Dart to the Heart*

Patrick Burns

8th March 2007

Most likely you know of the stock market dartboard game: some reputed experts are pitted against a portfolio that was selected "by throwing darts". This makes compelling journalism—especially when the darts win—but is less than perfect science.

However, a more rigorous version of this game *is* good science. The enhanced method generally goes by the name of "random portfolios" or "Monte Carlo simulation". It has the power to radically transform the practice of fund management—a dart to the heart.

We will start by taking a close look at performance measurement. We will then move on to some wider issues of fund management.

Perfect Performance Measurement

The aim of the dartboard game is to assess the skill of the experts. While the game in the newspaper involves the selection of only a few stocks, the real-world application is to judge the skill incorporated into a particular fund. We measure the skill during a specific time period.

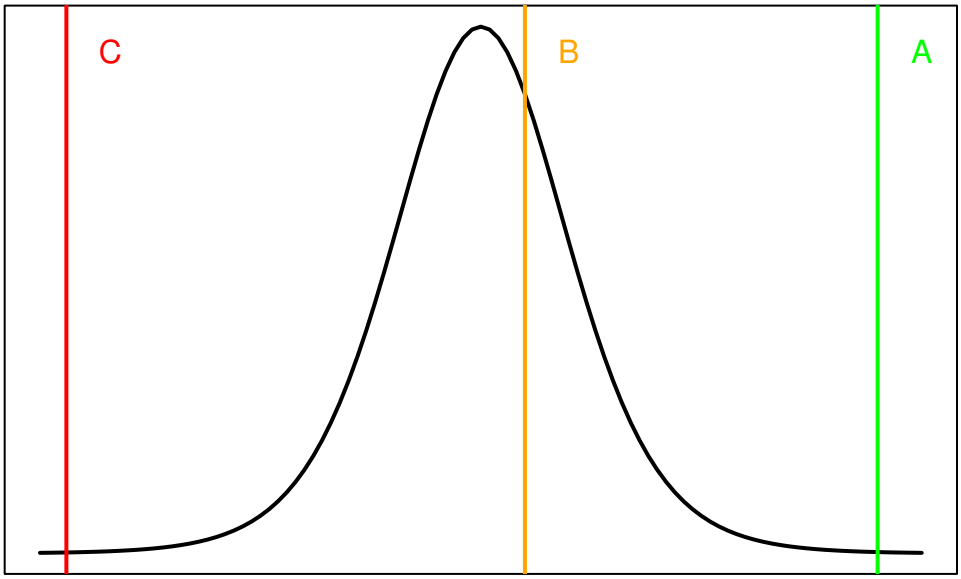
All funds operate under constraints. Some of these constraints are likely to be mandated—the universe of assets, maximum tracking error, perhaps maximum weights, and so on. Others will be self-imposed by the fund manager. Imagine we have a list of all of the portfolios that satisfy the constraints of a fund, and the list also includes the return over the time period for each of the portfolios. Comparing the actual return of the fund to the returns for all of the possible portfolios gives us all of the information we could hope to have about the skill of the fund.

(For simplicity we are assuming that the performance measure is the return, and that there is no trading during the period—relaxing these assumptions causes little if any additional difficulty in practice.)

Figure 1 shows a hypothetical example. The density represents the distribution of returns for all of the possible portfolios (that is, all of the portfolios that

*A mildly edited version of this appeared in the March 2007 issue of *Professional Investor* (http://www.uksip.org) under the title "Bullseye". This version may be found in the Working Papers section of the Burns Statistics website (http://www.burns-stat.com).

Figure 1: A hypothetical distribution of the returns of the set of all portfolios a fund might have held, along with three specific possible holdings.



satisfy the constraints). If the fund selected portfolio A, then the fund is exhibiting skill—of all of the portfolios that could have been chosen, it selected one of the ones that garnered the largest returns. If the fund selected portfolio B, it selected a portfolio that was about average and hence is not exhibiting much skill. If the fund selected portfolio C, then it is exhibiting negative skill—it selected a portfolio that was almost as bad as it could be.

