Issues Raised by Readers

Readers have asked questions about various aspects of the statistical tests described in the text and the Monte-Carlo Permutation Test document. I now address several of these issues.

Note that these notes are my own opinion. I cannot and will not be held responsible for losses due to errors or omissions. You accept all liability for your use or consideration of these opinions.

Using P-Values to Judge the Relative Performance of Trading Systems

"*How do I distinguish between systems that have different p-values, or that all have a p-value of zero? Which one is better?*"

This question concerns me, because it implies that the reader wants to use a p-value to judge the relative quality of a model. This is a common and serious error. A p-values does not measure performance. It measures the probability that a truly worthless model could have performed as well as the model being tested purely by good luck. This probability is heavily dependent on the variation among trade returns as well as the number of trades in the dataset. I would have no trouble producing a trading system that has an average annual return of just one percent, but has a p-value of nearly zero on any reasonable statistical test. I could then produce a system with an annual return of 15 percent and that has a large, insignificant p-value. To do this, I would only need to manipulate the number of cases or the variation in trade returns. Thus, you must never use a p-value to judge the relative effectiveness of a trading system. A small p-value is a necessary but not sufficient condition for a good trading system.

Selecting the Dataset for a Statistical Test

"My dataset has a return that is tremendously better than random systems generated in a Monte-Carlo Permutation Test, and its p-value is zero. I am sure that I have not overfitted the parameters in my model. Is a p-value of zero reasonable?"

When I hear this statement, I become suspicious that in-sample leakage has occurred. The dataset tested must be obtained from COMPLETELY out-of-sample data. By this, I mean that the data on which you tested your system is from a time period that has no overlap whatsoever with the time period in which you developed or confirmed proper operation of your model. You must not have even used this dataset to decide on which of several competitors or variations you choose. It must be totally virgin, never seen before the test is performed. If it has played any role whatsoever in the development of the trading system, any statistical test of its performance will exaggerate significance, often to an enormous degree. In this situation, you must use one of the methods described on Page 32 of the MC document, "Permutation Tests During Training."

So, assuming that your results are from a totally independent test set, let's think about what they say. As part of the MC test, you generated a huge number of random systems and computed the return of each. You then checked the return of the best of this multitude of random systems, the luckiest of the lucky, and you noted that it had a much lower return than your developed system. Think about what this means. Your system has fabulous performance, much better than the return of an extraordinarily lucky but ultimately worthless system. This is truly incredible. You have a real winner here! I claim that any statistical test that fails to profoundly reject the null hypothesis here is a ridiculously weak test. Of course, there is still the unlikely but possible caveat of the dual correlation problem discussed in the next section. But this can be dispelled by collecting more batches of out of sample data and trying the test again. Since the correlation problem exaggerates both tails, if this is the cause of the tiny p-value it will probably make itself known soon enough when an extraordinarily *bad* system appears.

The Monte-Carlo Permutation Test in Extreme Serial Correlation

"The market I am testing has extreme serial correlation. My trend-following system is able to capture many of these trends, and it performs well. However, when I do the permutations for the MC test, the trial position vectors are too random, so they do not benefit from being able to take advantage of contiguous long or short positions. The result is that my model's performance is much better than even the best of the random null hypothesis vectors, producing a p-value of zero. This would seem to be a weakness of the MC test in a serially correlated market."

This is discussed twice in the MC paper (Page 9 and Page 31). However, I will elaborate even further, since this is an important and potentially confusing issue. First, understand that this is as much a philosophical question as a rigorous mathematical question.

Here is the dangerous situation that I envision. (There may be more. This is the only one I can think of right now.) I will henceforth call this the *dual correlation problem*. Suppose we have a system for generating positions that are random and unrelated to the market data. Also suppose that its positions are inherently correlated. For example, one may roll a pair of dice. If a seven comes up, the next four positions are all long, and so forth. This system is obviously worthless. Now suppose we apply this serially correlated position vector to a serially correlated market. It would have an unusually high probability of getting lucky, with a bunch of longs lining up with a bunch of positive raw returns, and so forth. It would have an equally unusually high probability of being very unlucky, with strings of positions lining up with strings of the exact opposite signed returns. Thus, the distribution of returns would be unnaturally extended at both the winning and the losing ends. These broad tails would balance each other, producing a net expected gain of zero. Nonetheless, the broad right tail would result in excessive probability of rejecting the null hypothesis, a dangerous situation.

But what if the serially correlated positions result from an intelligent look at the serially correlated raw market returns? Certainly, the intelligent response to a serially correlated market is a serially correlated position vector. So in this case, we should (I believe) credit the intelligence.

