CFM
INSIGHT.DATA.CLARITY.

May 2016

IN-SAMPLE OVERFITTING

Avoiding the pitfalls in data mining

Executive summary

Probably one of the biggest pitfalls in investment, in-sample biases are often among the most difficult to convey and explain and also to avoid. Using simulations along with real world data we try to demystify this effect, explain the dangers and put forward suggestions as to how they can be mitigated.

Contact details

Call us +33 1 49 49 59 49
Email us cfm@cfm.fr

Introduction

Any newcomer to the world of investment may be forgiven for believing that achieving at least a positive return should not be overly difficult. Even small amounts of investable cash bring out offers of impressive levels of return to entice investors to the table. The limitless world of products available, each with high levels of realised performance, or in some cases simulated performance, can give the idea that tomorrow's returns will be just as rosy. Despite our desire to believe otherwise, common sense and intuition tell us it cannot be easy to achieve significant levels of return without assuming high levels of risk.

In-sample or data snooping biases arise from making investment decisions based on what has worked in the past. A decision taken to invest based on the prior knowledge that a strategy works will not likely lead to positive returns in the future and it is only when one begins to invest seriously that this realisation starts to sink in. In choosing strategies, investors more often than not invest in positive statistical fluctuations. A backtest will generally look better than its true performance due to the fact that we are distracted by unrealistically good performance, more likely due to good fortune in the past, which will not be reproduced in the future!

Of most importance in making an investment decision is, of course, whether the investment will make money in the future. We are all human, however, and as such susceptible to seeing backtests or the latest fashionable (and therefore recently performing) strategy as being an indication of short term, as yet unrealised, future riches. One can study the predictive power of in-sample backtested performance to see how well we can model the probability of future wealth. In so doing one finds that in-sample/past performance is not a good indicator of future results and, as such, should be taken with a big pinch of salt when allocating among managers or strategies.

We will proceed with this note as follows. We first present some empirical results from the reality of working in an investment firm, showing how in-sample returns are a highly biased estimate of out-of-sample performance. We then attempt to explain why this is the case with the help of random walks to simulate and model the process of finding strategies or managers. We also demonstrate the dangers for future performance of building overly complex strategies. Having shown, from many angles, the existence of the in-sample bias we finish by trying to advise on how best to mitigate the risk of in-sample overfitting.

The experience of working in an investment firm

The problem of in-sample bias is not restricted to systematic, quantitative investment as all investors are in some way guided by an observation of what worked well in the past. Even discretionary traders who invest based on "gut feeling", rather than backtests, still have this intuition built up through past market experience. What can differentiate systematic managers, however, is the existence of large data sets of past implemented models that can be used to study this in-sample bias in an effort to quantify the effect. CFM is no different in this respect with new models being regularly implemented in production and the go live date recorded in the database. A sub sample of this data is seen in **Figure 1**[1], which shows a recent backtest of a sub program of Stratus, CFM's flagship alpha program, with the realised performance of that same program superimposed.

This sub program is made up of seven independent clusters of models and approximately 80 individual (not necessarily independent) strategies. The simulated P&L obviously represents CFM's best effort to produce performance through this sub program and the simulation has indeed a very good Sharpe ratio of 2.6. The sub program, historically, has produced returns that are good compared to peers, but performance is someway off the levels seen in the paper traded backtest with a realised Sharpe ratio of 0.6. It is of course possible to argue that the firm has recently made a breakthrough and the performance is about to change for the better, the backtest being a taste of what is to come. However, this difference between "in-sample" past performance and realised performance has always existed over the history of the firm and the program. It could also be that our model of slippage severely underestimates the real costs of trading rendering our simulation an inaccurate representation of the realised performance. We have many years of experience of measuring and modelling costs, however, and have written extensively on the subject. We therefore feel that this is not a problem of cost estimation. With our database of trading strategies that have gone into production over the past 15 years or so we can also build a

[1] **Past returns are not a good indication of future results** as the CFTC disclaimer goes! This is precisely for the reasons discussed in this note.

picture of in-sample performance against out-of-sample performance averaged over all models.

