Permuting Super Bowl Theory

Patrick Burns[*]

2nd January 2004

Abstract

The quality of stock market predictions based on the winner of the Super Bowl is examined using permutation tests. These tests are very easy to perform in modern computing environments like the R language. One key point that comes to light is that the success rate of a prediction is not a good measure of its usefulness. Statistically significant success in prediction does not automatically lead to economically profitable strategies.

1 Introduction

Super Bowl theory is one of the better known and longer lasting market prediction tools. The theory says that the U.S. market will end the year up if a "National" team (as opposed to an "American" team) wins the Super Bowl—the championship game in professional American football. The theory's fame, of course, derives from the disparity between its seemingly remarkable record and the lack of any logical connection.

An article in the New York Times in 1978 by Leonard Koppett apparently was the birth of the idea. It gained currency as its winning streak continued.

There is a claim that the theory has been correct 30 times out of 37. However, that record defines "the market going up" as at least two of the Dow Jones Industrial Average, the S&P 500 and the New York Stock Exchange Composite going up. This definition of the market seems suspiciously like data snooping. While these days some enlightenment has been achieved and the market is often thought of as the S&P 500, for most of the Super Bowl's history—and certainly when the theory was put forward—the "market" meant the Dow Industrials.

There is also some hocus-pocus about what constitutes a "National" team. The current National Football League is the result of a merger of two leagues. In the theory a team is "National" if it was a member of the pre-merger National Football League, and it is "American" if it was a member of the American Football League. A more straightforward definition would be whether the team

[*] This report can be found in the working papers section of the Burns Statistics website http://www.burns-stat.com/. Input from Matthew Dowle is gratefully acknowledged.

at the time of the game comes from the National Conference or the American Conference. It is left to the historians to determine if this is after-the-fact theory migration or was a part of the original.

By the way, this is American-style football. It is not the game that the rest of the world calls football, which Americans call soccer.

2 Super Bowl Analysis

Our analysis uses the Dow Jones Industrial Average as the market. The genealogy of the data for the winner of the game is uncertain—it might be the actual conference that the team was in at the time of the game, but that is supposition only. In any event, the data used in the analysis are listed in Table 1.

For those who want to change the data to their own definitions, there is computer code freely available on the Burns Statistics website. The code is written in the R language [R Development Core Team, 2003] which can be downloaded for free via http://www.r-project.org.

The task of the analysis is to see just how good of a predictor the Super Bowl is. The theory—given the data of Table 1—is correct 25 out of 37 times. While substantially far from the best claims for the theory, this is still an ostensibly impressive 68% success rate.

The analysis consists of doing a permutation test—this mixes up the order of the winners (and/or the market results) numerous times. With each random ordering of the winners we note how many predictions are correct. After a set number of reorderings—1000 are used here—the results are tabulated. The fraction of random orderings that have the number of correct predictions equal to or greater than what actually happened is noted as the p-value.

The p-value says how likely it is that we would see what we did merely by chance. More specifically the p-value is the probability of observing something as extreme or more so if there really is no predictability. A p-value of 1 would mean that there is no chance of predictability, a p-value of 0 would mean that there is predictability without any doubt. The closer a p-value is to 0, the more evidence there is that something real is happening.

What is the permutation test really doing? In a test we need a "null hypothesis", a statistic, and the distribution of the statistic assuming the null hypothesis is true. The p-value depends on the location within this distribution of the statistic computed on the data.

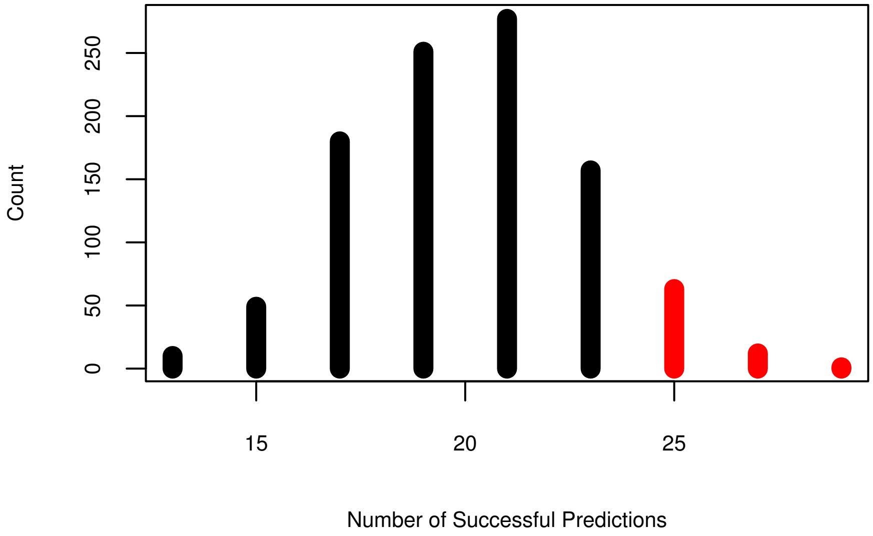
In our case the null hypothesis is that there is no predictability, and the statistic is the number of successful predictions. The exact distribution of interest (given the number of up years and of National wins) would tally the statistic for each permutation of the winners. There is an extraordinarily large number of such permutations so we merely take a random sample of permutations. We don't need acute accuracy—1000 random permutations is, in general, quite sufficient.

