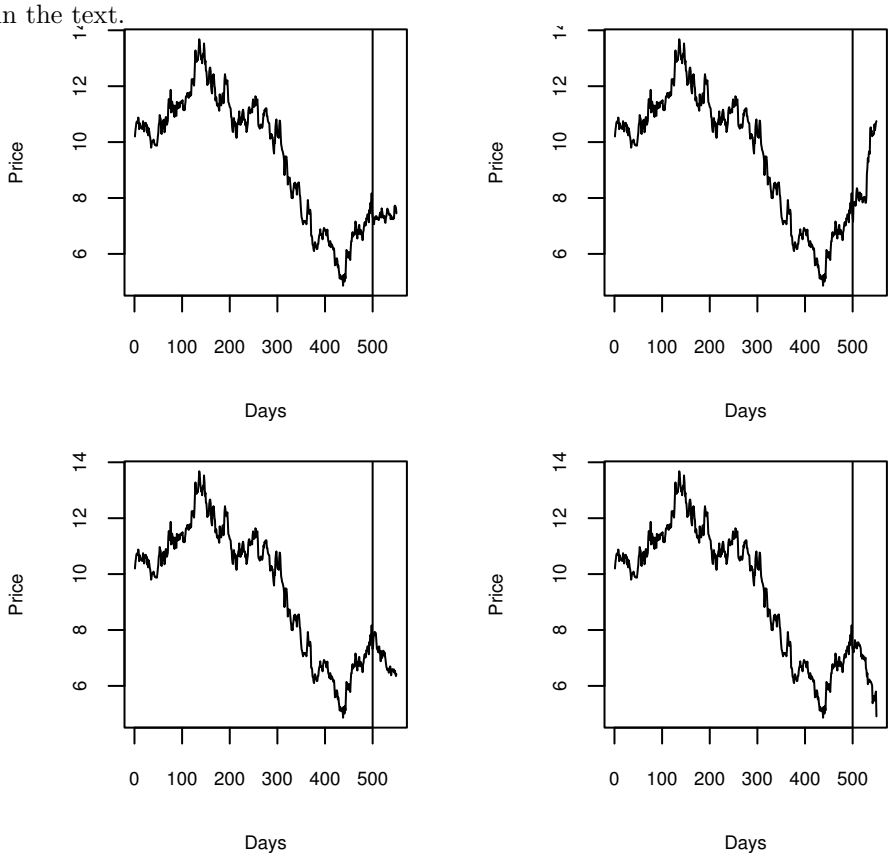
Participants were offered the choice between an anonymous submission and a public submission in which their identity would be revealed. The selection of problems to answer was at the discretion of the individual analyst, though a minimum of 10 answers was demanded of public submissions.

2 Overview of Submissions

There were a total of 19 submissions—a very small number given the number of people who looked at the study, and the even larger number who knew of it. During the study period there were over 2200 requests for the webpage describing the study, 1200 requests for the series plots, 600 for the series-plus-extensions plots, 500 for the extensions plots, and about 100 requests for the

*This report can be found in the working papers section of the Burns Statistics website http://www.burns-stat.com/.

Figure 1: Example series with its four extensions—the true extension is given in the text.



underlying data. Note, though, that there can be multiple requests from one person.

The response of quite a number of technical analysts was that there were not enough data for a meaningful analysis. A few would have been satisfied merely with a longer history. A large number wanted volume data. Some wanted open-high-low-close data as opposed to merely closing prices.

Of the 19 submissions, 2 are classified as "rogues" and are discussed in the next section. The remaining 17 are denoted "all TA" (all technical analysts) in tables of results. Another 4 submissions are classified as being given "under protest" meaning that they thought volume or some other set of data was needed for a meaningful analysis yet they still made submissions. Not all submissions specified a technique—these are accepted into the "all TA" group. Hence the "no protest" group consists of 13 submissions.

3 Rogue Techniques

While the intention was that participants use the same techniques that they would in practice, the artificial nature of the test allowed methods to be used that would not have a practical counterpart.