The problem with this perfect method is that the list, even if written very efficiently, would in most cases use a large portion of the matter in the universe. For any real fund this method will be impossible to implement.

Three Measures of Performance

Although perfect measurement can not be carried out in practice, a close approximation of it can be. Instead of looking at all the portfolios the fund might have held, we look at a random sample of those portfolios. The sample needs to be no bigger than a few thousand—this is quite feasible.

While random portfolios provide an almost perfect performance measurement, the two most commonly used methods are benchmarks and peer groups. We'll first look at a summary of the three methods, and then compare them.

The inputs to the random portfolio technique are the fund's constraints, and the time period of interest. A number of random portfolios that obey the constraints are generated, then their returns over the time period are found. The statistic of interest is a slight modification of the fraction of random portfolios

that outperform the fund. The statistic is interpreted as the probability of the fund performing at least as well as it did if it had zero skill. A statistic much less than one-half is evidence of skill. (The statistical jargon for this is that the statistic is a p-value of the null hypothesis of zero skill.)

The input for the benchmark technique is the difference between the fund returns and the benchmark returns over several periods. If the mean difference is large relative to the variability of the differences, then the hypothesis of zero skill can be rejected. (In statistical jargon, a t-test is performed on the differences in returns.)

The inputs for the peer group technique are a time period of interest and the group of funds that are similar to the fund that is being tested. The percentile of the fund return among the returns of the peers is found. A large percentile close to 100 is an indication of skill.

While random portfolios and peer groups test a single time period, the benchmark test requires several time periods. The need for multiple time periods means that the benchmark test has poor power to distinguish good funds from bad.

Tables 1 and 2 show the power of the benchmark test. For example with 12 quarters of data the power at 5% significance is 49% when the information ratio is 1—this test has a 5% chance of declaring skill when there is none, but only about a 50% chance of declaring skill when exceptional skill exists. If the information ratio is 0.5—a more likely value for a good manager—the power at 5% significance is less than one-half with 40 quarters of data. (The power when the information ratio is zero should always be equal to the significance level, but is not always exactly equal in the tables because these results are from a simulation.)

Another problem with benchmarks is that the ease of beating them is not constant. When the assets with the largest weights in the benchmark happen to perform relatively well, then the benchmark will be hard to beat. Conversely when the assets with the largest weights happen to do relatively poorly, the benchmark will be easy to beat. This effectively puts random numbers into the differences of returns, and hence makes the benchmark test even less powerful than Tables 1 and 2 indicate.

The peer group method seems very similar to the random portfolios technique. In spite of the similarity there are quite important differences. First off, with random portfolios we have a well-defined statistical hypothesis that is being tested. Nothing specific is being tested with a peer group—the percentile of the fund depends on the skill of the fund, the skill of the other funds and the noise (relative to skill) in the returns of the funds.

Consider a test over a single day. Certainly the percentile of a fund among its peers over one day is going to be almost pure luck (unless the funds trade an extraordinary amount). Over such a short time the noise in the returns is going to swamp the effect of skill in the funds. While we get a percentile for our fund, we see that it has little or nothing to do with skill.