Figure 1 shows the distribution of the number of correct predictions in the randomly permuted data. The p-value, which is slightly less than 7%, is the

Table 1: Super Bowl data and logarithmic returns for the Dow Industrials.

Year	**Winner**	**Dow Industrials (%)**	**Theory**
1967	National	+14.1	correct
1968	National	+4.2	correct
1969	American	-16.5	correct
1970	American	+4.7	wrong
1971	American	+5.9	wrong
1972	National	+13.6	correct
1973	American	-18.1	correct
1974	American	-32.4	correct
1975	American	+32.4	wrong
1976	American	+16.4	wrong
1977	American	-19.0	correct
1978	National	-3.2	wrong
1979	American	+4.1	wrong
1980	American	+13.9	wrong
1981	American	-9.7	correct
1982	National	+17.9	correct
1983	National	+18.5	correct
1984	American	-3.8	correct
1985	National	+24.4	correct
1986	National	+20.4	correct
1987	National	+2.2	correct
1988	National	+11.2	correct
1989	National	+23.9	correct
1990	National	-4.4	wrong
1991	National	+18.5	correct
1992	National	+4.1	correct
1993	National	+12.9	correct
1994	National	+2.1	correct
1995	National	+28.9	correct
1996	National	+23.1	correct
1997	National	+20.4	correct
1998	American	+14.9	wrong
1999	American	+22.5	wrong
2000	National	-6.4	wrong
2001	National	-7.4	wrong
2002	American	-18.3	correct
2003	National	+22.6	correct

Figure 1: Distribution from the permutation test of Super Bowl wins. The critical area is 25 and above.



fraction of the distribution that is 25 or more successes. While 7% is not large, it is seldom considered statistically significant.

3 Alternative Bowl Games

The Nursery Bowl

If the NFC played the Timbuktu Nursery School every year, then—assuming they never got overconfident and let the toddlers win—the prediction would have been right 26 out of 37 years. So the Nursery Bowl is an even better predictor than the Super Bowl.

Or is it?

We can do a permutation test again to find out. In this case the test can be done in our head because all permutations give the same result of 26 correct. The p-value for the permutation test is equal to 1. This is a gratifying result—the prediction is silly and the test says so.

The Solar Bowl

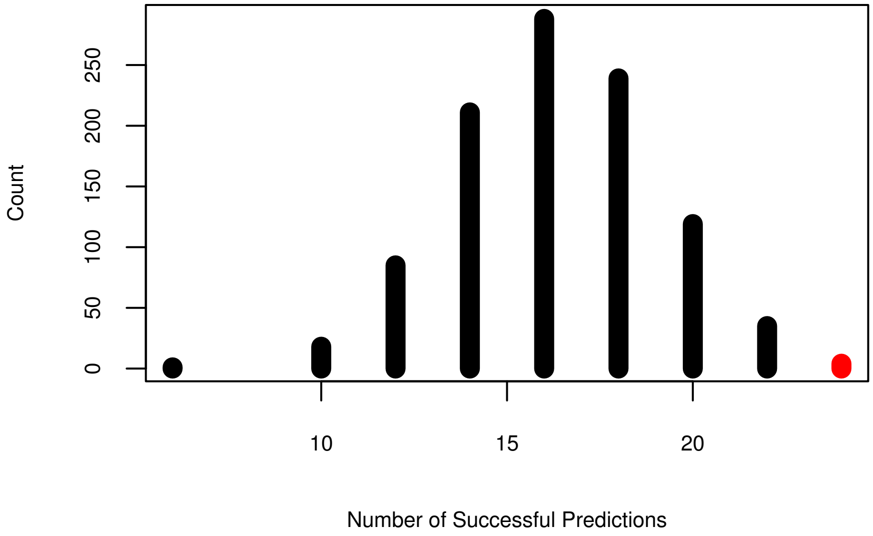
Consider a theory where the market is predicted to go up if the Venusians beat the Martians in the Solar Bowl. The history of the game is given in Table 2.

