portfolio construction forum

A lousy way to evaluate active managers

Michael Kitces | Pinnacle Advisory Group | 12 November 2013

Executive summary

Determining whether investment results are due to luck or skill is no small task for even skilled analysts evaluating an active investment manager. Given the amount of randomness inherent in markets, it can be very difficult to determine which is which. Fortunately, the field of inferential statistics exists specifically to analyse such situations and help to distinguish the signal from the noise, and determine when results are likely randomness and luck versus when there is at least a high probability that skill or some other factor is at play.

But, when inferential statistics is applied to evaluating active management, the results are questionable at best.

Given just how incredibly volatile markets really are, searching for "statistically significant" outperformance may actually be a lousy approach for evaluating managers. Even if a manager really does outperform for an extended period of time, the available tests simply are not capable of distinguishing skill from market noise given the tenure of even long-standing managers. In fact, when tested to determine the effectiveness of the approach in the first place, the reality is that even if a manager is adding several hundred basis points of outperformance, annually, for more than a decade, there is still a more-than-90% likelihood that inferential statistics will FAIL to identify the signal that really is there.

In other words, if the goal is actually to determine which active managers really DO add value, searching for statistically significant outperformance is an approach with an overwhelming likelihood to fail, even in situations where it should be succeeding! Which means in the end, failing to find statistically significant outperformance amongst active managers may actually be less a failure of active management itself, and more a problem with using an approach that was unlikely to successfully identify good managers in the first place!

UNDERSTANDING INFERENTIAL STATISTICS

Determining whether investment results are due to luck or skill is no small task for even skilled analysts evaluating an active investment manager – given the amount of randomness inherent in markets, it can be very difficult to determine which is which.

Fortunately, the field of inferential statistics exists specifically to analyse such situations and help to distinguish the signal from the noise, and determine when results are likely randomness and luck versus when there is at least a high probability that skill or some other

factor is at play.

But when inferential statistics is applied to evaluating active management, the results are questionable at best

The basic principle of inferential statistics is fairly straightforward – to try to draw conclusions from data that is muddled by randomness. Viewed another way, its goal is to find signals amidst noise.

Of course, sometimes there is a lot of noise, so the process of inferential statistics is fairly conservative in its approach, to avoid drawing inappropriate conclusions. If the difference between A and B is not significant – not statistically significant – it's assumed to just be a result of noise and not a signal. In other words, inferential statistics tries to minimise the risk that we make a mistake – called it a Type I error – of saying that A and B are different when in fact they're not.

On the other hand, given that we only draw a conclusion about whether A and B are different in scenarios where the magnitude of the difference – the signal – is larger than the magnitude of the randomness – the noise – it can be very difficult to draw much of a conclusion about anything. Fortunately, as the number of measurements increases, the randomness tends to cancel itself out. As a result, larger sample sizes (assuming they're sampled appropriately) tend to have less randomness, which makes it easier to differentiate the signal from the noise.

For instance, if I'm trying to determine whether the average height of men is taller than the average height of women in my local neighborhood, it's not so clear if I just measure one or two people. The fact that the first two men to be measured happen to average 5' 5" and the two women average 5' 3" doesn't mean we can draw a conclusion that men are taller than women where I live. After measuring only four people, it's possible the conclusion would be wrong just due to random chance (maybe I bumped into some especially tall or short neighbors). Yes, according to the sample the men are 2 inches taller than women on average, but given that human beings – both male and female – can vary from under 5 feet to over 7 feet, a 2 inch difference from measuring only four people just isn't enough to affirm that the difference is statistically significant. Or, viewed another way, if I determined after just a few people that the height of men in my neighborhood is 5' 5" plus-or-minus 6 inches, and the height of the women is 5' 3" plus-or-minus 6 inches, then clearly the 2-inch difference between them isn't all that significant given the +/- 6 inch bands of uncertainty.

On the other hand, if we keep growing our sample by measuring more people, on average, our estimates should move towards the true heights for all men and women in my area, and the variability should decline, as a few randomly tall or short people both tend to cancel each other out and become less of an impact on the overall average as the number of people grows. For instance, after several dozen measurements, we might find that the average height of men in my area is 5' 8" and that women are 5' 4", and that based on randomness alone we're 95% certain those estimates are accurate within +/- 3.5 inches. Notably, this

means we have now crossed the threshold of statistical significance – when the difference between the groups is 4 inches, and there's a less-than-5% chance that a difference larger than 3.5 inches could be due to randomness alone, we draw the conclusion that the men in the area are in fact taller than women because the odds the observed 4-inch difference is due to chance alone is small.

