The Search for Successful Active Asset Management

A Case of Skill or Just Pure Luck
Introduction

The asset management industry is an industry that delivers investing as a service. Large numbers of clients, like pension funds, insurance companies and wealthy individuals, use these services. The size of this industry is substantial. The assets under management by the world’s largest 500 asset managers at the beginning of 2009 alone are already 69.4 trillion dollars. In this large industry differences exist in how skilled the managers are in investing. This is indicated by big performance differences between different managers. Skilled managers add value and unskilled managers lose value on the long run. This is something of big concern with the clients who pay significant fees for the investment services. Also socially this is a question of concern, if for example pensions funds get into trouble due to poor investments, this directly affects the people who depend on the pension payments for their old days.

Unfortunately for clients, it is generally difficult to distinguish skilled managers from unskilled managers. It is not just a case of looking at good and poor performance and conclude that the a manager is skilled or unskilled, respectively. An important reason is that performance figures are also a result of luck. To reduce the risk of putting money in the hands of an unskilled manager, different analysis and testing methods have been developed. Although these analysis vary from very qualitative questions to various kinds of statistical analysis, they all look for the same answer: is the asset manager under investigation capable of achieving good performance on the long run, does he possesses skill in investing.

In the category of statistical analysis, tests have been developed that focus on the performance series of asset managers. These tests try to identify skill by looking for positive consistencies in these series. These kind of tests are widely used, though they easily can lead to wrong conclusions about asset managers. The main problem generally is that statistical testing procedures often demand that the data meets stringent requirements, else the outcomes are invalid and unreliable. Also an important problem concerns the fact that test methods, even when the test are valid, can identify managers with good performance as skilled, while in reality it is a result of luck. In statistics this problem is referred to as data-snooping. Finding skilled managers is a problem of distinguishing the seemingly skilled but lucky managers from the real skillful managers.

The goal of this thesis is to make inference on a broad spectrum of statistical tests. The analysis described looks at how tests behave when used testing performance data, how they react to survivorship bias and how sensitive they are to data-snooping bias. Subsequently also an empirical study will be done to see if there is evidence in favor of skill.

In the first chapter the concepts of skill and luck are elaborated, the universe of asset managers it applies to is established and a literature overview is given that describes the theoretical and historical background of

---

1 Source: Pensions & Investments/ Watson Wyatt World 500 ranking
luck, skill and of performance analysis in this context. Chapter 2 will proceed by describing the general and statistical characteristics of performance data that can influence the testing. In chapter 3 the different tests are described. The theoretical implications are given if the different test are used on data, that have the characteristics described in chapter 2. In part 4 of the thesis the simulation is clarified and the results are treated. In chapter 5 the results of the empirical study on set of 108 managers will be presented.
Chapter 1. Skill and luck, testing and data-snooping

In the asset management industry the terms skill and luck are of central debate. Also the process of testing to distinguish skill from luck has been subject to a lot of research. This thesis addresses problems when testing for skill with the following question: ‘how well do performance measures, based on historical returns, behave when tested and how well do these tests distinguish lucky managers from skilled managers’. This first chapter will clarify on the concepts of skill and luck. An important problem in testing for skill, the data-snooping bias, will be discussed and finally at the end of the chapter a short literature overview is provided on the theoretical and historical background of statistical analysis based on historical performance.

1.1 Luck and skill in investing

In this thesis the term asset management will refer to the professional management of financial assets by organizations, asset managers, for clients. Examples of clients are (wealthy) individuals, institutions, insurers and pension funds. Asset managers have a general goal to add value by utilizing a strict philosophy and investment process, referred to as a product, aimed at the long term. Note that asset managers can offer more than one product. Important of this definition is that it excludes any investor or player in financial markets who does not offer asset management as a service, who does not utilize a strict investment philosophy and investment process and who does not aim at the long term. Examples are speculators and market makers.

The examined products use an index as a benchmark in the investment process. This benchmark represents the market the product is invested in. For example, if an asset manager manages a product that is benchmarked against the S&P 500 index, the investment universe of the product are the companies held in the S&P 500 index. It must be noted that these kind of products often invest a proportion of the portfolio outside the benchmark, though in general these securities are comparable to the securities in the benchmark.

The benchmark products can be divided into two categories (Grinold et al, 2000): passive management and active management. Passive management, also called passive investing or indexing, is an investment approach with the aim of replicating the benchmark returns. This management style is considered passive because portfolio managers don't make active bets on securities; they simply try to track the index returns. Active management, or active investing, is when the manager tries to ‘beat’ the benchmark. In other words the manager tries to achieve returns in excess of the benchmark returns (excess returns). In this approach an investor invests in securities from the benchmark, but intentionally deviates from the composition of this benchmark by giving different weights to the securities in the benchmark. The weights are determined by the investment process and philosophy. Next to the two main categories, also strategies have evolved that lie between active and passive investing, called enhanced management.
In the context of this thesis the concepts of luck and skill relate to investment products that are benchmarked and actively managed. Managers who have a solid long term investment approach that generates excess returns, are defined as managers that have skill. Skill implies that the manager achieves outperformance due to a solid conviction and competence in active investing. Managers who do not have skill, will not be able to consistently beat the benchmark in the long run. Luck concerns the chance of accidentally achieving higher returns than the benchmark. Managers who do not have skill but who have good luck can outperform the market, which can lead to the false conclusion that the manager has skill. The opposite is also possible. Due to bad luck a skilled manager can show underperformance.

The two concepts can be shown in two dimensions, see figure 1.1. With this two dimensions managers can be divided in four categories. There are the managers that fail to beat the benchmark, in the figure they are represented under the horizontal axis. In the bottom left quadrant is the category of unlucky and unskilled managers, the doomed managers. These managers do not have skill to outperform and they also experience bad luck. In the bottom right quadrant are the skillful but unlucky managers, the forlorn managers. The managers in this group unfortunately do not achieve superior returns although they have skill. The two categories in the top quadrants present the managers who beat the benchmark. These are the unskilled but lucky managers, the ‘Insufferable Managers’, and the lucky and skillful managers, the ‘Blessed managers’.

![Figure 1.1. Skill and luck in two dimensions (source: Grinold et al, 2000)](image)

When managers are tested for skill with performance measures and tests methods based on past performance, the problem of identifying skillful managers becomes the challenge of separating the ‘Insufferable Managers’ from the ‘Blessed Managers’. The managers in the upper and lower quadrants are easily identified from each other by respectively negative relative performance and positive relative performance. The ‘Doomed Managers’ and ‘Forlorn managers’ are unfortunately not distinguishable using analysis based on past performance. Both types of managers historically did not produce returns that reflect any skill. The ‘Insufferable Managers’ and ‘Blessed Managers’ are subject of research when using statistical methods based on past performance.
1.2 Data-snooping Bias

With the use of statistical methods, an important issue that is generally encountered is data-snooping (White, 2000). In the case of analyzing performance of managers, data-snooping refers to falsely distinguishing a lucky, not-skillful manager as skillful. In other words, after analysis it is concluded that manager X has skill while in reality he is of the ‘Insufferable Manager’ category.