The number of successful predictions by the Solar Bowl is one worse than the Super Bowl—24 out of 37. So naively we expect that this is a worse predictor

Table 2: Solar Bowl data.

Year	**Winner**	**Dow Industrials**	**Theory**
1967	Venusian	+	correct
1968	Venusian	+	correct
1969	Martian	-	correct
1970	Martian	+	wrong
1971	Venusian	+	correct
1972	Martian	+	wrong
1973	Martian	-	correct
1974	Martian	-	correct
1975	Martian	+	wrong
1976	Martian	+	wrong
1977	Martian	-	correct
1978	Martian	-	correct
1979	Martian	+	wrong
1980	Venusian	+	correct
1981	Martian	-	correct
1982	Venusian	+	correct
1983	Venusian	+	correct
1984	Martian	-	correct
1985	Venusian	+	correct
1986	Venusian	+	correct
1987	Martian	+	wrong
1988	Venusian	+	correct
1989	Venusian	+	correct
1990	Martian	-	correct
1991	Martian	+	wrong
1992	Martian	+	wrong
1993	Venusian	+	correct
1994	Venusian	+	correct
1995	Martian	+	wrong
1996	Martian	+	wrong
1997	Venusian	+	correct
1998	Martian	+	wrong
1999	Martian	+	wrong
2000	Martian	-	correct
2001	Martian	-	correct
2002	Martian	-	correct
2003	Martian	+	wrong

Figure 2: Distribution from the permutation test for Solar Bowl wins. The critical area is 24 and above.



than the Super Bowl. However, Figure 2 shows this to be a very good predictor—the p-value is about 0.3% which is quite significant. An important fact about the Solar Bowl is that whenever the Venusians win, the market goes up.

4 In-sample, Out-of-sample

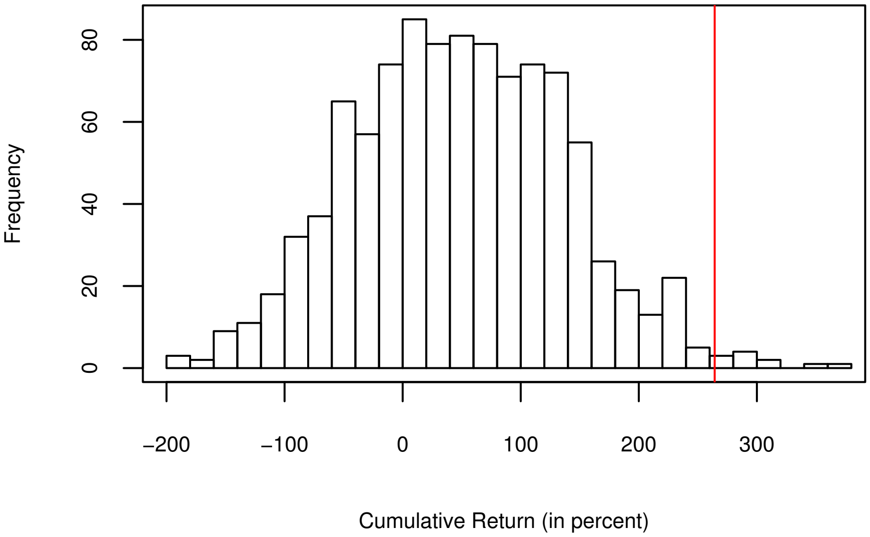
When testing a theory, the data that gave rise to it should not be used to confirm it. In statistical parlance the data that suggest a theory are called "in-sample". Since Super Bowl theory seems to have been proposed in 1978, we assume that the first eleven years (1967 through 1977) are the in-sample period, and that the subsequent data are out-of-sample. Using out-of-sample data only, there are 18 correct predictions in 26 years, so 69% correct. However, the p-value from the permutation test is about 27%—decidedly not significant.

What evidence was there for the theory when it was first proposed? The theory was right 7 out of 11 years or about 64% of the time. The p-value for the permutation test is about 21%. Compared to the out-of-sample period, it has a slightly better p-value with a slightly worse success rate and fewer observations.

There is a major qualitative difference in the two periods. In both periods the market usually went up. In the in-sample period the market always went up the few times that the National team won (similar to the Solar Bowl). National teams have usually won during the out-of-sample period (similar to the Nursery Bowl).

In general it should be a cause for concern when out-of-sample behavior is

Figure 3: Distribution from the permutation test of returns from the long-short strategy on Super Bowl wins (using both in-sample and out-of-sample data).



different than the in-sample behavior.