In **Figure 2** we align all models against the date that they went into production, with zero being the first date they were used. Everything less than zero is therefore in-sample and everything greater than zero is out-of-sample. One sees in the plot a factor of 1.5 between the two - meaning the in-sample, past, simulated Sharpe ratio is 50% higher than that actually observed once the model is put into production. The plot itself is very difficult to build in an unbiased fashion as models can be slightly adapted, post implementation, and therefore one has a degree of freedom in the choice of the in-sample date, an effect which would only increase the factor of 1.5!

So, all of this begs the question, what is going on? Why on earth are we systematically unable to reproduce the encouraging returns promised in the backtest?

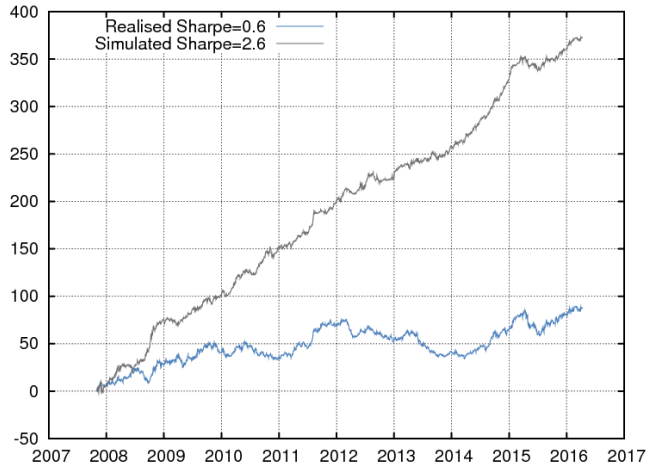

Figure 1: The realised performance for a sub-strategy of CFM's alpha program. Also plotted is the latest simulation net of costs and fees (the above is not net of fees! This should reduce Sharpe to ~2!) This difference has always existed and is due to in-sample bias.

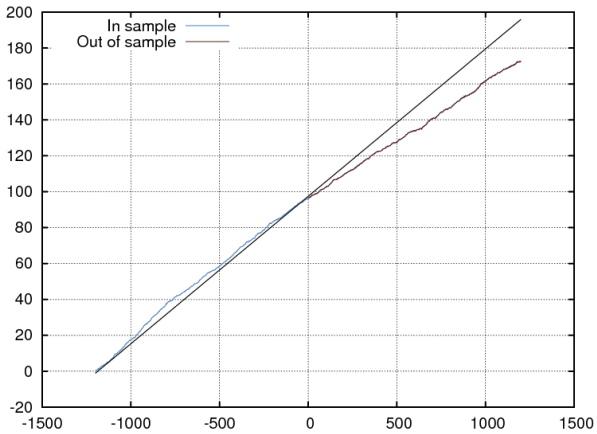

Figure 2: The performance of strategies implemented over the past 10 years or so in the same sub-strategy of CFM's alpha program. The implementation date for each strategy is lined up to fall at day 0. Days less than zero correspond to the average over all strategies' in-sample, simulated performance while days greater than zero correspond to the average over all data that was actually seen in production. We include strategies on all timescales, from intraday models to slower fundamental based models. Because the lines correspond to the average over many systems in the absence of costs, the Sharpe ratios are high – 15 in-sample and 10 out-of-sample i.e. in-sample performance is 50% better than out-of-sample performance!

Buying positive statistical fluctuations

We will use the technique of *random walks* to illustrate a few points concerning in-sample biases and the reader is invited to consult the appendix of this paper for a description of the technique. In essence, random walks represent independent realisations of strategies and are useful for simulating the environment within which we typically do research. For example, we may have a choice between two strategies, one with a Sharpe ratio of 1 and the other with a Sharpe ratio of 0. They may both have a real Sharpe ratio of 0.5 and the first has fluctuated up while the second has fluctuated down, but how would we know when just presented with the result of a backtest?

Let's illustrate further by showing some fictitious simulations of strategies from which we can choose to invest our hard earned cash. **Figure 3** shows the result of generating a number of zero Sharpe, zero expected gain random walks over a limited history of 10 years of data. Given that we know the Sharpe ratio is zero, it is obviously a no brainer that we would not invest in any of them. However, if we focus on the one with the biggest Sharpe ratio on a standalone basis then it is possible to be fooled by an attractive looking upwards fluctuation. We know that future performance can only be flat (on average), especially over a long history, but the positive fluctuation has enticed us in. This is our first source of in-sample bias, that due to a lack of statistical significance. The effects of such a bias are reduced by focusing on the level of significance of the strategy – if the fluctuation is consistent with a zero Sharpe random walk then don't invest. Statistical significance can only come from a longer backtest, which is not always possible however.