Thus, the fundamental question in the dilemma comes down to this: If the serially correlated positions are an appropriate response to the serially correlated market, we want to credit this intelligence and reject the null hypothesis. To do otherwise would cripple the test, perhaps rendering it nearly powerless. But if the serial correlation in the positions is an artifact of the design of the system that is not related to the serial correlation in the market, we do not want to credit it. To do so would increase the likelihood of erroneously rejecting the null hypothesis, a serious error. Unfortunately, we often do not know the true nature of the situation, so we have to weigh the price and benefit of each alternative and choose wisely, being ready to accept the consequences of a bad choice.

It is important to note that this phenomenon is not a problem with only the Monte-Carlo permutation test. The deeper fundamental problem is that when a serially correlated but otherwise random position vector is mated to a serially correlated market vector, the result will be a preponderance of very lucky and very unlucky systems. This is a problem with the data, not the test. Different tests will reflect the dual-correlation problem in different ways:

> In the Monte-Carlo Permutation Test, most random permutations result in position vectors that do not have the serial correlation necessary to significantly overlap serial correlation in the market. The result is that the null distribution is narrower than what would be expected in the real-life population, a serious violation of the permutation principle. Hence, the null hypothesis will be rejected too often. A crude fix is to limit the random permutations to those having serial correlation similar to that of the test system. This is discussed later.
>
> In the ordinary *t*-test, a fundamental assumption is that the variance of the mean of a set of *n* independent observations is the original variance divided by *n*. But in the dual-correlation problem, the returns themselves are serially correlated. If a set of observations is correlated this way, the variance of the mean does not drop by a factor of *n*. It drops more slowly. The result is that the *t*-score is inflated, resulting in excessive rejection of the null hypothesis. A crude fix is to estimate the serial correlation and apply a simple formula to correct the variance estimate. This is risky.
>
> In a bootstrap test of data having the dual correlation problem, the individual bootstrap replications will have a distribution that is too narrow relative to the population distribution of the test statistic, due to the clustering of serially correlated observations. The result is that the test statistic will lie too far outside the null distribution, causing excessive rejection of the null hypothesis. In fact, if you look at Figure 2 in the MC document, which displays the results of a simulation of this problem, you will see that the bootstrap test is actually more adversely impacted than the MC test. There do exist some bootstraps that can partially compensate for serial correlation, and I have not tested them in this scenario, but I strongly suspect that they would not provide any material help here. The process of estimating serial correlation is risky, and these tests do have significantly lower power than the normal bootstrap.

Here is another way of looking at the dilemma. Suppose we have a market that has strong serial correlation. A successful trading strategy will probably have two characteristics:

1) Its positions will have strong serial correlation so as to match the extended trends

2) Among all possible similarly correlated systems, it will be superior.

The straightforward MC permutation test, as well as other tests like the *t*-test and the bootstrap, are sensitive to *both* of these characteristics. In other words, the null hypothesis is that the system is worthless, versus the alternative that *something* is making it outperform a worthless system. This something may be serial correlation, or superior intelligence among serially correlated systems, or both (serial correlation created as the result of an intelligent model). In the likely event that the observed positions are serially correlated, do you really want to limit your statistical test to the second characteristic, relative intelligence, discounting the possibility that the observed serially correlated positions are actually an intelligent response to the serially correlated market?

You might be able to modify the MC test to force it to employ a null hypothesis distribution consisting of only position vectors having serial correlation comparable to that of the test system. This could be done by rejecting unsatisfactory permutations. This will, of course, extend the width of the null distribution as unusually lucky and unlucky systems replace many middling systems. This will result in a more conservative test, reducing the impact of the dual correlation problem discussed above. At the same time, it will reduce or eliminate the credit given to characteristic 1 above, serially correlated positions. It will test only the degree to which this particular system is superior to other similarly correlated systems, ignoring the degree to which serial correlation in the positions helps the return.

Which is the better approach? You need to be the judge. My own feeling is that since a serially correlated position vector is the intelligent response to a serially correlated market, we want to give it credit by performing the full MC permutation test, building the null distribution based on all possible permutations, including those with negligible serial correlation. The price paid is susceptibility to the dual correlation problem, an admittedly dangerous situation. On the other hand, the price paid for limiting the null distribution to serially correlated positions, aside from the computational difficulty of doing so, is a test that is so conservative that the component of success due to producing correlated positions is discounted. This could so weaken the test that it becomes worthless. You pays your money and you takes your choice.

I conclude this section with a plea for more research on this topic. I don't have time to do the extensive testing that would be required to cast more light on this important and confusing subject. Here is what I would like to see: Collect historical returns for a variety of correlated markets. Also generate a few synthetic markets having known serial correlation. Define a variety of random (hence worthless) serially correlated position vectors. Then use vast replications of experiments to assess the degree to which the dual correlation problem impacts the full Monte-Carlo Permutation Test, as well as the several stationary bootstrap tests. The results of these tests would go a long way toward resolving this dilemma.