Unmasking

The most obvious of these is to try to unmask the identity of the series. This consists of writing a program that loops through stocks and times looking for a match of the pattern of returns. Once such a match has been found, then it is a trivial matter to identify the correct extension.

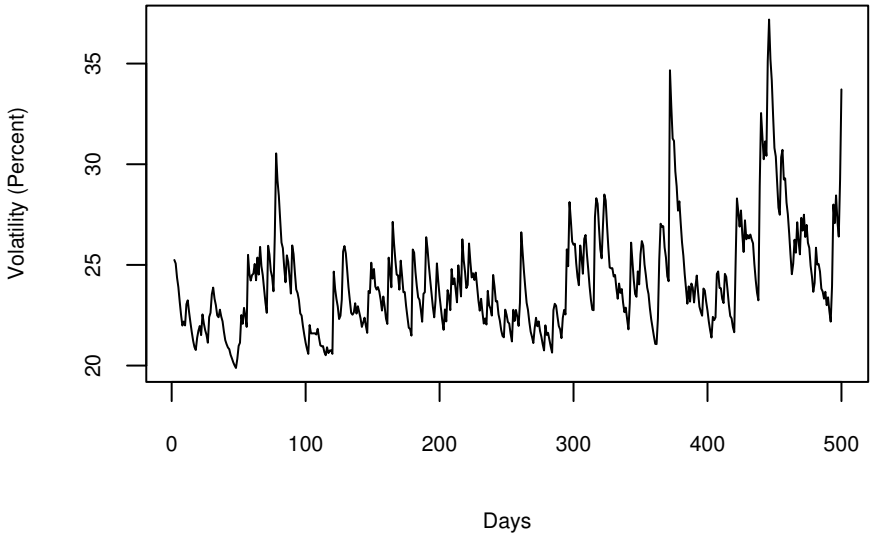
There was a submission by "Mr. D" that has all of the appearances of such a procedure. 89 answers were given and all were correct. The actual series were US equities—90 NYSE and 10 NASDAQ. Mr. D answered none of the NASDAQ and the remaining unanswered series was AOL.

Volatility Modeling

A less obvious method is to examine the volatility of the series. It is well-known that market data tend to have volatility clustering—periods of high volatility which gradually subside. GARCH models are a popular means of modeling market volatility. There are numerous variants, the first was [Engle, 1982]. An accessible explanation of the most popular GARCH models can be found in [Alexander, 2001]. Figure 2 shows a GARCH estimate of the volatility (annualized standard deviation) of the example series. The actual extension of the example series is in the lower left corner of the previous figure.

To use this knowledge of market data to select extensions, volatility models must be used for both the series and each of the extensions. Since GARCH models are designed to predict volatility, it is natural to predict the volatility over the extension period, and then compare that to whatever volatility model

Figure 2: GARCH estimate of volatility of the example series.



is created for each of the extensions. The main problem with this scheme is that there are very few observations with which to perform the estimates. Since volatility is an unobservable quantity that continuously changes, it takes a lot of data to get reasonable estimates—a minimum of 1000 daily observations is a recommendation for fitting GARCH models.

A part of the design of the study which was not made public was that the random extensions for half of the series were created by simulating GARCH models. The other half were merely collections of random returns with no attention paid to the volatility. Hence a method that tries to distinguish the true extension from the random ones via volatility should do better for the series that do not use GARCH simulations.

Dr. Allan White of the University of Birmingham made a submission based on volatility modeling. He expressed a lack of confidence, due at least partially to time constraints. To some degree his reservations were justified—his overall accuracy was just under 25%. His record was slightly better (30%) for the series not using GARCH simulations. He chose not to answer 7 of the "easy" series, and 12 of the "hard" series.

4 Test of the Number Correct

A key element of the design of the study was to have a straightforward statistical test for skill. If the analyst has no skill, then the distribution of correct answers will follow the binomial distribution with probability of success equal to 0.25. Under the alternative hypothesis that the analyst does have skill, the

Table 1: Binomial tests of the number of correct responses.