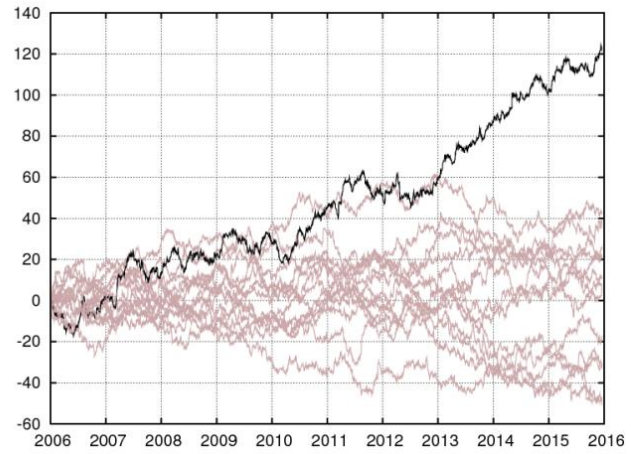

Figure 3: Zero Sharpe ratio random walks representing what one is trying to avoid in selecting genuine strategies. The black line is one such random walk which will easily be confused with a genuine strategy. The chances of selecting a positive fluctuation like this will be minimised by using longer backtests.

Now let's move on to some random walk simulations, generated this time with a low but positive Sharpe ratio of 0.5 as in **Figure 4**. This would be the typical level of true (future i.e. out-of-sample), expected (or hoped for!) Sharpe ratio for a strategy added to one of CFM's products. As is the case for any investor, in researching strategies, CFM looks to have an in-sample Sharpe ratio which is greater than a given threshold before being accepted; it is a brave man that invests in a flat or negative backtest! This threshold is typically higher than 0.5 in order to get statistical significance (being careful not to confuse the strategy with that zero Sharpe random walk!)

Let's apply a threshold of 0.7 (not far off what we would do in reality when researching slow fundamental based models) meaning, therefore, that we are likely to accept strategies which have fluctuated upwards and will reject those which have fluctuated downwards. Now we can look at the difference in Sharpe ratio between the in-sample period, prior to today but with a minimum Sharpe ratio, and the out-of-sample Sharpe ratio, which in this case is then exactly 0.5. **Figure 5** shows the plot with a result which resembles the real data in **Figure 2** with a ratio between in-sample and out-of-sample Sharpe ratios that looks very similar. This is our second source of in-sample bias, that due to only picking strategies with Sharpe ratios greater than a minimum. In so doing, we are more likely to accept strategies that have fluctuated upwards, while future performance can only be in line with the real Sharpe ratio. Even if we lower our minimum Sharpe ratio to say 0.5 i.e. the true Sharpe ratio, we will still find a kink in the plot between in-sample and out-of-sample performance, due to the fact that we reject the strategies that have fluctuated downwards. One has to wonder if the situation is improved by using longer backtests? It is true that using more data gives more significance and therefore we, in principle at least, could accept lower, and yet still statistically significant Sharpe ratios. However, we always impose a given minimum level of Sharpe ratio as being a pre-requisite for a strategy to be accepted – excessively low Sharpe ratios, even if they are statistically significant, come with their own problems!

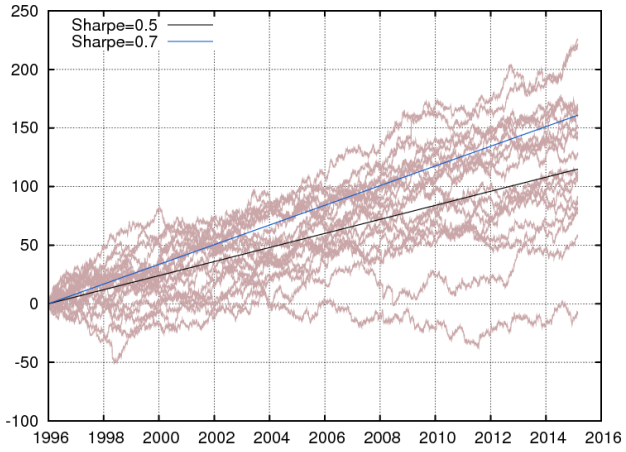

Figure 4: Random walks generated with a Sharpe ratio of 0.5. The black line corresponds to the average of the "spaghetti" of lines while the blue line corresponds to the level of Sharpe ratio researchers would typically demand as a minimum before being accepted into the portfolio of strategies.