In the context of hypothesis testing data-snooping, also referred to as the data-snooping bias, is a form of type 1 error. When a manager is tested for skill the null hypotheses generally states that the manager has no skill. When it is falsely concluded that the manager has skill, the null is falsely rejected and therefore a type 1 error is made. The data-snooping bias refers to the multiple type 1 errors made when a set of managers is tested.

The problem has been often addressed, though it was not solved for a long time (Griffioen, 2003). Finally White (2000) came up with a procedure for testing performance while correcting for the data-snooping bias. Later Hansen (2005) adjusted this test and corrected for some statistical issues.

1.3 Previous research and literature in performance measurement

In Academics it remains a question of extensive research whether active management is capable of outperforming benchmarks in the long run.

Empirical research in the first half of the 20st century was mainly concentrated on technical analysis and forecasting methods. This research did not find any evidence that the capability existed to outperform a benchmark or market. This gave rise to theories and conclusions that investing strategies based on past information do not generate excess return on the long term. Academics started to define return series as random walks. This ‘Random Walk Hypotheses’ states that trends in prices are spurious and purely accidental (Griffioen, 2003). According to this theory it is not possible to consistently generate excess returns. Academics started to pursue theories that explained the random walk behavior of returns and stock prices. This led to the development of the efficient market hypotheses (EMH) by Fama. Three forms can be distinguished according to Fama (1970): the strong, the semi-strong and the weak form. The weak form states that investors cannot outperform the benchmark or market using only historical prices. The semi-stringent form states that investors cannot outperform the market using all publicly available information. The strong form states that the prices on the market contain all relevant information, so outperformance is not possible.

In the same period the Capital Asset Pricing Model (CAPM) was developed by Sharpe (1964) and Lintner (1965). According to the CAPM, expected excess returns of portfolios other than the market portfolio are entirely determined by the sensitivity of these portfolios to the excess market return. In this context the returns are in excess of the risk free rate. According CAPM all
return differences that cannot be explained by this sensitivity to the market portfolio, are the result of chance. As a consequence on the long term these excess returns should tend to zero.

Under the market efficiency assumption\(^2\) CAPM indicates that no investor in the market that deviates from the market portfolio is capable in the long term of adding any value uncorrelated with the market return. Although CAPM is not the same as the efficient market hypotheses, they are consistent. According the two theories it is not possible to achieve consistent outperformance (Grinold et al, 2000).

A lot of empirical research in active asset management concentrated on finding evidence whether skill existed or not. These studies for skill can be divided into two approaches.

The first approach is about persistence in managers performance data. Testing for performance persistence is about testing if a winner in one period is a winner in a subsequent period. Mixed results have been found. Some studies do not find evidence for skill. Examples of these studies are Jensen (1968), who looked at 115 mutual funds over the 1945-1964 period and Carhart (1997), who analyzed 1892 diversified equity mutual funds over the 1962-1993 period. Examples of studies that do find evidence of skill are Grinblad and Titman (1992) who looked at 157 mutual funds from 1968 to 1982 and Goetzman and Ibbotson (1994), who studied 728 mutual funds over the 1976-1988 period. Survivorship bias and the effect on the power of the different test methods used also have been subject to analysis. Different studies have studied how the tests react on this problem. Examples are Brown et al (1992) and Carpenter et al (1999).

A second approach looks for skill by identifying consistent uncorrelated return with the market return through regression analysis. The regression based analysis are variations of the original CAPM specification. In general the studies did not provide significant evidence in favor of skill. Examples of regression analysis are Jensen (1968), who developed Jensen’s alpha, Fama and French (1993) who developed the three factor model and Carhart (1997) who developed the four factor model. Also performance measures have been developed to test for skill. These measure try to identify skill by testing for consistent excess returns. Examples of performance measures developed are the Sharpe Ratio (Sharpe, 1966) and the Information Ratio (Goodwin, 1998).

\(^2\) Market Efficiency Assumption: all information available is known by all investors and thus all investors are the same (Grinold, 2000)
Chapter 2. Mutual Fund Performance Data and simulation

The behavior and results of statistical tests on historical performance depend mainly on general and statistical characteristics of the examined data. To make inference on the behavior of these tests through a simulation, the simulated data should be in line with the data generating process (DGP); the general and statistical characteristics should be known and accounted for.

In this chapter the general and statistical characteristics of performance series of products will be addressed. The general characteristics indicate if the return series are comparable in analysis. The characteristics in this context are:

- Characteristics of the product portfolio: styles, capitalization and the benchmark
- Source of outperformance: stock picking and benchmark timing

The statistical characteristics of the data influence the validity of different test methods. Besides the more regular financial time series characteristics as heteroskedasticity and serial correlation there are the following statistical issues:

- Survivorship bias and fund attrition
- Crossectional Correlation and Crossectional heteroskedasticity

2.1 Data characteristics

Investment styles, correct benchmarks and the dataset

When testing for skill in beating a benchmark, tests based on historical performance look at historical excess returns. To get valid test results, the excess returns should be correctly derived. In this context it is important that the benchmark is correct. Characteristics of the benchmark should be in line with the characteristics of the examined product.

Common discrepancies between benchmark and portfolio are on the level of capitalization grade and in terms of investment style. Capitalization grade concerns the average capitalization of the portfolio stocks. Investment styles refer to the style bias of a product. A portfolio has a style bias when it is particularly concentrated in certain style stocks. Three main styles are defined: value, growth and core. When a manager has a value bias he concentrates investments in value stocks, stocks that tend to trade at a lower price relative to its fundamentals (i.e. dividends, earnings, sales, etc.). A growth bias refers to a tendency to concentrate investments in stocks of companies whose earnings are expected to grow at an above-average rate relative to the market. The core style refers to an investment style without a significant bias to value or growth stocks. An example of this
kind of mismatch is when a manager tries to beat the Russel 2000 index, but he does so by creating a value bias. By doing so the manager takes on exposure that is more correlated to the Russel 2000 Value index then to the Russel 2000 index. The Russel 2000 value index is a version of the Russel 2000 index that is more concentrated in value stocks. The excess returns of the product are better calculated with the value version of the Russel 2000 then with the normal version.