Analysts/Series	Successes	Answers	Fraction	p-value
all TA / all	257	979	.263	.19
no protest / all	194	764	.254	.41
all TA / garch	132	491	.269	.18
all TA / no garch	125	488	.256	.39
no protest / garch	100	382	.262	.32
no protest / no garch	94	382	.246	.59

distribution will be a binomial with probability of success greater than 0.25.

The p-value of such a test states the probability of getting data as good or better than we did assuming there is no skill. So a p-value near 1 means that there is no evidence of skill, a p-value near 0 means that there is evidence of skill. We can never have certainty of skill (or no skill)—we can only have varying degrees of evidence of skill.

The results from the submissions show no indication at all of being able to distinguish the true extension from random numbers. Tests for the number of correct responses are given in Table 1. As can be seen, the fraction correct in all of the tests is very close to 0.25. No individual did particularly well (with the possible exception of "Mr. D" who was classified as a rogue).

5 Confidence of the Analysts

A required field in the submissions was an estimate of the percentage of answers that the analyst thought were correct. Figure 3 shows the success rate predicted by the "no protest" group compared to the success actually achieved. Clearly the majority were over-confident.

Regardless of the level of skill, over-confidence is bound to be problematic. No doubt fundamental and quantitative analysts experience excessive optimism as well as technical analysts. It might be argued that over-confident people were more likely to make submissions to the study. That is true, but over-confident people are probably more likely to manage funds as well.

6 Distance Test

We've seen no evidence that the random series are distinguished from the real extensions. However, it is more important to be able to predict the direction of prices correctly. In practical terms selecting a random path that looks similar to the real path is almost as useful as selecting the real one. Though more ambiguous, a test of the proximity of the selected path to the real path is a fairer test.

There were 43 of the 100 series where two of the extensions moved up and two moved down. For these series we can produce a binomial test (with probability

Figure 3: Realized versus predicted success—the size of the symbol relates to the number of answers given.

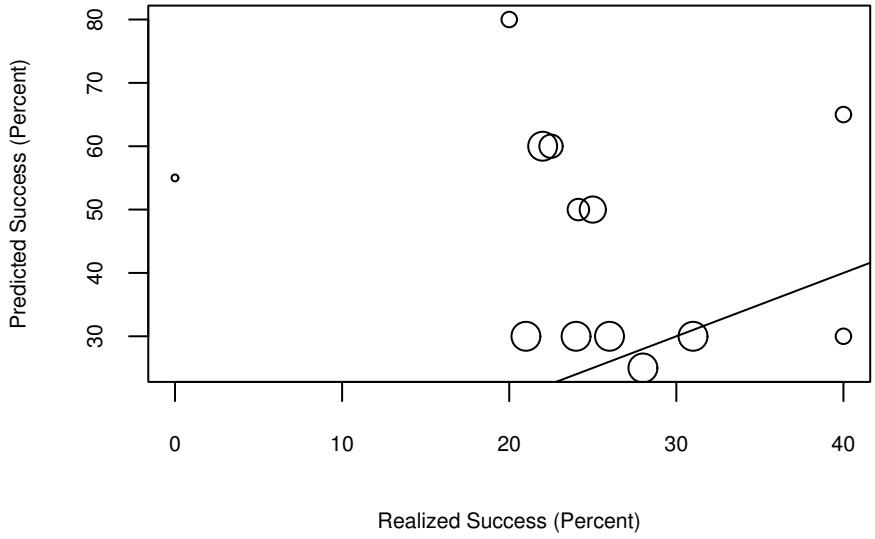


Table 2: Example computation for distance test.

	boak	hank	matt	shaw
return	0.106	-0.042	0.077	0.143
difference from true	0	-0.148	-0.029	0.037
absolute value	0	0.148	0.029	0.037
mean zero	-0.0535	0.0944	-0.0241	-0.0168
standardized distance	-0.951	1.680	-0.428	-0.300

0.5) that the selection has the same sign as the true answer. For the "all TA" group just under half of the selections are of the proper sign—211 out of 425. So there is no evidence that it is better than expected.

We can also test if the selections are close to the true answer. Specifically we test the distance of the logarithmic return over the 50 trading day extension of the selected answer from the true answer. The absolute difference in the log returns is found for the four extensions from the true extension, then these values are standardized to have mean 0 and variance 1. Table 2 shows the computations for the first series.