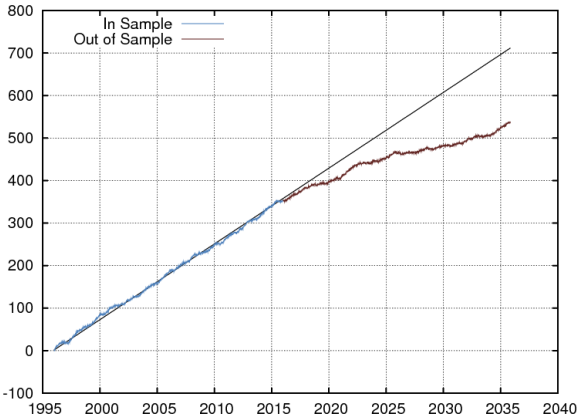

Figure 5: Shows a replication of **Figure 2** with random walks. In imposing a minimum Sharpe ratio of 0.7 for random walks generated with a Sharpe ratio of 0.5, our in-sample performance, the blue curve on the left half of the plot (data prior to 2016), has a Sharpe ratio closer to 1. The out-of-sample performance we see, as a result of our fictitious research, reverts back to the original 0.5.

A further source of in-sample bias arises from data over fitting or fine tuning strategies. Let's imagine a researcher finds a viable strategy and begins to tweak parameters to "finesse" the strategy, by adding small, what he feels to be valid and justified, incremental changes. He observes for instance that his strategy works better at the beginning and at the end of the week and therefore decides to de-weight the strategy on a Wednesday. He also observes

that the strategy works slightly better on high volume days and therefore gives extra weight to these days too.

In **Figure 6** we can try to represent this type of scenario again with fictitious random walk strategies. The researcher makes the first change and finds a Sharpe ratio improvement from 0.7 to 0.8, while the second change pushes the Sharpe ratio to 0.9. However, changes made to strategies are only accepted if they add positive performance (obviously) but one always has to be careful that these incremental changes are adding more than something statistically consistent with noise (those dreaded zero Sharpe random walks again!). If enough of these changes are made, each one being justified but not statistically significant, the in-sample performance will get better and better but will diverge from that seen in the future. This is another example of buying into positive fluctuations and being fooled into thinking that one is improving a strategy. This bias should be mitigated by using longer backtests but one still needs to be careful to make sure that changes to strategies are genuinely adding real, statistically significant information.

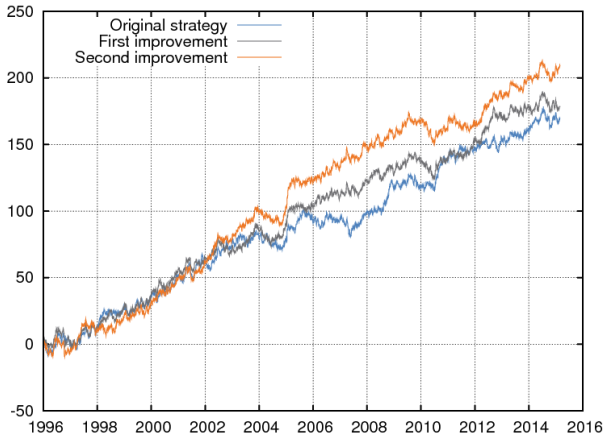

Figure 6: A fictitious scenario where a researcher starts with a strategy and makes incremental tweaks to improve it, passing from a Sharpe ratio of 0.7 to 0.8 and then to 0.9. These changes are, however, only accepted because the performance improves! In reality each change only adds information in a statistically insignificant fashion (zero Sharpe random walks that have fluctuated upwards again!). For this reason, the improvement will not be observed going forwards i.e. out-of-sample!

The dangers of complexity in building strategies

> With four parameters I can fit an elephant, and with five I can make him wiggle his trunk
>
> John von Neumann

The previous example of a researcher finessing a strategy is typical of an excessive data mining or parameter tuning scenario. The choice and range of parameters for any trading strategy is vast and that brings with it the temptation to play with these parameters and to complexify a strategy in order to improve the backtest.