In contrast a random portfolio test over one day is perfectly valid—it always

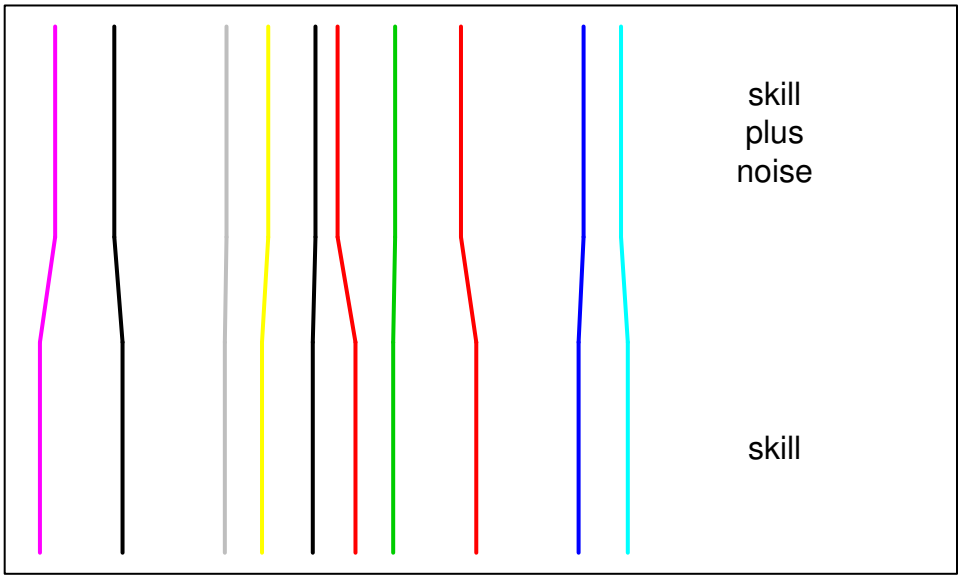
Table 1: The power of the benchmark test with quarterly data. The last three columns show the fraction of tests that have a p-value smaller than the column heading.

Quarters	**Info ratio**	**5% significance**	**1% significance**	**0.1% significance**
8	**0**	5.0%	1.0%	0.1%
8	**0.1**	6.6%	1.4%	0.2%
8	**0.2**	8.1%	1.8%	0.2%
8	**0.5**	15.9%	4.1%	0.5%
8	**1.0**	35.7%	12.4%	2.0%
12	**0**	4.9%	1.0%	0.1%
12	**0.1**	6.9%	1.5%	0.2%
12	**0.2**	9.3%	2.1%	0.2%
12	**0.5**	20.3%	6.0%	0.8%
12	**1.0**	49.0%	21.0%	4.4%
20	**0**	5.1%	1.0%	0.1%
20	**0.1**	7.7%	1.7%	0.2%
20	**0.2**	11.3%	2.8%	0.3%
20	**0.5**	28.3%	9.7%	1.8%
20	**1.0**	69.6%	40.2%	12.9%
40	**0**	5.1%	1.0%	0.1%
40	**0.1**	9.3%	2.2%	0.3%
40	**0.2**	15.3%	4.3%	0.6%
40	**0.5**	46.5%	21.2%	5.6%
40	**1.0**	92.9%	76.6%	45.2%

Table 2: The power of the benchmark test with annual data. The last three columns show the fraction of tests that have a p-value smaller than the column heading.

Years	**Info ratio**	**5% significance**	**1% significance**	**0.1% significance**
2	**0**	5.0%	1.0%	0.1%
2	**0.1**	5.9%	1.2%	0.1%
2	**0.2**	7.0%	1.4%	0.1%
2	**0.5**	10.5%	2.1%	0.2%
2	**1.0**	18.0%	3.6%	0.3%
3	**0**	5.0%	1.0%	0.1%
3	**0.1**	6.6%	1.4%	0.2%
3	**0.2**	8.2%	1.6%	0.2%
3	**0.5**	15.3%	3.3%	0.3%
3	**1.0**	32.3%	7.7%	0.8%
5	**0**	5.1%	1.0%	0.1%
5	**0.1**	7.2%	1.5%	0.2%
5	**0.2**	10.1%	2.2%	0.2%
5	**0.5**	23.9%	6.4%	0.8%
5	**1.0**	57.8%	21.8%	3.2%
10	**0**	5.0%	1.1%	0.1%
10	**0.1**	8.8%	2.0%	0.2%
10	**0.2**	14.4%	3.8%	0.5%
10	**0.5**	42.9%	16.6%	3.0%
10	**1.0**	89.7%	64.0%	23.7%

Figure 2: An ideal peer group analysis in which skill plus noise is virtually identical to skill.



