System evaluation based on past performance: Random Signals Test

Alex Strashny
24 May 2002
Affiliation: University of California, Irvine. Graduate student in Economics.

Contact

- E-mail: [astrashn@uci.edu]
- Phone: 949-
- Mail: Alex Strashny, Department of Economics, University of California – Irvine, 3151 Social Science Plaza A, Irvine, CA 92697-5100
- Web: [http://www.ags.uci.edu/~astrashn/]

JEL: G1, C12
Keywords: system evaluation, hypothesis testing, trading

Abstract

This paper introduces a new method for evaluating a trading system based on its past performance. The method is a hypothesis test that asks whether the system is making random trades. The test controls for price behavior during the test period and the trade characteristics of the system being tested. A system should be traded only if the null hypothesis of random trading is rejected.

Introduction

Many traders use concrete trading rules, called *system*s, to make trading decisions. The performance of a system depends on both (1) the merits of the system itself and (2) market conditions under which it is used. A system might perform well historically only because it is well suited to specific market conditions during the test period. For example, the buy-and-hold strategy works extremely well when the price is going up; swing trading works well when the price is in a range.

A system might achieve good performance simply through random trading. However, such a system should not be traded because in the long run random trading has poor performance. The procedure described in this paper tests whether a system can be distinguished from random trading. It is called the *Random Signals Test* because it is a hypothesis test based on randomly generated trading signals. A system should only be traded if its performance is so high that the null hypothesis of random trading is rejected.

The hypothesis test is based on a performance measure that describes trader preferences. An example of such a performance measure is Rate of Return. First, we construct the probability distribution of a performance measure under the null hypothesis of random trading. The distribution is constructed through randomly issued trading signals. Then, based on this distribution, we find the appropriate critical value. If a system's performance is greater than this critical value, we reject the null hypothesis of random trading.

The Random Signals Test controls for price behavior during the test period and so the results of the test are not influenced by it. We control for price behavior by comparing a system's performance to the performance of random trading on the same price data.

The test also controls for the trade characteristics of the system being tested. The trade characteristics of random trading are set equal to those of the system being tested. For example, if the system being tested trades a variable number of contracts, so should random trading. We do this so that performance is the only potential distinguishing factor between the system being tested and random trading.

Hypothesis Test

We evaluate a system by asking whether its high past performance can be achieved by random trading with a reasonably high probability. This question is formally viewed as a hypothesis test. The null hypothesis is that the system in question is making random trades. If this hypothesis cannot be rejected, the system should not be traded. The alternate hypothesis is that, since the system's performance is so high, the trades it makes are not random. A system should be traded only if the null is rejected in favor of the alternate.

To perform this test, we must know the probability distribution of a performance measure under the null hypothesis of random trading; call it the *performance distribution* for short. Based on this performance distribution, we calculate a critical value. If c is this critical value, the hypothesis test is

- H_0: System is bad. System's performance is indistinguishable from performance of random trading. $Performance(System) \leq c$.
- H_A: System is good. System's performance is better than performance of random trading. $Performance(System) > c$.

Distribution of Performance Under the Null Hypothesis

Define the *random system* as a series of random trades on the same price series that is used to calculate the performance of the system being tested. The performance from one run of the random system is one draw from the distribution under the null hypothesis of random trading. Make an arbitrarily large number of such draws thus reconstructing the performance distribution.

Distribution of random system's trades

The trades issued by the random system are randomly picked from some distribution. This distribution should match closely the distribution of trades issued by the system being tested. This is so that the random system and the system being tested could not be distinguished solely on the types of trades they issue; performance is the only potential distinguishing factor.

All trades are defined by three *trade characteristics*, namely

1. Number of contracts (for example, -2 for short two contracts);
2. Transaction cost (including commission, slippage, etc.); and
3. Trade duration.

This paper assumes that the three trade characteristics are independent of each other, except that if a trade is of one type (short, flat, or long), then the next trade has to be of a different type. A trader can model trade characteristics by introducing other dependencies as well. For example, in another model, trade duration for flat trades could on average be different than trade duration for long or short trades.

Estimating distribution of trade characteristics of the system being tested
Estimate the distribution for number of contracts simply from observed probabilities. For example, if there are 40 trades in all and 15 of them are long one contract, the probability of being long one contract is $^{15}\!/_{40} = 37.5\%$.