Figure 8 shows another example of an overfitted strategy. Prior to our in-sample date of the beginning of 2008 we look at the average returns of the days of the month for a range of developed market equity indices and sovereign bonds. These markets have performed well over the period and our researcher tries to improve the long only strategy by modulating the size of his position as a function of past returns for each day of the month.

One such example is plotted in **Figure 7**, for the T bond future, with the points showing the average return for that particular day, grouping together the first day of the month and calculating the mean return, then the second day, then the third day etc. In the in-sample backtest, using these averages to tune our position gives an improved strategy, that is until one moves forwards past the date that was used to fit the data. The fact that we improve the backtest is not that surprising as we are adjusting positions on a day by day basis knowing that it gave better performance in the past. This is an example of a strategy with as many parameters as days, 30 in this case, and using these measured points indeed delivers good performance up until 2008[2], beyond which the out-

[2] Of course, in reality this strategy will trade every day and will not be profitable once trading costs are accounted for.

of-sample performance is just equivalent to our bare long strategy.

We can now try to "improve" the strategy by fitting a line to the points as seen in **Figure 7**. This fit is an attempt to reduce the number of parameters in the system, and seems like a good way to isolate only the biggest or most significant effects. This idea is indeed reflected in the fact that as we reduce the number of parameters to the fit we find that, as expected, the in-sample performance prior to 2008 worsens, but we converge towards the most significant effect – that of being long only. The in-sample period shows that more parameters gives better performance, whereas out-of-sample, as we reduce the number of parameters, the performance gets closer to that seen in-sample. This result could have been predicted before we performed the analysis – we did not need to fit anything at all to see that on average all points are positive and therefore to get the best forward looking performance, the best strategy is just to be long across all contracts.

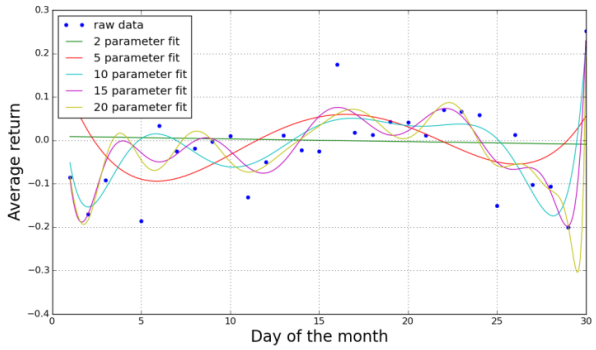

Figure 7: The average return of the T-bond future as a function of the day of the month. Day 1 for example is the average returns of all the 1st of the months, day 2 the average returns of the 2nd of the month etc. The lines correspond to a multi-parameter fit through these points. As we reduce the number of parameters we lose more and more of the detail but gain in robustness. The most robust pattern is that all average returns are slightly above zero, meaning that being long the T-bond future delivered positive performance over the backtest history.

[3] The t-stat can be considered as the distance from the zero Sharpe ratio random walk and calculated as the product of the Sharpe ratio and the square root of the number of years used in the backtest.

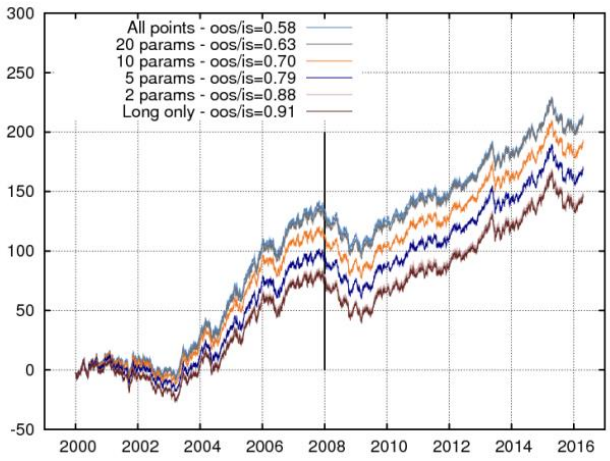

Figure 8: The result of using the fitted average returns per day of the month to modulate exposure to a long position in each bond and index future. On the left of the black line is the performance with the data used in the calculation of the average returns. One clearly sees that the more parameters that are used, the better the in-sample performance gets. On the right of the black line is the out-of-sample performance. We see no such improvement out-of-sample and see that the ratio of out-of-sample/in- sample performance Sharpe ratios get closer to 1 as we use less and less parameters.