To make sure the excess returns do not get blurred by benchmark misspecification, the products used throughout the thesis must apply to the following restrictions:

- Only equity products
- The products are benchmarked against the S&P 500
- Only products with a reported core investment approach
- The primary capitalization of the product portfolio is large cap, thus companies with a capitalization grade of $5 billion or more
- The strategy is actively managed, not passively or enhanced
- The product is long only, no long/short extensions

Stock picking or benchmark timing

Sources of outperformance in active management can be separated in benchmark timing and stock picking. Benchmark timing is the practice of altering the exposure to the systematic risk of the benchmark. The exposure is raised or lowered, based on expectations of the direction of the benchmark returns. Managers that apply this technique can successfully beat the benchmark with superior returns. Stock picking refers to buying and selling the right stocks at the right time, in order to generate excess return uncorrelated with the benchmark. The scope of this thesis is to identify skill in stock picking, not in benchmark timing. This choice is made to keep the products comparable (Grinold et al, 2000).

Removing managers that apply benchmark timing, can be done through regression analysis. For this thesis the following regression is used, proposed by Grinold et al (2000):

$$r_{p,t} = \alpha_p + \beta_p * r_{b,t} + \gamma_p * Max(0, r_{b,t}) + \epsilon_p(t)$$

In this regression the return of portfolio, $r_{p,t}$, is regressed on the benchmark return, $r_{b,t}$. The $\alpha_p$ indicates the average uncorrelated return with the benchmark. The $\beta_p$ stands for the exposure to systematic risk of the benchmark, the beta. A second systematic risk exposure parameter is added, $\gamma_p$. The equation includes a downward market beta, $\beta_p$, and a upward market beta, $\beta_p + \gamma_p$. When $\gamma_p$ is positive and significantly different from zero, it cannot be rejected that the manager applies benchmark timing in
the product. When this is the case the product will not be used or analysed in this thesis.

This is one method to identify benchmark timing. Other model specifications with an upward and downward beta for the regression are also used in practice. Although the method of a simple regression seems quick and dirty, it has been shown affective in previous empirical studies (Grinold et al, 2000).

2.1 Statistical characteristics of return data

Survivorship bias and fund attrition

One of the main issues addressed in the literature on performance data of investment products, is the widespread use of survivorship biased datasets. A dataset is survivorship biased when products that are shut down, are not present. It is very common that databases the sets are collected from suffer from survivorship bias and in general they do not allow to either study or correct for it. As a consequence a lot of earlier studies ignored fund attrition, what makes the conclusions less reliable (Elton et al, 1996).

Statistically survivorship bias is equivalent to truncation from below. The main problem with data that is truncated from below is that multivariate statistical tests on such samples are biased and misleading. When assumed that in general a product shuts down when a manager experiences poor performance, multivariate testing of performance measures often will overstate measured performance when the studied universe is survivorship biased (Elton et al, 1996).

To see how the tests behave when testing survivorship biased data, this will be controlled for in the simulation. In chapter 4 this will be elaborated.

Crosssectional Correlation and Crosssectional heteroskedasticity

In the investment products asset managers apply different investment processes that lead to different portfolios. A direct consequence is that the volatility will differ between the portfolios. In a multivariate system of performance series this leads to cross sectional heteroskedasticity. Though asset managers utilize different investment approaches, managers show herding behavior (i.e. buying and selling of the same (kind of) stocks at the same time) (Grinblatt et al, 1995). This herding behavior can result in significant correlation between the portfolios of the different products. This correlation leads to cross sectional correlation between different return series. In literature it is found that the herding
behavior is mainly due to the tendency to buy the same stocks that were ‘winners’ in the past (Grinblatt et al, 1995). Data series in datasets that show cross sectional correlation will result in correlated performance measures.

Cross sectional correlation and cross sectional heteroskedasticity result in a multivariate distribution with a covariance matrix that has nonzero off diagonal elements and diagonal elements that are not the same. So the distribution needed for multivariate testing of a system of performance series is directly affected by the cross sectional correlation and the cross sectional heteroskedasticity.

When testing and simulating, this correlation and heteroskedasticity will have to be taken into account. Chapter 4 will elaborate on how these characteristics are incorporated in the simulation.
Chapter 3. Test procedures and test statistics

To identify skillful managers with analysis of the historical performance series, different performance measures and test procedures have been developed. To make inference on the reliability and behavior of these performance measures and tests, they are simulated. This chapter highlights the performance measures and test procedures examined. Their statistical properties and problems will be discussed.

The tests range from simple methods to more sophisticated test procedures. A broad range is chosen to give clear insights on how the measures react to different levels of testing. Especially simple invalid methods are widely used. To illustrate their shortcomings they are also incorporated in the simulations.

This chapter consists of three parts. The first part will handle basic statistical tests of three performance measures, that are only (asymptotically) valid under very stringent assumptions on the performance data. These assumptions generally do not apply to this data, consequently this leads to invalid inference. The second part highlights tests that are based on stationary bootstrap simulations instead of asymptotic theory. These methods have the advantage that they do not rely on stringent assumptions about the performance data to be asymptotically valid. The last part will elaborate on test methods that also take into account data-snooping bias. This test are White's Reality Check method (RC) and Hansen's Superior Predictive Ability method (SPA). These tests are based on the same stationary bootstrap simulations as the tests in the second part. The main difference with the ordinary stationary bootstrap is that they use the whole dataset of performance series to make a more conservative estimation of p-values, therefore reducing the data-snooping bias.

3.1 Simple tests

The mean excess return and a simple t-test

The first performance measure that is tested is the mean excess return. Skillful managers should have a significant positive true value of the mean excess return. The sample value of the mean excess return looks as follows:

\[
\text{Sample mean excess return } = \bar{er} = \frac{1}{T} \sum_{t=1}^{T} (r_t - r_{b,t})
\]

The null hypotheses of no skill is \( H_0: E(er) = 0 \). The null hypotheses will be tested using a standard t-test. The t-statistic:

\[
T_{mer} = \sqrt{T} \cdot \frac{\bar{er}}{\hat{\sigma}_{er-r_b}}, \quad \text{where } \hat{\sigma}_{er-r_b} = \sqrt{\frac{\sum_{t=1}^{T} (r_t - r_{b,t} - \bar{er})^2}{T-1}}
\]
Although the simplicity of this measure and procedure seem appealing, this measure has some statistical shortcomings. The test is subject to data-snooping, by chance a manager without skill can show a significant positive mean excess return. Another issue is the possible presence of heteroskedasticity and serial correlation in the data. These problems influence the estimate of the standard deviation, \( \hat{\sigma}_{r-r_f} \), which is often underestimated. In turn the power is not accurate anymore.

**The Sharpe ratio criterion and a simple t-test**

Besides the mean excess return another well known performance measure is the Sharpe ratio. This ratio measure relative performance over the risk free rate, per unit of risk. So the higher the ratio, the better the manager is capable of achieving excess return per unit of risk. Whether the manager has skill or not can be evaluated by the Sharpe criterion. The criterion states that if the manager has a significantly higher Sharpe ratio then the benchmark has, the manager has skill.

\[
\text{Sample Sharpe criterion} = \bar{sr} = \frac{\frac{1}{T} \sum_{t=1}^{T} (r_t - r_f)}{\hat{\sigma}_{r-r_f}} - \frac{\frac{1}{T} \sum_{t=1}^{T} (r_{b,t} - r_f)}{\hat{\sigma}_{b-r_f}}
\]

The null hypotheses of no skill, \( H_0: E(sr) = 0 \), can again be tested with a simple t-test, based on the standard normal distribution, according to asymptotic theory.

\[
T_{sr} = \sqrt{T} \frac{\bar{sr}}{\sqrt{2}}
\]

This statistic suffers from the same statistical issues as the simple t-test on the mean excess return.