Assume that transaction cost has a normal distribution and estimate its mean and standard deviation. Since trade duration can only be positive, assume that it has a truncated normal distribution. Also, draws from this truncated normal have to be rounded to the nearest integer.

This approach for estimating the distribution of trade characteristics only works if the system being tested makes enough trades. As a rule of thumb, only apply this approach if the number of trades is *at least* 30.

Alternatively, we can estimate the distribution of trade characteristics using Bayesian estimation. In this approach, the empirical probabilities are combined with prior beliefs about the system. The approach is more involved but is particularly suited for situations in which the system being tested does not make many trades.

Performance Measures
Using the above technique, a trader can construct the performance distribution based on any performance measure. The trader should choose the performance measure that best reflects his preferences. As an example, consider Rate of Return.

Rate of Return
The most basic performance measure is the Rate of Return. With some types of trading instruments, such as futures, the actual amount of money that must be committed to trading is not clearly fixed. It is therefore convenient to use profit divided by some multiple of *maximum drawdown* (the largest decline in equity) as a proxy for Rate of Return. This paper uses profit divided by three times the maximum drawdown. Since maximum drawdown is a measure of risk, this calculation for Rate of Return can also be viewed as risk-adjusted profit. To facilitate comparisons, all figures given in this paper are annualized.

Custom Performance Measures
A trader can combine several existing performance measures into one to better describe his preferences. Here are some measures, in addition to Rate of Return, that traders might want to use:

- Time to recovery. The time from the beginning of a drawdown to the point at which the amount of money in the portfolio is recovered. Can use either the average time to recovery or some other value, such as a high percentile.
- Percent of winning trades. The number of profitable trades divided by the total number of trades (excluding flats).
- Maximum number of consecutive losing trades.

For example, if *ROR* is Rate of Return and *ATR* is Average Time to Recovery, performance measure *U* that best describes a particular trader's preferences might be:

$$U = 1.0 * \log ROR - 0.5 * \log ATR \qquad (1).$$

This performance measure basically says that the trader remains indifferent if the Rate of Return increases by 5% while Average Time to Recovery increases by 10%.[1]

Example: S&P 500 Futures

As an example, we do the Random Signals Test for a hypothetical system that trades on S&P 500 March 2002 futures. The data used here is daily closes from 18 March 2000 to 15 March 2002 – 498 trading days in all. Figure 1 shows the price series as well as the signals issued by this hypothetical system.

[** Figure 1**]

The hypothetical system produces a profit of 543 points; its maximum drawdown is 126 points. Therefore, its (annualized) Rate of Return is 72%.

The estimated trade characteristics for this system are given in Table 1. The estimation is done by informally combining empirical probabilities with prior beliefs about the system. Transaction cost is roughly based on Wolff's [2002] slippage estimates. According to Wolff, average slippage in the S&P futures market is between about $140 and $230 per contract round turn. One point in this market is $250.

[** Table 1 **]

We use the estimated trade characteristics from Table 1 to run the random system on the price series. The random system is run 100,000 times. After each run, calculate the Rate of Return. Use these calculations to construct the distribution of Rate of Return under the null hypothesis of random trading. This distribution is shown in Figure 2.

[** Figure 2 **]

The 95th percentile of this performance distribution is 39.8%. This means that the probability that random trading can achieve a Rate of Return of 39.8% or higher is only 5%. Based on this distribution, the p-value of our hypothetical system is 0.8%.[2] When only one system is being tested, as is the case here, the p-value is the probability that, given some performance measure, the system is indistinguishable from random trading.

Critical Values

The level of significance is the probability of rejecting the null hypothesis when it is true. In our case, it's the probability that the Random Signals Test mistakenly picks a system when it shouldn't. If the Random Signals Test is used to test just one system, then the critical value of 5% significance is just the 95th percentile of a performance distribution.

However, a trader could be data mining – he could run the hypothesis test on several different systems and then pick the system that passes the test. In this case, the conventional critical values cannot be used. Intuitively, if there are 100 systems that trade randomly, about 5 of them will have performance greater than the 95th percentile of a performance distribution. Thus, if 100 random systems are tested, about 5 will be picked with conventional critical values, even though none should be picked.