Mitigating in-sample bias

In-sample bias is subtle and our hope is that we have managed to convey some of the nuances of the effect. It is of course fair to wonder how one can mitigate the effects of the bias in building strategies and portfolios. This depends on the context, but in general, getting higher levels of statistical significance will help. One good way of achieving this is to get higher Sharpe ratios, which is of course easier said than done. Higher Sharpe ratios are further from the noise and therefore one is less likely fooled by positive fluctuations. With higher Sharpe ratios, a researcher can also use shorter data sets to do his research and can then split data into in-sample periods to do research and out-of-sample periods to check research results, for example, by splitting a data set into two halves.

This is however more difficult when dealing with lower Sharpe ratios, which is more often the case for most investors. Taking the example of a strategy with a Sharpe ratio of 0.5, statistical significance (t-statistic[3] of 3-4) is only achievable with a backtest of 36-64 years, the length depending on the level of significance we need. This makes it difficult to play the game of splitting the data in two halves, with an in-sample research half and an out-of-sample test half, in that we severely limit our chances of achieving statistical significance in researching the

strategy. We need another technique, therefore, to deal with in-sample bias when researching low Sharpe ratio strategies. Our preferred way to deal with the problem at CFM is to try to minimise the amount of data mining. We try to conceive an idea before having looked at the data. We then build the strategy, using a standard statistical tool kit with no freedom for parameter choice, and construct the backtest. If the strategy is positive then it is accepted, if not then we drop the idea. This requires a good degree of self-discipline to not fall into the trap of trying to finesse the strategy to get that little bit more positive performance that will not be seen once the strategy is implemented.

For the slower, modest Sharpe ratio strategies we also try to gain in significance by gauging the level of plausibility of the strategy. This is perhaps the unscientific approach strategy selection but is complementary to just relying on statistical significance. This plausibility can come about through, say, getting an understanding of the driving forces behind the strategy, the effect being arbitraged may be anchored in human behaviour, something that can perhaps be demonstrated with empirical evidence. Plausibility can also come through out-of-sample tests using other related data. This qualitative plausibility is a key component to selecting strategies that can help to reduce in-sample bias and gain confidence and comfort with a particular strategy.

Conclusions

In-sample bias is a difficult concept to understand and convey and even more difficult to avoid in selecting trading strategies, in particular those with modest Sharpe ratios. We have tried to demonstrate and describe the effect from a few angles, explaining the source of the bias in strategy selection and suggesting techniques for minimising the risk. This bias doesn't only affect the selection of systematically implemented strategies, however, discretionary traders are also susceptible to applying trading ideas that have worked in the past and having their world view shaped by past events. The selection of managers or investment advisors is also prone to in-sample bias, in choosing a manager one is attracted to those who have performed well in the recent past, even if the outperformance is statistically insignificant. For the selection of a manager or a strategy it is imperative that investment decisions be based on statistically significant information and that, where possible, that information be supplemented by a less statistical plausibility argument – does the strategy make sense? Is it arbitraging an effect that is inherent to human behaviour and likely to persist? Have others made money in this way before? When selecting a manager, does his investment philosophy make sense? If the manager is underperforming, is it just because he has been unlucky but will likely perform in the future due to his sound and disciplined approach? These are all questions that will help to decipher positive statistical fluctuations from real signal in the case of a strategy or real skill in the case of a manager.

Additional reading

Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample Performance, DH Bailey, JM. Borwein, M. Lopez de Prado, and Q J Zhu, May 2014, Notices of the American Mathematical Society.

Predicting anomaly performance with politics, the weather, global warming, sunspots, and the stars, Robert Novy-Marx, Journal of Financial Economics, Volume 112, Issue 2, May 2014, Pages 137–146

. . . and the Cross-Section of Expected Return, Campbell R. Harvey, Yan Liu, and Heqing Zhu, Review of Financial Studies, October 9, 2015

Appendix
Random walks: a useful tool for explaining in-sample biases

The following may be considered technical by some readers; it may be safely skipped in order to get to the key results but we would like to try to answer these questions by simulating the process of strategy selection. We first begin by introducing the basic tool of numerical simulations - the random walk. In order to keep things as simple as possible we will only study time-series with constant levels of risk and Sharpe ratio.