**Jensen’s Alpha and simple regression analysis**

The development of the CAPM model and the discussion about market efficiency in the 1960’s led to the rise of regression-based performance models. The most basic analysis of this kind is Jensen’s alpha. Jensen first proposed this measure in 1968 (Grinold et al, 2000). The measure looks for the return component of the gross returns that is uncorrelated with the benchmark.

\[
\text{Sample Jensen’s Alpha} = \hat{\alpha} \text{ based on } r(t) = \alpha + \beta \cdot r_b(t) + \epsilon(t)
\]

Jensen’s alpha is the estimator of the intercept of the regression. The null hypotheses of no skill, \( H_0: E(\alpha) \leq 0 \), can be tested using a standard t-test.
Although the wide use of this measure, it is subject to data-snooping. Also the estimator and the power of the t-test are invalid by the presence of heteroskedasticity and serial correlation. A third issue with this measure is the assumption that the data generating process can be modeled according the CAPM relation. If this assumption is wrong and the model suffers from misspecification, the estimate is biased and the test is invalid.

3.2 Test based on the stationary bootstrap method

The Stationary bootstrap

Testing methods based on asymptotically approximated distributions are only valid when stringent assumptions apply to the tested data. In the case of performance series of investment products these assumptions are not met. An alternative to approximate the distributions of the performance measures is the bootstrap method. The original bootstrap method was introduced by Efron in 1979. Instead of using asymptotic approximations of distributions, the bootstrap method approximates distributions by resampling the available data (L. Horowitz, 2000).

Often it is difficult and involves hard mathematical computation to find the asymptotic distributions of test statistics. The bootstrap method substitutes these mathematics. Under the appropriate conditions the bootstrap yields an approximation to the distribution that is at least as accurate, and in finite samples often more accurate, as the approximation obtained from first-order asymptotics. Besides the computational difficulties that are relieved, an advantage of the bootstrap is that the empirical character of the method makes it valid under mild regularity conditions. This is an appealing property for the performance series, that have the statistical characteristics described in chapter 2. In this thesis the stationary bootstrap of Politis and Romano(1994) is applied. This is a variant of the original bootstrap methodology that preserves serial correlation in the underlying data.

The bootstrap estimation method can be described as follows. Take the return sample of product $r_t$, $t = 1, \ldots, T$, and the benchmark return sample, $\{r_{b,t}: t = 1, \ldots, T\}$. These returns are used to calculate a performance measure $S_n = S_n(r_{n,1}, \ldots, r_{n,T}; r_{b,1}, \ldots, r_{b,T})$. Let $G_n(\tau, F) = P(S_n \leq \tau)$ be the cumulative distribution function (cdf) of $S_n$, which usually depends on the cdf of the underlying return data, $F$. To make inference on the performance measure $S_n$ the CDF $G_n(\tau, F)$ has to be approximated. The bootstrap method provides an approximation to the cdf by replacing the unknown distribution function $F$ by an estimate for this distribution, referred to as $\tilde{F}$. In this thesis $\tilde{F}$ is estimated by the empirical distribution function (edf) of the underlying data. To find the approximation $G_n(\tau, \tilde{F})$, many bootstrap samples are generated. A bootstrap sample of size $T$ is generated by sampling the
underlying data randomly, with replacement. For every sample $S_n$ is calculated. The realizations of the performance measures from the bootstrap samples are indicated by an astrix. $K$ samples are generated. Eventually from the $K$ replications of the test statistic, the empirical distribution derived and used as the estimate for $G_n(\tau, F)$.

Stationary bootstrap realization of $S_n = S_{n,k}^*$ where $k = 1, ..., K$

The minimal criterion for the estimated distribution to be valid is that it has to be consistent. The following condition, by Mammen (1992), provides necessary and sufficient conditions for the bootstrap estimate of the cdf of the statistic to be consistent (Horowitz, 2000).

**Theorem 3.1 (Mammen 1992):** Let $\{X_i = 1, ..., n\}$ be a random sample from a population. For a sequence of functions $g_n$ and a sequence of numbers $t_n$ and $\sigma_n$, define $\bar{g}_n = n^{-1} \sum_{i=1}^n g_n(X_i)$ and $T_n = (\bar{g}_n - t_n) / \sigma_n$. For the bootstrap sample $\{X_i^* : i = 1, ..., n\}$, define $\bar{g}_n^* = n^{-1} \sum_{i=1}^n g_n(X_i^*)$ and $T_n^* = (\bar{g}_n^* - \bar{g}_n) / \sigma_n$. Let $G_n(\tau) = P(T_n \leq \tau)$ and $G_n^*(\tau) = P^*(T_n^* \leq \tau)$, where $P^*$ is the probability distribution induced by bootstrap sampling. The $G_n^*(\cdot)$ consistently estimates $G_n$ if and only if $T_n \overset{d}{\rightarrow} N(0,1)$.

The main condition this theorem states is that the data has to be independent over time and identical distributed (iid). The performance series of products do not satisfy this property. As described in chapter 2 the series are not independent over time and also not identically distributed due to the presence of heteroskedasticity. It is known that the bootstrap works reasonably well when the data is not identically distributed (Politis et al., 1997). Therefore the heteroskedasticity in the data will be assumed not to impose problems. That the data is not independent is of bigger concern. It impacts the validity of the estimate of the distribution. Politis and Romano came with an adapted bootstrap method called the stationary bootstrap that addresses this problem.

The stationary bootstrap randomly resamples blocks of observations instead of individual observations. The resampling of blocks of succeeding observations mimics the dependence between the observations of the original data. Early block resampling techniques chose a fixed length of block-length. Politis and Romano developed the stationary bootstrap which resamples blocks of random length. The length is randomly generated according a geometric distribution. With this method they solved stationarity issues of the fixed block methods. A second improvement was made by connecting the last observation to the first observation in the original data set when resampling. A consequence is that when the last observation of the original data set falls in a block, the next observations will be the first observations from the original data set. This makes sure that every observation has the same chance of appearing in the resampled bootstrap set.
Note that for estimates to be consistent, the average block length must increase with an increasing sample size. Also the statistics still have to satisfy theorem 3.1 for consistency (Horowitz, 2000).