Technically, we still haven't PROVEN it, because we haven't measured everyone. In fact, our measurements haven't even precisely predicted the actual height of men and women nationwide (which is actually 5' 9.5" vs 5' 4" respectively in the US). Nonetheless, when we reach the point where the differences are so large that there's a less-than-5% chance it's due to randomness alone, we assume we've got a signal. Technically, that less-than-5% chance also represents our probability of a Type I error, also known as (statistical) alpha – that is, that the men around here really aren't taller than women, and that our sampled difference really is just random noise.

Notably, though, back when my sample of four people wasn't large enough to determine whether the 2-inch difference was noise or just a signal, it would be wrong to conclude that "the men are not taller than women here" just because we didn't have a statistically significant difference. Failing to have a large enough sample to separate noise from signal doesn't mean there is NO signal, just that we don't have enough data to draw a conclusion that there IS a signal. This is important, because in some situations we are limited to the number of measurements we can make. For instance, if I only had the time to measure half a dozen people total, there is a significant risk that, even though the men here really are taller than women, the differences between our small sample of men and women wouldn't be large enough to affirm statistical significance. This scenario – where there really is a signal, but we fail to successfully detect it amidst the noise – is called a Type II error, and is especially common when we don't have a large enough sample to minimise the noise (and accordingly, the noisier the data, the larger the necessary sample).

CONFIDENCE INTERVALS, TYPE I AND TYPE II ERRORS, AND STATISTICAL POWER

Ultimately, we try to affirm that we're not making a Type I error by determining not just the average of our samples and whether one is bigger on average than the other, but the (typically 95%) confidence intervals around those averages, and we don't draw a conclusion that group A is different than group B until the magnitude of the difference between them is greater than what we might expect from merely randomness within our confidence interval alone. Of course, sometimes the actual difference between the groups – called the effect size – isn't all that large to begin with, so it may require an extremely big sample to hone the randomness down to the point where the confidence interval is so small, we're finally able to distinguish a signal from the noise. For instance, if the truth was that men really were taller than women in my neighborhood, but the average difference was actually only half an inch, I'd need to measure a lot of people before I could safely draw a conclusion about such a

small difference.

On the other hand, sometimes this itself can actually be a problem. If the truth is that the effect size really IS fairly small, but the groups have a lot of variability, and we're limited in how big of a sample size we can gather in the first place, there's an increasingly high likelihood that we will fail to find a signal in the noise and make a Type II error. Not because there isn't a signal, but simply because it was very difficult to detect the signal due to the combination of small effect size, high variability, and limited sample. In fact, inferential statistics uses a measure called statistical power (also known as statistical beta) to calculate, given a certain anticipated effect size, an estimate of variability, and the available sample size, the likelihood that the researcher will make a Type II error, failing to detect the actual signal that was really there.

Notably, these measures all impact each other as well. The wider we draw the confidence intervals, the more we reduce the risk of a Type I error (by making it harder and harder to draw a conclusion about a signal by assuming there is a larger amount of noise), but the more we reduce the statistical power and increase the risk of making a Type II error (by making it so hard to distinguish real signals in the noise that we mistakenly fail to detect them when they're really there).

In general, statisticians err in the direction of not identifying signals that turn out to be noise - in other words, we set the risk of a Type I error at a fairly low level, even at the "cost" of reducing statistical power - but it's important to realize that, in some circumstances, statistical power may turn out to be very low indeed.

APPLICATION TO ACTIVE INVESTMENT MANAGEMENT

So what does all of this have to do with active management? The all-too-common approach for evaluating whether there is value in active management overall, or an actively managed fund in particular, is to measure the results of the fund against the results of the index, to see whether the results are "statistically significant" - in other words, given the underlying randomness inherent in the index itself, and the randomness that would occur from an active manager merely due to chance, is the difference between the two large enough that we can distinguish a signal from noise?

The problem, unfortunately, is that measuring the results of investment performance is a classic scenario where the variability is high (the standard deviation of equities is typically estimated around 20% or a bit more based on annual returns), and the effect sizes are likely modest at best. After all, even if the fund manager can outperform, an amazing manager still might "only" outperform by a few percent per year on average, which is dwarfed by the 20% annual volatility of equities. Of course, the measurement of random volatility alone will decline over time (technically, by the square root of the number of years), but that too is constrained - until a manager has a longer track record, there simply aren't all that many years available to measure in the first place. The net result is that even if there is a signal and

the active manager is creating outperformance, at a level that can still accrue a material amount of long-term wealth, it can be remarkably difficult to measure it using inferential statistics. Instead, the overwhelming likelihood is that a Type II error will be made - that is, failing to identify the signal even when there is one.