If all the systems under consideration are *a priori* independent of each other, the correct percentiles for critical values are easily calculated from the definition of level of significance.[3] Table 2 shows the percentiles for critical values for selected levels of significance α and numbers of systems *N*. For example, if 10 systems are being tested, a system should only be picked at the 5% significance level if its performance is greater than the 99.49th percentile of the performance distribution.

alpha

[** Table 2 **]

Because the percentiles of interest are so large and so close together, a lot of runs of the random system are needed to obtain accurate critical values. In this paper, the random system is run 100,000 times.

Table 3 shows the appropriate critical values for S&P 500 March 2002 futures based on the performance distribution in Figure 2 and percentiles in Table 2. Table 3 states that, for example, if 10 systems are tested, a system should only be picked at the 5% significance level if its Rate of Return is greater than 81.4%.

[** Table 3 **]

In our example, the Rate of Return (72%) is greater than all the Rates of Return in the $N = 1$ column. This means that, assuming we're testing only this one system, the Random Signals Test rejects the null hypothesis of random trading even at the 1% significance level.

Curve-Fitting

Traders often define a system in terms of a set of parameters and then find the set that maximizes past performance. For example, a trader dealing with the moving average crossover system might find the look-back period that maximizes historical Rate of Return. This is known as *curve-fitting* or *optimization*.

For the purpose of finding percentiles for critical values, each set of parameters defines a different system. However, the performances of these different systems are *a priori* correlated. For example, the performance of the moving average crossover system with the look-back period of 20 days is very similar to the performance of the system with the look-back period of 21 days.

Because of this correlation, the formula for percentiles mentioned above is no longer valid. If the performance of all the systems is independent, then N in Table 2 should be set to the number of systems. If the performance of all the systems is identical, then it does not matter how many systems there are and N should be set to 1. If the performance across systems is *a priori* positively correlated, as is the case with curve-fitting, N should be somewhere between 1 and the number of parameter sets considered. However, the exact value of N is unknown and depends on the system being tested.

We can approximate N by counting the number of clusters of parameter sets that *a priori* (1) have similar performance within the cluster, (2) have approximately independent performance across clusters. For example, consider the moving average crossover system. A trader who is testing it for look-back periods of 10 to 50 days might believe that there are two such clusters: from 10 to about 30 days and from about 30 to 50 days. That is, for example, he believes that the performance of the system with look-back period of 35 days is

1. Highly positively correlated to the performance of the system with look-back period of 40 days; but
2. Independent of the performance of the system with look-back period of 20 days.

In this case, the trader sets $N = 2$.

Multiple Markets

A trader might also test a system on several different price series and pick it if the null hypothesis of random trading is rejected on at least one of these series. In this case, the derivation of the formula for percentiles is the same as discussed above, with N now being the number of price series on which the system is tested. As long as all N price series are independent of each other, performances on each of these series are also independent and the above derivation applies.

Conclusion

This paper develops a way to decide if, based on a system's past performance, the system should be traded. The decision is made through a basic hypothesis test that compares the system's past performance to a critical value. The key is that the critical value is conditional on the price series and the system's trade characteristics. In this way, the test considers what performance could have been achieved given what the price actually did.

The Random Signals Test is crucial to traders. Traders come across many systems with good past performance. However, if a system's good past performance is less than some critical value, then there is a high probability that the performance is the result of chance, and so the system should not be traded.

Acknowledgments
Thanks to Mark Simms, Andreas Krysl, and Rick Luppy for helpful comments.

Endnotes

[1] If Rate of Return is zero or negative, performance measure *U*, as written, is meaningless. However, as described in the next section, traders are only interested in a high percentile of the performance distribution, not in the distribution itself. Thus, when Rate of Return is less than some very small positive value (such as 0.01 or 1%), set *U* to some very negative number (such as -5).

If *ROR* and *ATR* change in such a way that performance *U* remains constant, then the trader is indifferent to this change. At constant *U*, differentiating equation (1) gives

$$1.0 * \frac{dROR}{ROR} = 0.5 * \frac{dATR}{ATR} \qquad (2).$$

Since dX/X is approximately the percent change in *X*, at constant *U*

$$\%\Delta ROR \approx \tfrac{1}{2}\%\Delta ATR \qquad (3).$$

Delta

This means that the trader remains almost indifferent if Rate of Return increases by some percent while at the same time Average Time to Recovery increases by twice as many percent.