With this in mind, the simplest random walk for a price p can be written as follows:

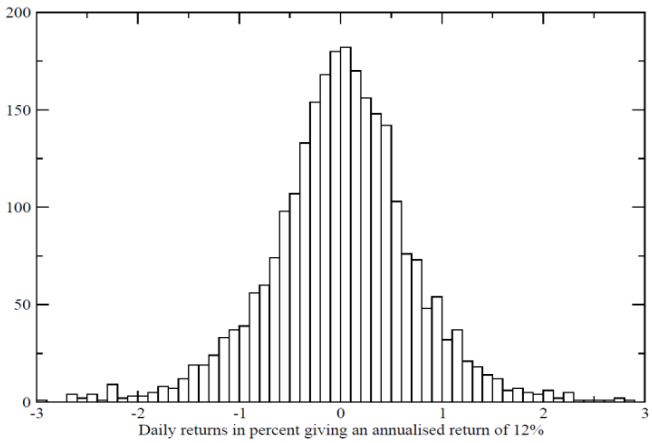

Figure 9: A histogram illustrating the bell shaped distribution of the random numbers used in the random walks. The random numbers are centered on zero and have tails that fit financial time-series well.

$$P_n = \sum_{n=0}^{N} (d + \eta_n)$$

where n is the counter, say the days for a daily return and N is the total number of days in the time-series of returns. The η term is simply a zero mean noise term or random number generator with a bell shaped distribution that best models the returns of the investment strategy. A histogram of these random numbers can be seen in Figure 9 showing a distribution centered on zero with tails representative of financial returns[4]. The d term is a constant added to the unpredictable "noise" η_n at every time step to generate a random walk with a "drift," or positive return. Figure 10 shows the results of generating random walks with Sharpe ratios of 0, 0.5 and 1 by varying the drift term to achieve the Sharpe ratio we require. Obviously, a Sharpe ratio of zero is generated by applying no drift term at all *i.e.* setting d to zero and allowing the zero mean of the ηn random numbers to generate a flat (on average) random walk with a Sharpe ratio of zero.

We now have a framework within which to simulate many random walks with any particularly desired Sharpe ratio, each realisation being different due to the existence of the ηn term. The time-series in Figure 10 shows how these random walks resemble different return streams, such as investment indices or individual funds.

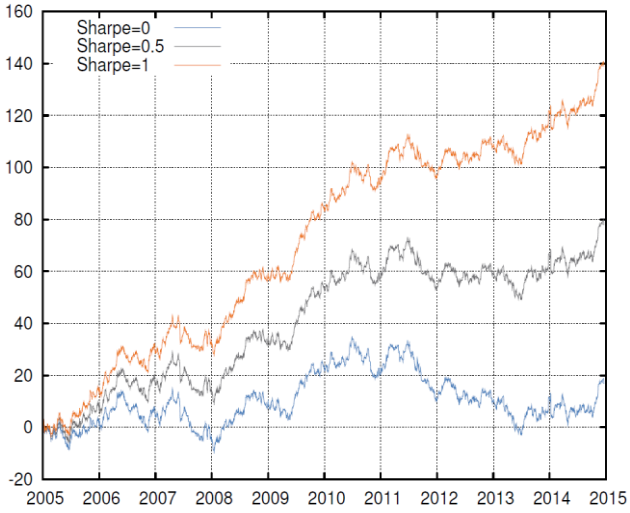

Figure 10: Random walks generated with three Sharpe ratios, illustrating how varying the d parameter allows us to easily change the drift and hence the Sharpe ratio.

[4] The choice of the distribution of returns can change the results of the study. Here we use a student's distribution with 4 degrees of freedom, a distribution which is naturally "fat tailed" and fits financial time-series well. For the purpose of this short note, however, we will neglect the effects of these fat tails on the calculation of correlations. Once could use the commonly known Gaussian distribution to achieve very similar results.

Important disclosures

This document is being provided at the request of the recipient for information and discussion purposes only, and should not be redistributed. This document does not convey an offer of any type and is not intended to be, nor should it be construed or used as investment, tax or legal advice or an offer to sell, or a solicitation of an offer to buy, an interest in any trading strategy or investment vehicle managed by Capital Fund Management SA ("CFM").