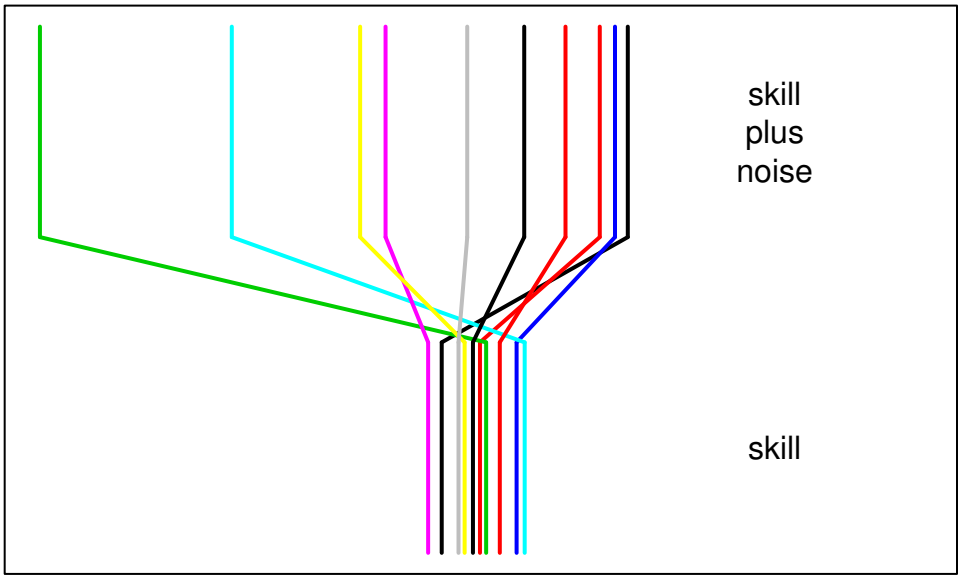
tests against zero skill portfolios so short time scales do not invalidate the test. The one-day test is generally not of much importance, but it is a true test of the skill of the fund. The p-values for a series of one-day random portfolio tests can be combined into a test for a longer period.

Figure 2 shows a perfect peer group analysis in which the order of the skill plus noise of the funds (that is, the returns) is the same as the order of the skill of the funds. In contrast Figure 3 shows an example where the order of skill plus noise is significantly different from the order of skill. Which of these two figures more accurately represents an analysis depends not only on the length of the time period but also on the dispersion of skill of the peer funds.

The lesson is that the time period over which a peer group analysis is performed needs to be long enough that the distribution of skill among the funds dominates the noise. In the absence of using random portfolios, it seems impossible to know how long the time period should be for this condition to hold. The results of a peer group analysis give no indication of how affected they are by noise.

The breadth of the peer group as well as the expanse of time is important. Being at the 90th percentile among 200 funds is more precise than being at the 90th percentile among 10 funds. Hence there is a tension between including a lot of funds in order to have more precise percentiles, and restricting the group to only those funds that are most like the target fund. In contrast, precision in the random portfolio method is not a problem—it is likely to be only a matter of a few minutes of computer time to add an additional 1000 random portfolios,

Figure 3: A peer group analysis in which noise dominates skill.



and the addition only makes the test more trustworthy, not less.

Figure 4 shows a distribution of returns for the year of 2005—this is from random portfolios that satisfy a certain set of constraints. The stocks in the portfolios are from the FTSE 350, the portfolios of 90 to 100 stocks are long-only, the maximum weight of a stock is 5%, between 10% and 30% of the weight is from the FTSE 100, 50% to 70% of the weight is in high yield, and the weight of each of the five sectors is between 10% and 30%. If we had a fund that satisfied these constraints, then we could use this distribution to judge its skill for the year.

It is not unusual to know (or be able to get) the composition of the portfolio at the beginning of the time period. While this information is of no use when using a benchmark or a peer group, it can be very valuable for an analysis with random portfolios.