[2] The 95th percentile of profit distribution under the null hypothesis is 433 points. Based on profit, the system's p-value is 1.9%.

[3] Let a be the desired level of significance, $Perf_i$ the performance of the i^{th} system, *N* the total number of systems being tested, *c* the appropriate critical value, and *p* the corresponding percentile of a performance distribution. Then, from the definition of level of significance,

alpha

$$\begin{aligned} a &\equiv \Pr(\text{reject } H_0 \mid H_0 \text{ is true}) \\ &= \Pr(Perf_1 > c, \text{OR } Perf_2 > c, \ldots, \text{OR } Perf_N > c \mid H_0) \\ &= 1 - \Pr(Perf_1 \le c, \text{AND } Perf_2 \le c, \ldots, \text{AND } Perf_N \le c \mid H_0) \\ &= 1 - [\Pr(Perf_i \le c \mid H_0)]^N \\ a &= 1 - (p/100)^N \end{aligned} \qquad (4).$$

alpha

The transition from line 3 to line 4 is from the assumption that the *N* systems are independent of each other. From line 4 to line 5, the probability that a random variable is less than some value is the percentile (divided by 100) corresponding to that value.

Solving for the percentile *p*,

$$p = 100 * (1 - a)^{1/N} \qquad (5).$$

alpha

For a given level of significance a, as the number of systems *N* increases, so does the appropriate percentile *p*.

Parameter	Distribution
Number of Contracts	0 with probability 40% -1 or 1 with probability 30% each
Trade Duration	Normally distributed with mean 45 days and standard deviation 25 days
Transaction Cost	Normally distributed with mean 0.75 points and standard deviation 0.25 points

Table 1: Estimated trade characteristics for the hypothetical system whose signals are shown in Figure 1. These parameters are used by the random system to construct the distribution of Rate of Return shown in Figure 2.

Level of Significance (a)	Number of candidate systems (N)				
	1	2	5	10	20
10%	90.00	94.87	97.91	98.95	99.47
5%	95.00	97.47	98.98	99.49	99.74
1%	99.00	99.50	99.80	99.90	99.95

Table 2: Percentiles corresponding to selected levels of significance a and numbers of candidate systems N.

alpha

When one system is tested, the critical value of 5% significance is taken from the 95th percentile of a performance distribution. When more than one system is tested, critical values must be adjusted upward to account for data mining. For example, if 10 *a priori* independent systems are tested, the critical value at 5% significance is from the 99.49th percentile.

In the case of curve-fitting, the systems are not independent of each other; N is thus somewhere between one and the number of parameter sets considered in the fitting.

When one system is tested on different price series, N is the number of those price series.

Level of Significance (a)	Number of candidate systems (N)				
	1	2	5	10	20
10%	27.5	39.4	55.9	68.2	81.0
5%	39.8	52.0	68.7	81.4	93.6
1%	69.1	81.7	97.8	109.5	123.6

Table 3: Critical values for Rate of Return for selected levels of significance a and numbers of candidate systems N.

alpha

Values calculated for S&P 500 March 2002 futures based on trade characteristics in Table 1 and percentiles in Table 2.

For example, if 10 *a priori* independent systems are being tested, pick a system at the 5% significance level if its Rate of Return is 81.4% or more.