The tests

The performance measures tested with the stationary bootstrap are the mean excess return, \( \overline{r} \), and the Sharpe criterion, \( \overline{s} \). An additional measure that will be tested is the Information Ratio. The information ratio measures the mean excess return per extra unit of risk taken over the benchmark. The information ratio, simply the ratio of the mean excess return and the standard deviation of the excess returns, seeks to summarize the mean-variance properties of a portfolio (Godwin, 1998).

\[
\text{Sample Information Ratio } = \frac{\text{TR}}{\text{IRR}} = \frac{1}{T} \sum_{t=1}^{T} \frac{(r_t - r_{b,t})}{\sigma_{r-r_f}}
\]

To make sure that the bootstrap approximations will be accurate, the statistics have to satisfy theorem 3.1. For this, the central limit theorem is required. To ensure that the central limit theorem is satisfied, the following assumption must apply (Hansen, 2005):

Assumption 3.1. The series \( \{d_t : t = 1, ..., T\} \) is (strictly) stationary and \( \alpha \)-mixing of size \( -(2+\delta)(r+\delta)/(r-2) \), for some \( r > 2 \) and \( \delta > 0 \), where \( E|d_t|^{r+\alpha} < \infty \) and \( \text{var}(d_t) > 0 \).

This stationarity assumption is made to ensure that the population moments are well-defined. The alpha-mixing part of the assumption implies that the observations temporally far apart from one another are assumed to be nearly independent. The series \( \{d_t : t = 1, ..., T\} \) in the context of the performance measures apply to:

- mean excess return: \( d_t = r_t - r_{b,t} \)
- Sharpe ratio: \( d_t = \frac{(r_t - r_{f,t})}{\sigma_{r-r_f}} - \frac{(r_{b,t} - r_{f,t})}{\sigma_{r_b-r_f}} \)
- Information ratio: \( d_t = \frac{(r_t - r_{b,t})}{\sigma_{r-r_f}} \)

Under these conditions the central limit theorem can be applied to the three statistics:

\[
S_1 = T^{1/2}(\overline{r} - E(\overline{r})) \overset{d}{\rightarrow} N(0, \sigma_1^2), \quad \text{where } \sigma_1^2 = \lim_{t \to \infty} \frac{1}{T} \sum_{t=1}^{T} \text{var}(r_t - r_{b,t})
\]

\[
S_2 = T^{1/2}(\overline{s} - E(\overline{s})) \overset{d}{\rightarrow} N(0, \sigma_2^2), \quad \text{where } \sigma_2^2 = \lim_{t \to \infty} \frac{1}{T} \sum_{t=1}^{T} \text{var}\left(\frac{(r_t - r_{f,t})}{\sigma_{r-r_f}} - \frac{(r_{b,t} - r_{f,t})}{\sigma_{r_b-r_f}}\right)
\]

\[
S_3 = T^{1/2}(\text{TR} - E(\text{TR})) \overset{d}{\rightarrow} N(0, \sigma_3^2), \quad \text{where } \sigma_3^2 = \lim_{t \to \infty} \frac{1}{T} \sum_{t=1}^{T} \text{var}\left(\frac{(r_t - r_{b,t})}{\sigma_{r-r_f}}\right)
\]
According to theorem 3.1, the bootstrap methodology provides consistent approximations to the distributions of $S_1$, $S_2$ and $S_3$. Using theorem 3.1 the following bootstrap realizations have to be calculated for the $K$ bootstrap replications:

$$S_{1,k}^* = T^{1/2}(\overline{s}_{k}^* - \overline{s})$$, where $k = 1, ..., K$

$$S_{2,k}^* = T^{1/2}(\overline{st}_{k}^* - \overline{st})$$, where $k = 1, ..., K$

$$S_{3,k}^* = T^{1/2}(\overline{tR}_{k}^* - \overline{tR})$$, where $k = 1, ..., K$

The statistics tested under the null are given by the central limit theorem. These statistics require the true value of the mean under the null. This imposes problems, because negative true values also confirm the null. These values are not known. To solve this problem the ‘Least favorable to the alternative’ configuration is used. This configuration takes the element that confirms the null hypotheses and that is least favorable to the alternative and uses this with the null hypotheses. This element is zero. The null hypotheses for the three measures become $H_0: E(\text{er}) = 0$, $H_0: E(\text{sr}) = 0$ and $H_0: E(\text{IR}) = 0$. Now the test statistics that are tested under the null become $T^{1/2}\overline{st}$, $T^{1/2}\overline{st}$ and $T^{1/2}\overline{tR}$.

### 3.3 Test that correct for data snooping

The stationary bootstrap is a good alternative to other methods in testing financial return series. It is valid under very mild regularity conditions, what is a very convenient property. Also in finite datasets the bootstrap often is more accurate than first-order asymptotic approximations. The issue the method does not address is the data-snooping bias. In the literature this problem has been addressed. Two methods, the Reality Check (White, 2000) and the Superior Predictive Ability (Hansen, 2005), have been developed to correct for data snooping bias.

**The Reality Check**

The Reality Check (RC) procedure was developed by White (2000). It is developed to reduce the data-snooping bias of testing the quality of forecasting models and performance series. The method compares $N$ models with a benchmark model and tries to determine if the best model is really better than the benchmark model.

The Reality Check uses the whole multivariate system of performance series. This in opposite of the stationary bootstrap tests that evaluates products in isolation. For this purpose the performance measures of the individual
managers are grouped in a vector. The null hypotheses of the RC method states that the best manager has no skill over the benchmark:

\[ H_0: \max_{n=1,\ldots,N} (E(f)) \leq 0, \text{ } f \text{ being the vector of } N \text{ performance measures} \]

The performance measures examined are the mean excess return, the Sharpe criterion and the Information Ratio. This method uses the stationary bootstrap method in approximating p-value’s, thus it is required that the performance measures apply to the central limit theory. Assumption 3.2 can be generalized to the vectors of performance measures, so the central limit theorem assures:

\[ T^{1/2}(\overline{\alpha} - E(\alpha)) \overset{d}{\rightarrow} N(0, \Omega_1) \]
\[ T^{1/2}(\overline{\pi} - E(\pi)) \overset{d}{\rightarrow} N(0, \Omega_2) \]
\[ T^{1/2}(\overline{\mu} - E(\mu)) \overset{d}{\rightarrow} N(0, \Omega_2) \]

Note that in this context the symbols stand for the vector equivalent of what they stood for in the former parts of this chapter. The asymptotic variances are replaced by asymptotic \( N \times N \) covariance matrices.

For the RC method White developed a test statistic based on the following assumption (White, 2000):

**Assumption 3.2:** Suppose that \( T^{1/2}(\overline{f} - E(\overline{f})) \overset{d}{\rightarrow} N(0, \Omega) \) for \( \Omega \) positive semi-definite. A) If \( E(f_k) > 0 \) for some \( 1 \leq k \leq N \), then for any \( 0 \leq c < E(f_k) \), \( P[f_k > c] \rightarrow 1 \) as \( T \rightarrow \infty \). B) If \( l > 1 \) and \( f_i > f_k \), for all \( k=2,\ldots,N \), then \( P[f_i > f_k \text{ for all } k=2,\ldots,N] \rightarrow 1 \) as \( T \rightarrow \infty \).