Figure 5 shows a distribution of random portfolios that obey the same set of constraints as before, but they are generated by starting the year with a specific portfolio and then trading 8% of the portfolio value every other week. Hence these portfolios have traded more than 200% by the end of the year. The distribution is substantially narrowed by using the additional information of the holdings of the fund at the beginning of the year and an estimate of its turnover. The performance measurement based on the distribution of all portfolios that satisfy the constraints is already excellent relative to traditional methods, but conditioning on the initial portfolio greatly improves the measure.

The return for the initial portfolio with no trading is at the 35th percentile

Figure 4: The distribution of the 2005 returns for portfolios satisfying the set of constraints.

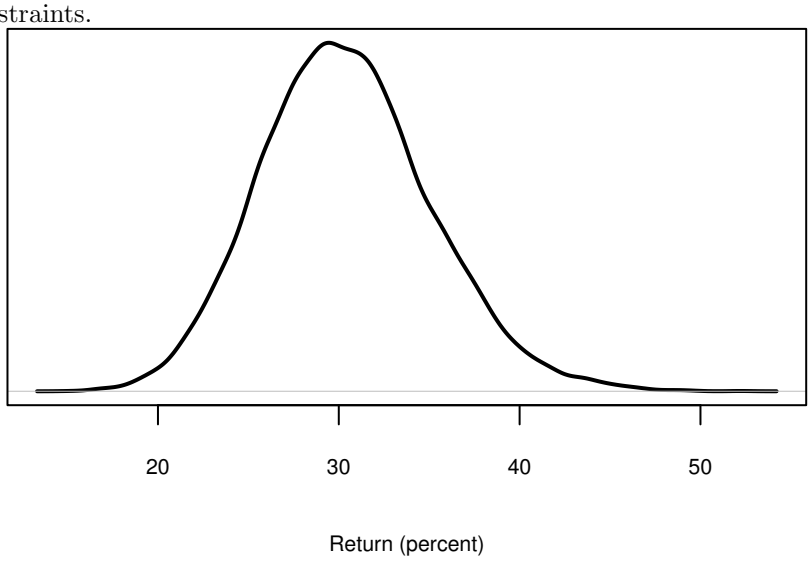
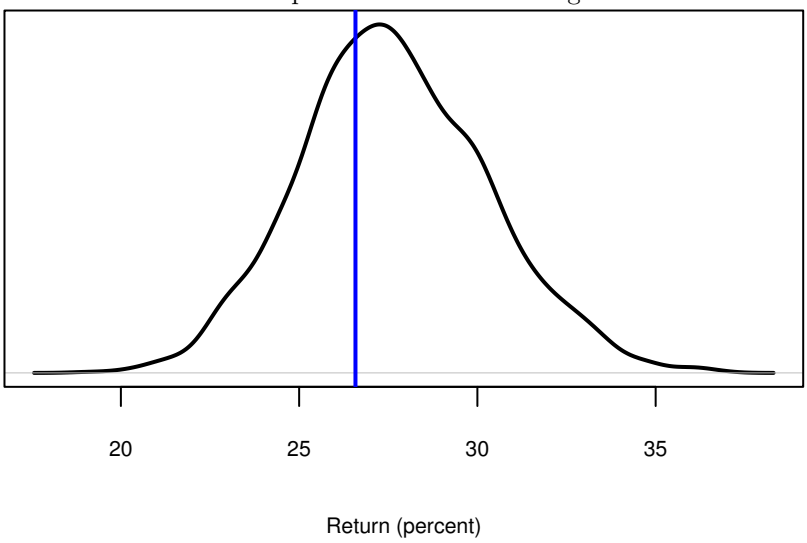


Figure 5: The distribution of the 2005 returns for portfolios that trade 200% over the year from a particular portfolio and satisfy the constraints. The vertical line is the return for the initial portfolio with no trading.



of the distribution in Figure 5. Since this portfolio has a low return among all of the portfolios that satisfy the constraints, trading away from it is more likely to improve the return than hurt it. A fund that performs better than what would have happened with no trading over the period is not necessarily exhibiting skill.