Figure 1: SPH02: S&P 500 March 2002

Figure 2: Distribution of ROR under H_0

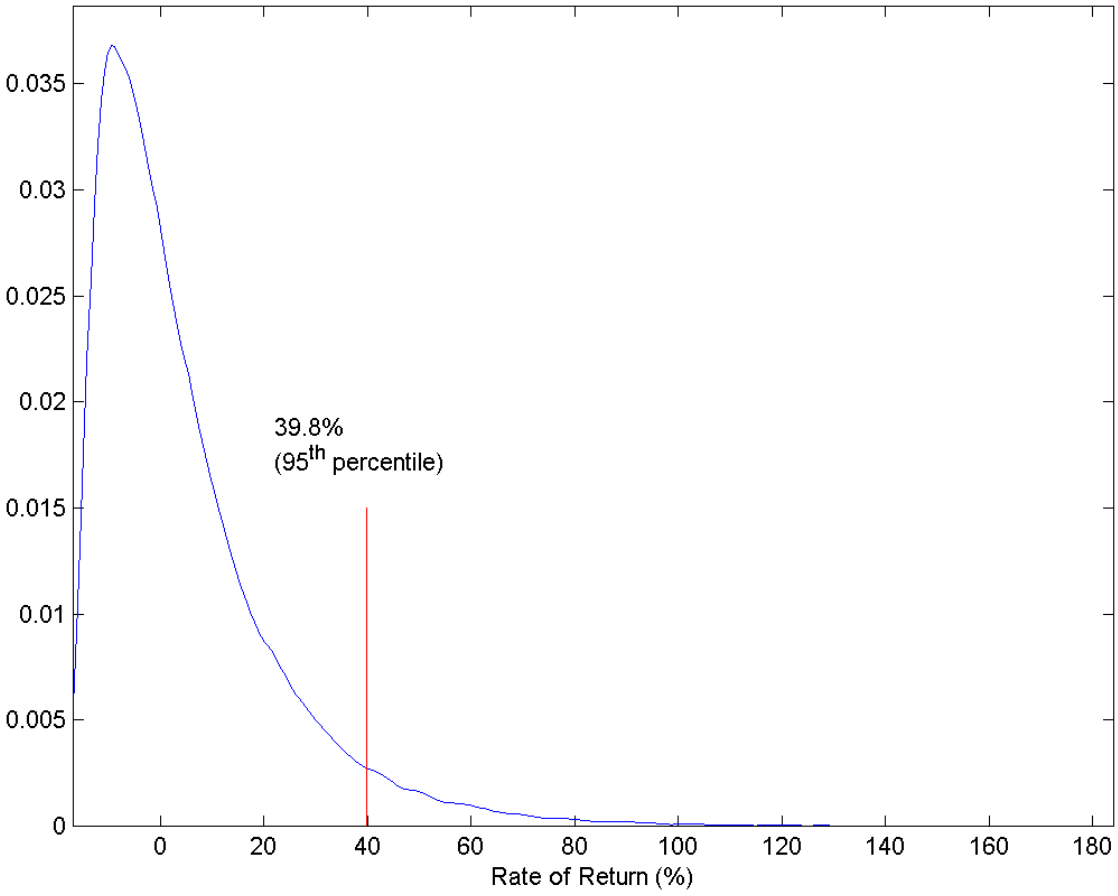

Figure 1: S&P 500 March 2002 futures. Also shown are signals from a hypothetical trading system. Triangle up indicates long one contract; triangle down indicates short one contract; x indicates a flat signal.

Figure 2: Probability distribution of Rate of Return under the null hypothesis of random trading. The distribution is constructed by applying the random system to the S&P 500 March 2002 price series and so is specific to it. The random system is run 100,000 times.

The vertical line at 39.8% shows the 95^{th} percentile of the distribution. There is only a 5% chance that random trading achieves a Rate of Return this high or higher.

Glossary

Curve-Fitting- Finding the set of parameters for a system that maximize its past performance. Also called optimization. When curve-fitting, the critical value used in the Random Signals Test has to be adjusted upwards. Though bounds on the critical value can be derived, its exact value is unknown.

Maximum Drawdown- The largest peak to trough decline in portfolio value. Maximum drawdown is a measure of risk. Profit divided by three times the maximum drawdown is a popular proxy for Rate of Return.

Performance- The measure of a trading system's worth in the eyes of a trader, such as Rate of Return, Sharpe Ratio, or a trader constructed measure.

Performance Distribution- The probability distribution function of a performance measure under the null hypothesis of random trading. This distribution is conditional on both the price behavior during the test period and the trade characteristics of the system being tested.

Random Signals Test- The method of system evaluation based its on past performance that is developed in this paper. The method is a hypothesis test that is based on randomly issued trading signals.

Random System- A series of random trades on the same price data that is used to calculate the performance of the system being tested. The performance from one run of the random system is one draw from the performance distribution.

System- A set of rules for making trading decisions.

Trade Characteristics- The characteristics that define a trade: number of contracts, trade duration, and transaction cost. In the Random Signals Test, the distribution of trade characteristics of the random system is set equal to the corresponding distribution of the system being tested.

Bibliography

Wolff, Jesse. “Frictional costs in futures trading.” *Technical Analysis of Stocks & Commodities*, April 2002, 20:4, pp. 68-70.