This assumption states that it is asymptotically valid to select the model that shows the highest value of the sample performance vector \( \overline{f} \). Asymptotically this will identify the best model when there is one. Based on this result White proposed the following test statistic to identify the best model:

\[ Q_{rc} = \max_{n=1,\ldots,N} \left( T^{1/2}(\overline{f}_n - E(f)) \right) \]

Also here the ‘Least favorable to the alternative’ configuration is used, for the same reasons as with the ordinary stationary bootstrap method. This implies using the least favorable element to the alternative under the null, which is \( E(f) = 0 \). With this configuration the test statistic under the null hypotheses, becomes:

\[ Q_{rc} = \max_{n=1,\ldots,N} (T^{1/2}(\overline{f}_n)) \]
Under this configuration the null becomes:

\[ H_0: \max_{n=1,...,N} E(f) = 0 \]

To test the statistic \( Q_{rc} \), a distribution has to be consistently estimated. With the following assumption White indicated which distribution should be looked for.

Assumption 3.3. Suppose that \( T \frac{1}{2}(\hat{f} - E(f)) \overset{d}{\to} N(0,\Omega) \) for \( \Omega \) positive semi definite. Then as \( T \to \infty \):

\[
\max_{n=1,...,N} \left( T \frac{1}{2}(\hat{f}_n - E(f_n)) \right) \Rightarrow V_N \equiv \max_{n=1,...,N} \{ Z_n \}
\]

and

\[
\min_{n=1,...,N} \left( T \frac{1}{2}(\hat{f}_n - E(f_n)) \right) \Rightarrow W_N \equiv \min_{n=1,...,N} \{ Z_n \}
\]

Where \( Z \) is a \( Nx1 \) vector with components \( Z_n, n = 1,...,N \), distributed as \( N(0,\Omega) \).

The assumption shows that the asymptotic distribution of \( Q_{rc} \) is the distribution of extremes of the asymptotic distribution \( N(0,\Omega) \). Under assumption that performance measure \( \hat{f} \) satisfies the central limit theorem, the distribution of the statistic \( T \frac{1}{2}(\hat{f} - E(f)) \) is consistently approximated by the stationary bootstrap. The distribution of the set of the maximum bootstrapped performance measures from every replication now consistently approximates the distribution of these normal extremes.

White proposed to apply the stationary bootstrap method to the original dataset and to repeatedly draw the following realizations of the \( K \) bootstrap samples:

\[ T \frac{1}{2}(f_{n,k}^* - \hat{f}_n) \], where \( k = 1,...,K \) and \( n = 1,...,N \)

Now the RC statistic can be evaluated by comparing it with the quantiles of:

\[ Q_{rc,k}^* = \max_{n=1,...,N} \{ T \frac{1}{2}(f_{n,k}^* - \hat{f}_n) \}, \quad k = 1,...,K \]

The RC tests if the best model is better than the benchmark. The test is seeking to control the simultaneous rate of error under the null hypothesis. It has to be noted that the scope of this thesis is about distinguishing the managers that have skill from the unskilled managers. The test has to control for the average rate of error and not for the simultaneous rate of error under the null hypotheses. To adapt the test for the average rate of error, three alternative specifications of the method are proposed. The approximations of the distributions of these statistics follow the same line of arguments White used to show his statistic is consistent, so the bootstrap approximations are also consistent.
The first alternative proposed is the stepwise summation and testing of the performance measures. This test will be referred to as the summation test. First the method of White is followed to find if the largest statistic really is significant. If the null hypotheses is not rejected the conclusion of the test is that there is no skill present in the tested data. If null is rejected a new null hypotheses is adopted. The new null hypotheses states that the combination of the two managers with the highest performance measures has no skill over the benchmark, \( H_0: E(\sum_{j=1}^{J} \bar{f}_j) = 0 \). \( J \) is the number of managers in the test with the highest performance measures \( \bar{f}_j \), in this case two managers. The new statistic under investigation:

\[
Q_{\text{sum},j} = \sum_{j=1}^{J} T^2(\bar{f}_{j[N+1-j]}), \text{ where } j = 1, \ldots, J
\]

The square brackets around the subscript of performance measure \( \bar{f} \) indicate that it is about the order statistics of the vector of performance measures with length \( N \). The stationary bootstrap method can be repeated to estimate the distribution of \( Q_{\text{sum},j} \). From every bootstrap simulation the summation of the \( J \) highest performance measures is calculated:

\[
Q_{\text{sum},j,k} = \sum_{j=1}^{J} T^2(\bar{f}_{j[N+1-j],k}) - \sum_{j=1}^{J} T^2(\bar{f}_{j[N+1-j]}), \text{ where } j = 1, \ldots, J \text{ and } k = 1, \ldots, K
\]

The resulting series give an approximation of the distribution of this statistic. If the null hypotheses is rejected, adopt a new null hypotheses where the \( J \) is raised with one. The iterative testing goes one until the moment a null is not rejected, now the null hypotheses of no skill for the combination of the \( J-1 \) managers is rejected.

The second alternative is to stepwise check managers against their own relevant approximated order distribution. This test will be referred to as the order test. The first step is to follow the method of White to find if the largest found statistic really is significant. If the null is not rejected, the conclusion is that there is no skill present in the dataset. If the null hypotheses that the ‘best’ performing manager has no skill is rejected, the manager with the second highest performance measure is tested. The new null hypotheses is that the manager with the second highest performance measure has no skill. The performance measure is now tested not by comparing it to the quantiles of the maximum bootstrap realizations of the statistics, but to the second highest realizations of the bootstrap statistics. If again this manager is found skillful, repeat the procedure with the third highest performance measure. In general the test statistic for the \( J^{th} \) highest found manager looks as follows:

\[
Q_{\text{order},j} = T^2(\bar{f}_{j[N+1-j]})
\]

The square brackets around the subscript of performance measure \( \bar{f} \) indicate that it is about the order statistics of the vector of performance measures
with length $N$. The stationary bootstrap estimates the relevant distribution with the following realizations:

$$Q_{\text{order},j,b}^* = T^j(f_{[N+1-j],k} - f_{[N+1-j]})$$

where $j = 1, \ldots, J$ and $k = 1, \ldots, K$.

The iterative testing goes own until the moment a null is not rejected anymore, the null hypotheses of no skill for the $J-1$ managers is rejected.

The third and last alternative will be referred to as the removal test. This method performs the reality check on the whole data set. When for the ‘best’ manager the null of no skill is rejected, this manager is removed from the dataset. The reality check is performed on the new data set. Again if for the ‘best’ manager in this data set the null of no skill is rejected, remove the manager from the data set. Keep repeating this step until a the null of no skill is not rejected. The conclusion is that the null hypotheses of no skill can be rejected for the managers that where tested before the last tested manager.