Performance and skill are really distinct concepts. Performance is about what has already happened while skill is a window on what the future holds. The random portfolio analysis conditional on the initial portfolio allows us to separate skill from performance.

The Tiger Woods Problem

Another way of phrasing the performance measurement problem is that we need to distinguish skill from luck. Unfortunately we will never be able to do that definitively.

I think that Tiger Woods is an average golfer who happens to have been very, very lucky. You may think that he is skilled. There is no way to objectively declare one of us wrong—where to draw the line between skill and luck is a matter of opinion.

The p-value that random portfolios provide is a useful input for the decision. The p-value is sometimes interpreted as the chance of there being no skill—*this is wrong*. P-values are conditional statements: it is the probability of seeing a return at least this big *if there were no skill*. But we don't know if there is skill or not, that is what we are trying to find out.

There is another complication as well. Just because a fund underperforms during a time period doesn't mean that it doesn't have skill. It does not *exhibit* skill during the period, but it may be in a very good position that hasn't yet paid off. The best we can do is to know if a fund appears to be skillful in given time periods.

Performance measurement with random portfolios does not eliminate the need for judgement. What it does do is provide better evidence so that judgements are much more informed.

In the Service of the Investor

The role of fund managers is to maximise, to the best of their ability, the utility of their investors. At present fund management seldom fulfils this.

The investor's utility depends not only on the returns that the fund manager provides, but also on the correlation of those returns to the rest of the investor's portfolio. In a specific asset class an investor usually has the choice of buying an index fund for a very small management fee or buying an active fund for a considerable management fee. It is almost always the case that the best choice of the investor is to put at least part of the money into the index fund.

Typically the active fund is mandated to have a small tracking error to the index. Now, either the fund manager reliably outperforms the index by enough to repay the management fee, or not. If not, then the investor is obviously better off only buying the index fund. If the fund manager does have skill, then the investor is best off when the active fund has a small correlation with the index. A small correlation is the same thing as a large tracking error. In terms of investor utility it makes more sense to have a minimum tracking error constraint rather than a maximum tracking error constraint.

Maximum tracking errors exist because of the issue of performance measurement. Suppose two active funds have the same return and volatility, but the second fund has twice the tracking error. In reality the second fund is more beneficial to the investor, but it will be much easier to declare the first fund skillful if performance relative to the index is used as the performance measure.

A fund unconstrained by tracking error should be in the best interests of the investor. A reasonable argument against unconstrained mandates has been that investors have a hard time knowing if the fund manager is performing well. Random portfolios solve this—the investor can have a good estimate of skill no matter what the tracking error. The abolition of a tracking error constraint leaves the fund manager free to add more value to the investor. Performance of the unconstrained portfolio has potential for greater gains, and even with the same returns the investor is better off because of the lower volatility of the entire portfolio due to the reduced correlation between the fund and the index.

It is possible to go even further—the active fund can be constrained to have a low correlation with some proxy for the portfolio of the investor. Random portfolios can be used to gauge what a good value for the correlation limit might be. There will be a trade-off between the return of the fund and the strictness of the correlation constraint. Random portfolios can show how the distribution of returns is affected as the constraint bound is changed.

Since random portfolios measure performance more accurately, they facilitate performance fees. Performance fees can of course be based on a fund's performance relative to an index. However, this is probably more of a bet on market capitalisation (or whatever it is that determines the constituent weights) than it is a reward for investment skill. Creating a performance fee relative to an index is very much like trading an option on the factor that determines the weights in the index.

Without performance fees there is no reason for fund managers to restrict the size of their funds. Fund managers that have had good performance tend to increase the size of their funds until their performance falls back to average—good fund managers benefit from their skill, but the investors tend not to benefit. Performance fees can allow investors to share in the benefits of the skill.

Random portfolios provide close to the best performance measurement possible. The availability of accurate performance measurement allows additional changes in fund management to ensue. In particular, the major complaint against unconstrained mandates is removed. Throwing darts, it turns out, is more than a game.