5 Super Returns

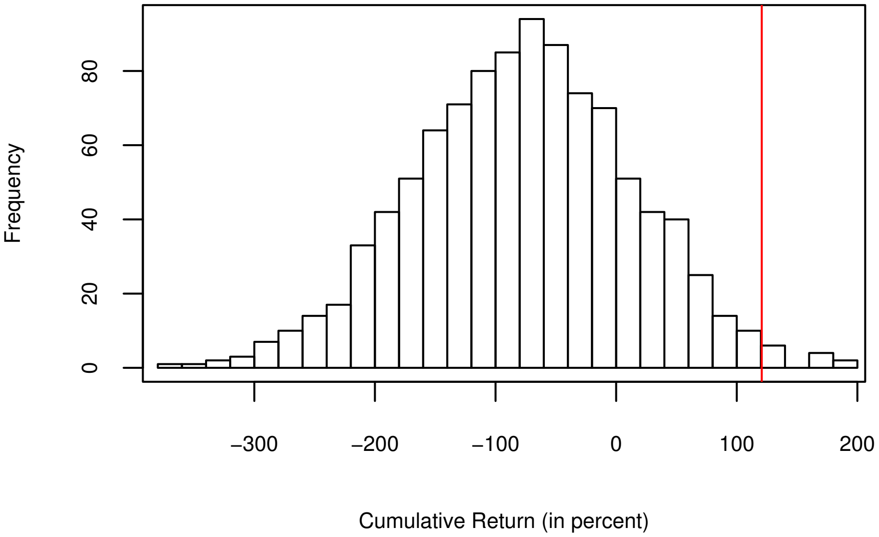
To this point we have focused on the market going up or down. What really matters, though, is how much up or down it goes. We can perform permutation tests using the market returns, rather than just the sign of the returns.

In order to do such a test, we need to decide how to use the theory. The strategy that we investigate invests in the Dow Jones if the National team wins (we get the return from the Dow Jones for the year), and goes short the Dow Jones if the American team wins (we get -1 times the Dow Jones return). Logarithmic returns are used since they can be added over time periods. The permutation test on the full dataset for this scenario gives a p-value of about 1%, a result that is traditionally considered significant. Figure 3 shows the permutation distribution in this case. While the test is significant in the statistical sense, the return achieved by the strategy is only marginally better than always holding the index. The strategy using the Super Bowl provides a cumulative return of 264% and the buy and hold strategy produces 259%.

Of course in-sample data should be excluded from tests. The test on only the out-of-sample data yields a p-value of about 7%.

A test of Solar Bowl returns—shown in Figure 4—has about the same significance as the corresponding test for the Super Bowl. However, the economic

Figure 4: Distribution from the permutation test of returns from the long-short strategy on Solar Bowl wins.



value of the strategy is quite poor. There are two reasons for this:

- While the Super Bowl has tended to correctly predict returns with large magnitude, the Solar Bowl is not especially good at getting the large returns right.
- The Solar Bowl strategy is short the market most of the time, but the market goes up most years.

There may be profitable ways of using the predictability of the Solar Bowl, but this strategy is not one of them.

6 Discussion

Permutation tests are an easy and transparent means of testing predictors. They are especially useful given that the fraction of correct predictions need not be a good indicator of the quality of the predictor. The minimal number of assumptions of permutation tests means their results are trustworthy.

Data have been presented as either in-sample or out-of-sample. Reality is more subtle and problematic—“out-of-sample” is often a matter of degree. The Super Bowl data we’ve stated to be out-of-sample have influenced our decision to perform the data analysis. To be truly out-of-sample, the data must be observed after the decision to perform a specific analysis. In finance this virtually always means future data. There are techniques that can help—see for instance [White, 2000] and [Sullivan et al., 1999].

While the analysis of the Super Bowl data does not yield p-values that are small enough to be typically accepted as significant, what if it did? Would there be a lot of money invested on the basis of the relative frequency of a leather-encased packet of air passing two lines 91.44 meters apart? Probably not.

If we had the same significant results but it involved interest rates or economic growth, then would a lot of money be invested? Probably.

What's the difference?

The difference is that people are perfectly willing to believe that economic variables may have predictability for stock markets, but sporting variables are given close to zero credibility. It would take overwhelming evidence to get most people to invest based on the outcome of a ball game. To put it into statistical jargon, investors act like Bayesians—results from data are tempered by prior beliefs.

References

[R Development Core Team, 2003] R Development Core Team (2003). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3.

[Sullivan et al., 1999] Sullivan, R., Timmerman, A., and White, H. (1999). Data snooping, technical trading rule performance, and the bootstrap. *Journal of Finance*, 54:1647–1692.

[White, 2000] White, H. (2000). A reality check for data snooping. *Econometrica*, 69:1097–1127.