**Test for Superior Predictive Ability**

The Superior Predictive Ability method (SPA) is developed by Hansen (2005) and is an improvement of the Reality Check method. With the SPA method Hansen addressed and solved three problems. first the power of the Reality Check can be driven to zero when adding irrelevant and poor alternatives. The following proposition, made by Hansen (2005) elaborates on this.

**Assumption 3.4.** Let $m_0 \leq N$ be the number of models with $E(f_k) = 0$, define $\Sigma$ to be the $N_0 \times N_0$ submatrix of $\Omega$ that contains the $(i,j)$th element of $\Omega$ if $\mu_i = \mu_j = 0$, and let $\zeta_\Sigma$ denote the distribution of $Z_{\max} = \max_{j=1,\ldots,m_0} Z_j^0$, where $Z^0 = (Z^0_1, \ldots, Z^0_{m_0}) \sim N_{m_0}(0, \Sigma)$. Then $Q_{\text{rc}} \xrightarrow{d} \zeta_\Sigma$ if $\max_{i=1,\ldots,N} \mu_i = 0$, whereas $Q_{\text{rc}} \xrightarrow{p} -\infty$ if $\mu_i < 0$ for all $i = 1, \ldots, N$. Under the alternative where $\mu_k > 0$ for some $k$, it holds that $Q_{\text{rc}} \xrightarrow{p} -\infty$.

This proposition shows that only the binding constraints, $\mu_k = 0$, matter for the asymptotic distribution. A consequence of the proposition is that by adding irrelevant alternative models, so $N$ is increased but $m_0$ does not, the power of the Reality Check statistic is driven down. In other words by adding more bad alternatives, the chance grows that in the bootstrap simulation extreme values come up. Due to these extreme values, the RC method will lose power to distinguish a significant value of the RC statistic.

A second problem of the RC method addressed by Hansen is that the statistics of the different series, in the real data set as well as in the bootstrapped datasets, are measured in different units of standard deviations. This can lead to a situation where an unskillful manager, that
shows very a high performance measures due to high variance, gets selected by the RC statistic over a skillful manager that has a small variance.

The third problem concerns the ‘Least Favorable to the Alternative’ configuration. This configuration can invoke serious power problems, because negative values also confirm the null hypotheses of no skill. This problem is illustrated in figure 3.1. By enforcing the null with the element least favorable of the alternative, point 0 becomes the center of the confidence region. The critical values of the Reality Check are indicated by the dotted line that starts in $C_{rc}$. Although the true value of the performance measures is $\bar{d}$, which indicates that manager 1 indeed has skill, the Reality Check fails to reject the null. If the alternative of point $a$ is imposed under the null, the null will be correctly rejected.

![Figure 3.1](source: Hansen (2005))

Correcting for the afore mentioned problems, Hansen proposed the following alternative statistic and null distribution:

1. $Q_{SPA} = \max [\max_{n=1,...,N} \frac{\sqrt{T} \cdot \hat{f}_n}{\hat{\omega}_n}, 0]$  
   Where $\hat{\omega}_n^2$ is a consistent estimator of $\omega_n^2 = var(\sqrt{T} \cdot \hat{f}_n)$.  

2. Invoke a null distribution, multivariate of order $N$, that is based on $N(\hat{\mu}^c, \hat{\Omega})$, where $\hat{\mu}^c$ is the estimator for $\mu$, the real mean of the statistic $\sqrt{T} \cdot \hat{f}_n$:  
   $$\hat{\mu}^c = \hat{f}_n \cdot \left(\frac{\sqrt{T} \cdot \hat{f}_n}{\hat{\omega}_n} - \sqrt{2 \log \log T}\right)$$

Let $\hat{\omega}_n^2 = K^{-1} \sum_{k=1}^{K} (\hat{f}_{n,k}^* - \hat{f}_n)^2$, calculated from the $K$ bootstrap realizations of the performance measure. Due to the fact that the bootstrap simulations consider the $\hat{f}_n$ as the population value, the law of large numbers ensures that this estimator is consistent for the bootstrap population value of the
variance, which in turn is consistent for the true variance $\omega_n^2$ (Hansen, 2005).

Because of the new distribution under the null, the population mean under the null is not zero anymore for some of the performance series. Therefore the following transformation on the bootstrapped performance measures must be followed to make sure that the bootstrap estimations of the distribution of the performance measures are still consistent:

$$\tilde{Z}_{n,k}^* = \tilde{f}_{n,k}^* - g(\tilde{f}_n^*)$$

For $n = 1, \ldots, N$, $k = 1, \ldots, K$ and $g(\tilde{f}_n^*) = \tilde{f}_n^* \left( \sqrt{\frac{T_{n,k}}{\hat{\sigma}_n}} - \sqrt{2\log\log T} \right)\}

The estimate of the distribution of the transformed performance measure is again consistent under the null. The bootstrapped SPA values are:

$$Q_{spa,b}^* = \max\{0, \max_{n=1,\ldots,N} \frac{\sqrt{T}Z_{n,k}}{\hat{\sigma}_n} \}$$

A valid $p$-value can be found by comparing the SPA statistic to the quantiles of $SPA_b^*$, for $k = 1, \ldots, K$.

Also the SPA method of Hansen described tests only if the best model is better than the benchmark. To modify this to control for the average rate of error the proposed three alternatives discussed under the Reality Check will also follow the Hansen method.
Chapter 4. The simulation procedure and simulation results

To make inference on the reliability of the different testing methods, discussed in chapter 3, the methods are simulated. In the simulation a controlled environment of gross return series is generated, so it is exactly known which simulated product has skill and which product does not. This chapter discusses the model used in the simulation that imitates the real life data generating process (DGP) of the return series. It is explained how the returns are generated and how the characteristics of the data, discussed in chapter 2, are mimicked. The last part of this chapter will discuss the results of the simulation. The tests will be compared on their capabilities to distinguish skilled managers and on the number of falsely rejected null hypotheses of no skill.

4.1 The model for the return generating process

The model

A common model specification for the DGP of financial return series is the Capital Asset Pricing Model framework (Grinold et al, 2000). This framework presents a good starting point to specify a model for the DGP. According CAPM:

\[ E(r_p) = r_f + \beta_p (r_m - r_f) \]

The relation states that the expected return on portfolio \( p \), \( r_p \), is the risk free rate, \( r_f \), plus the sensitivity of the portfolio to the expected market risk premium, represented by \( \beta_p \). The market risk premium is the difference between the market return and the risk free rate. This relation can be modified according the following specification:

\[ r_{p,t} = r_{f,t} + \alpha_p + \beta_p (r_{m,t} - r_{f,t}) + \varepsilon_{p,t} \]

This equation represents the gross return of portfolio \( p \), in period \( t \) as indicated by the subscripts. Two new terms have been added in this representation. There is \( \alpha_p \), which represents the structural return component that is not related to the market risk premium, and the residual term \( \varepsilon_{p,t} \), which is just the return component due to volatility in the market. The residuals are often assumed to be white noise.