Any description or information involving investment strategies, process or allocations is provided for illustrations purposes only, may not be fully indicative of any present or future investments and is not intended to reflect anticipated performance.

Targets or objectives, including with respect to returns, volatility or leverage, if any, are used for measurement or comparison purposes only. Such targets or objectives reflect subjective determinations based on a variety of factors, including, among others, investment strategy, prior performance, volatility and leverage measures and expectations and market conditions. There can be no assurance that such targets or objectives will be met or met over any particular time horizon. Performance may fluctuate as can volatility and leverage. Targeted returns are not intended to be actual performance and should not be relied upon as an indication of actual or future performance.

All of the figures presented in this document are for illustrations purposes only, are unaudited and may not reflect the full cost structure of any investable fund or product. No representation is made that CFM's or a fund's risk management, investment process, trading performance, investment objectives or the control of operational risks, credit risks and other risks involved in any trading strategy will or are likely to be achieved or successful or that any fund or underlying investment strategy managed by CFM will make any profit or will not sustain losses. Any measure of risk is inherently incomplete and does not account for all risks or even all material risks such as risks due to unforeseen catastrophic events.

Any statements regarding market events, future events or other similar statements constitute only subjective views, are based upon expectations or beliefs, involve inherent risks and uncertainties and should therefore not be relied on. Future evidence and actual results could differ materially from those set forth, contemplated by or underlying these statements. In light of these risks and uncertainties, there can be no assurance that these statements are or will prove to be accurate or complete in any way.

HYPOTHETICAL PERFORMANCE RESULTS HAVE MANY INHERENT LIMITATIONS, SOME OF WHICH ARE DESCRIBED BELOW. NO REPRESENTATION IS BEING MADE THAT ANY ACCOUNT WILL OR IS LIKELY TO ACHIEVE PROFITS OR LOSSES SIMILAR TO THOSE SHOWN. IN FACT, THERE ARE FREQUENTLY SHARP DIFFERENCES BETWEEN HYPOTHETICAL PERFORMANCE RESULTS AND THE ACTUAL RESULTS SUBSEQUENTLY ACHIEVED BY ANY PARTICULAR TRADING PROGRAM.

ONE OF THE LIMITATIONS OF HYPOTHETICAL PERFORMANCE RESULTS IS THAT THEY ARE GENERALLY PREPARED WITH THE BENEFIT OF HINDSIGHT. IN ADDITION, HYPOTHETICAL TRADING DOES NOT INVOLVE FINANCIAL RISK, AND NO HYPOTHETICAL TRADING RECORD CAN COMPLETELY ACCOUNT FOR THE IMPACT OF FINANCIAL RISK IN ACTUAL TRADING. FOR EXAMPLE, THE ABILITY TO WITHSTAND LOSSES OR TO ADHERE TO A PARTICULAR TRADING PROGRAM IN SPITE OF TRADING LOSSES ARE MATERIAL POINTS WHICH CAN ALSO ADVERSELY AFFECT ACTUAL TRADING RESULTS. THERE ARE NUMEROUS OTHER FACTORS RELATED TO THE MARKETS IN GENERAL OR TO THE IMPLEMENTATION OF ANY SPECIFIC TRADING PROGRAM WHICH CANNOT BE FULLY ACCOUNTED FOR IN THE PREPARATION OF HYPOTHETICAL PERFORMANCE RESULTS AND ALL OF WHICH CAN ADVERSELY AFFECT ACTUAL TRADING RESULTS.

CFM has pioneered and applied an academic and scientific approach to financial markets, creating award winning strategies and a market leading investment management firm.

Contact us

Capital Fund Management S.A.
23, rue de l'Université
75007 Paris, France
T +33 1 49 49 59 49
E cfm@cfm.fr

CFM International Inc.
The Chrysler Building, 405 Lexington Avenue - 55th Fl.,
New York, NY, 10174, U.S.A
T +1 646 957 8018
E cfm@cfm.fr

CFM Asia KK
9F Marunouchi Building, 2-4-1, Marunouchi, Chiyoda-Ku,
100-6309 Tokyo, Japan
T +81 3 5219 6180
E cfm@cfm.fr

Capital Fund Management LLP
64 St James's Street, London
SW1A 1NF, UK
T +44 207 659 9750
E cfm@cfm.fr