How bad is the risk? Let's say we have an equity manager who is actually capable of outperforming his benchmark by 100 basis points per year (after netting all appropriate fees). While this is a fairly modest level of investment outperformance, it's nonetheless quite material over a long period of time. With stocks averaging about 10% annualised over the long run, a $10,000 indexed portfolio would grow to $174,494 after 30 years, but the manager with an 11% annualised return would grow to $228,923, a whopping 31.2% increase in wealth. In a lower return environment - for instance, if stocks only provide a return of 8% going forward - the 1% outperformance has a slightly greater impact on a lower base, resulting in future wealth of $100,627 in the index returning 8% and $132,677 from the fund manager earning 9% (a 31.9% difference in wealth).

Yet, placed against a backdrop of equities with 20% volatility, it turns out to be remarkably difficult to affirm that the results are an actual signal. Using this statistical power calculator to compare the manager (fund being evaluated) versus the population (the benchmark index) to measure a continuum of potential results, we find that after five years of a manager averaging 11% while the index averages 10% with a 20% standard deviation, the statistical power is a whopping... 3.2%. In other words, assuming the manager really IS capable of outperforming by 1% per year, there's only a 3.2% chance that our approach will correctly identify this - and a 96.8% chance that we'll fail to realise the manager as successful using inferential statistics. What happens if we wait 10 years? Not much better. The statistical power only rises to 3.6%. After 30 years? 4.6%.

Yes, that's right - using inferential statistics, even a manager who outperforms by 1% per year for an entire 30-years (what could be his/her whole career at that point!) still has a 95.4% chance of being "indistinguishable from noise" by the end. If the manager outperforms by 2.5% per year - which will double the investor's final wealth after 30 years of compounding - the statistical power is still only 10.1%. To put that in real dollar terms, that means if Investor A finishes with $1,000,000, and Investor B finishes with $2,000,000 because his investment manager is brilliant for 30 years, inferential statistics would still conclude there's an 89.9% chance that this was just random luck. After 30 years. If you were merely giving the manager a "typical" three to five years to establish a track record, the statistical power falls back below 5%, even assuming a whopping 2.5% per year outperformance effect size.

So what's the bottom line to all of this? Simply put, assessing whether a manager is "good" or "successful" or not by using inferential statistics to determine whether the outperformance is likely due to skill or indistinguishable from chance is a virtually useless way to approach the problem of manager assessment. Over what most investors would consider a generous time horizon, like three to five years of building a track record, the methodology has a whopping

95%+ probability of failing to identify a real manager who actually creates real enhancements to return. Even over multi-decade time periods, there's still a 90% failure rate for the approach to accurately detect material outperformance. That's simply the reality of trying to measure relatively modest differences of a few percentage points a year of outperformance against a backdrop of 20% standard deviations.

Fortunately, the volatility is less severe for some other asset classes - which improves the statistical power - but unfortunately less volatile asset classes also tend to have smaller effect sizes (less outperformance potential in the first place), which means the methodology is equally problematic there, too. Of course, in some cases, investment managers are so bad that their results statistically significant - to the downside - and despite the problems of inferential statistics, those are clearly managers to avoid (just as you have to be really good to be identified as such by inferential statistics, if the results are statistically significant in the other direction, the manager must have been really bad!).

But, for the overwhelming majority of funds - where technically, the results were not statistically significantly bad, but merely failed to be significantly significant in the positive direction - skill is not disproven but simply not able to be distinguished from luck. And given what constitutes "good" outperformance relative to the volatility of equities, distinguishing luck from skill using this kind of approach is actually almost impossible.

Which means the reality is that failing to distinguish luck from skill when evaluating a manager may be less a problem of the manager, and more a problem of the tool being used to do the measurement in the first place. Because, in the end, trying to evaluate active management with tests of statistical significance is, in fact, significantly likely to be wrong, even when the manager actually makes calls that are right. Which means in turn, if you want to evaluate the prospective value of an investment manager, it's time to focus on more qualitative methodologies (evaluating incentives, knowledge, experience, governance, process, etc.), because inferential statistics just isn't capable of doing the job.

Michael Kitces is a Partner and the Director of Research of Pinnacle Advisory Group, a US-based private wealth management firm that works with over 700 families and manages close to US$1 billion in assets for clients in the US and around the world. The above article is reproduced with permission from Michael's blog "The Nerd's Eye View". Michael is a member of PortfolioConstruction Forum's core faculty of leading investment professionals.

More about Michael