The statistic that is the sum of the standardized distances of the submitted answers will be approximately normally distributed with a known variance. If the statistic is close to zero, then that means the selected answers tend to be close to the average distance from the true answer. If the statistic is negative, that means the answers tend to be closer to the true answer than average—that

Table 3: Distance test results.

Analysts	**Number of answers**	**p-value**
all TA	979	.013
no protest	764	.060
protest	215	.036

Table 4: Binomial tests for the probability of selecting an answer that lands inside the range of answers for the series.

Analysts	Successes	Answers	Fraction	p-value
all TA	526	979	.537	.011
no protest	409	764	.535	.028
protest	117	215	.544	.11

is, evidence of skill.

Table 3 gives the results of distance tests. There is reasonable evidence that the analysts pick answers that have returns closer to the truth than random selections. This seems to be generally true—the effect is not concentrated in the GARCH or non-GARCH series, nor in the extreme series that are discussed later.

However, there is an alternative explanation for selecting answers with better than average distance. The analysts may tend to select answers that fall in the middle of the choices in preference to the answers that end highest or lowest. A binomial test with probability 0.5 can be used. Table 4 reports this test. The analysts do favor answers from the middle, which at least partially explains the good performance in terms of the standardized distances.

From a simulation in which the middle values are selected with probability 0.537, the p-value of 0.013 in the test of standardized distances should be about 0.06. That is, there is not much evidence that the analysts are really doing better than expected. The lack of evidence of getting the direction of the price moves right adds weight to the argument that the analysts are not exhibiting skill in this regard.

7 Test of Extremes

Another test is to see if analysts can effectively select the true extension if it is significantly far from the three random extensions. There were 7 series identified in which this occurred: E008, E011, E014, E019, E028, E031 and E041. Since this was done in terms of difference in returns, there is some selection bias for more volatile series.

Table 5 shows the results when we restrict the binomial test for successfully identifying the true extension to these 7 series. Though the tests are far from statistically significant with the small sample size, the fraction correct is above 25%. Most of the success is in the first 3 of the series, however.

Table 5: Test of correct selection when random extensions end far away.

Analysts	Successes	Answers	Fraction	p-value
all TA	24	74	.324	.092
no protest	18	59	.305	.20

If the tests had been significant, another explanation would have been that the analysts tended to pick an answer that was different from all of the rest. This seems not to be the case—12 series were found where a random answer was far from all of the others. The series and the extreme wrong answers are: E024, exel; E027, puma; E032, dott; E045, mike; E046, gold; E059, yano; E060, wayn; E061, serg; E066, poem; E069, fire; E076, cher; E091, rhye.

The analysts selected the extreme wrong answers only about 17% of the time—with a p-value of about 0.03 for having a probability of selection less than 25%. There is a significant difference between the fraction of extreme right answers and extreme wrong answers selected. The p-value for the difference (via a normal approximation) is 0.0081. The small number of series on which this result rests makes the inference of a real effect tenuous.

8 Discussion

There is no evidence at all that the analysts could generally distinguish random series from the true price series. Additionally, the analysts were in general hugely overconfident of their skill at doing so. While the submissions apparently did well in terms of picking extensions that ended closer to the true extension than should be expected, this is explained by a tendency to pick extensions that end inside the range of endings. Technical analysts might hold out hope for the effect seen with extreme answers.

A person who did not make a submission is Max Danzig, who trades full-time for his own account. He studied technical analysis intensely, but decided to move towards statistical models some time ago when he saw market patterns in the way that his wife had hung Christmas lights. There is general agreement that price series do have some predictability, but the patterns exhibited are very subtle. The human eye is unlikely to be the best tool to find the patterns.

Further research is to provide a test with more data available—volume for example. Burns Statistics is planning such a test.

References

[Alexander, 2001] Alexander, C. (2001). *Market Models: A Guide to Financial Data Analysis*. John Wiley & Sons.

[Engle, 1982] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of UK inflation. *Econometrica*, 50:987–1008.