For the model of the GDP, equation (1) will be changed on some points. First the benchmark must be represented in the model. The theoretical market portfolio that CAPM indicates consists of all assets traded globally. The investible universe of managers is just a subset of this portfolio, represented by the benchmark. The theoretical market portfolio returns, \( r_m \), will be replaced by the benchmark returns, \( r_b \), in this thesis that is the S&P 500 portfolio. A second adjustment is that the alphas are
not taken to be fixed. The alpha will vary per month according a normal distribution. The choice to let the alpha vary per month is made to bring in some extra variability in the model. This follows the idea that besides volatility of the portfolio, also the investment process will not be capable of adding the same amount of uncorrelated return to the performance every month. There are months when the process is more favored and months where the process is less favored, due to always changing market circumstances. Lastly, the model has to take into account the issues of heteroskedasticity and serial correlation. This issues relates to the residual term $\varepsilon_{p,t}$. To let the model specification be in line with these data characteristics, the residuals will not be distributed as white noise. The residuals have be distributed with mean zero and a covariance matrix that has nonzero covariances and differing variances, $\Omega$. The resulting specification:

$$r_{p,t} = f_t + \alpha_{p,t} + \beta_p (r_{b,t} - f_t) + \varepsilon_{p,t}$$  \hspace{1cm} (2)

Where $\varepsilon_p \sim D(0,\Omega_p)$, $\varepsilon_p$ is a $T \times 1$ vector of residuals, $\alpha_{p,t} \sim N(\alpha_p, \sigma_{\alpha_p}^2)$.

The characteristics of correlation among the different portfolios of the managers and cross-sectional heteroskedasticity between the portfolios have to be taken into account when the multivariate system is simulated. It is important that the simulated multivariate system of performance series incorporates these characteristics. This is done by generating the residuals in such a way that these characteristics are accounted for. How this is done, is described later in this chapter.

**Skill and the model**

To measure the reliability of the different tests in the simulation, the number of type 1 errors and correct rejections of the null hypotheses must be counted. To be able to count them, the skill of every simulated manager series must be known. This skill must be derived from the model for the dgp. This can be done according the null hypotheses for the different performance measures.

When testing Jensen’s alpha, the null hypotheses of no skill is $H_0: E(\alpha) \leq 0$. Jensen’s alpha, $\hat{\alpha}$, represents the estimate for the average returns uncorrelated with the benchmark. In model (2) this is represented by $\alpha_p$. So when testing by Jensen’s alpha, skill is equivalent to a positive value of $\alpha_p$.

For the mean excess return and the Information ratio, the null hypotheses associated with no skill are $H_0: E(\text{er}) \leq 0$ and $H_0: E(\text{IR}) \leq 0$. The expectation of the mean excess return can be rewritten as follows:
So skill is equivalent to a positive value of equation (3). This can also apply for the Information ratio to count the different type 1 errors and correct rejections of the null.

The null of no skill for the Sharpe Criterion is \( H_0: E(sr) \leq 0 \). When testing the Sharpe criterion, skill is equivalent to a significant higher return per unit of risk. The null can be rewritten:

\[
E(sr) = E\left( \frac{(r_p - r_f) - (r_b - r_f)}{\sigma_{r_p-r_f}} \right)
= E\left( \frac{r_f + \alpha_p + \beta_p (r_b - r_f) + \varepsilon_p - r_f}{\sigma_{r_p-r_f}} \right) - E\left( \frac{(r_b - r_f)}{\sigma_{r_b-r_f}} \right)
= \frac{\alpha_p + \beta_p \varepsilon_p}{\sqrt{\text{Var}(\alpha_p) + \beta_p^2 \text{Var}(r_b - r_f) + \text{Var}(\varepsilon_p)}} - \frac{\varepsilon_b}{\sigma_{r_b-r_f}}
\]

(4)

So skill is equivalent to a positive value of equation (4) for the Sharpe criterion.

4.2 The data generation

In the simulations, systems of 100 product performance series will be generated with a sample length of \( T \). To make reliable conclusions on the behavior of the different tests, they must be based on simulations that are in line with the real world data generating process. For the model representation of the dgp in model (2) this means using realistic values of alphas, betas and using appropriate approximations of the covariance matrices of the residual series. In the context of simulating a system of managers, the correlation between the performance series also must be taken into account.

The Alphas, betas and residuals

A realistic set of monthly alphas and betas is estimated from real data. The dataset is collected from the manager database Evestment Alliance. In total 108 performance series of different products are gathered. The series have a length of 144 observations. The products chosen apply to the general characteristics described in chapter 2. The data are regressed according the following model:
The 108 estimates for the alphas and betas, both the significant and insignificant ones, are gathered in an alpha and a beta set, respectively. In Table 4.1 some features of these sets are given. The product performance series that are simulated are assigned a beta and an alpha, randomly drawn with replacement from the beta set and the alpha set. The drawn alphas will function as the mean for the alpha distributions. For series \( n \) this is \( \alpha_n \).

Table 4.1

<table>
<thead>
<tr>
<th></th>
<th>Alpha</th>
<th>Beta</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>0,0214%</td>
<td>0,96</td>
</tr>
<tr>
<td>Minimum</td>
<td>-0,9095%</td>
<td>0,68</td>
</tr>
<tr>
<td>Maximum</td>
<td>0,9559%</td>
<td>1,13</td>
</tr>
<tr>
<td>Mean</td>
<td>-0,0217%</td>
<td>0,94</td>
</tr>
<tr>
<td>Number</td>
<td>108</td>
<td>108</td>
</tr>
</tbody>
</table>

The variance of the alpha distribution of series \( n \), \( \sigma_{\alpha_n}^2 \), will be generated randomly based on a distribution that is also used by Carpenter and Lynch (1999):

\[
\varphi_n \sim N(0.95; 0.25^2)
\]

\[
\sigma_{\alpha_n}^2 = 0.05349 \cdot (1 - \varphi_n)^2 \cdot \frac{1}{12}
\]

So the random alpha’s for manager \( n \) will be drawn from:

\[
a_{n,t} \sim N(\alpha_n, \sigma_{\alpha_n}^2), \text{ where } n = 1, \ldots, 100 \text{ and } t = 1, \ldots, T
\]

To get a good approximation of the system of heteroskedastic, serially and cross-sectionally correlated residuals, the different residual series resulting from regression (5) are stored in a matrix. The residuals in the simulation will be randomly sampled from this matrix, according to the stationary bootstrap method. The resulting matrix of residuals will be randomly assigned to obtain the residuals for the simulation. The reason the residuals of the real data set are used through the stationary bootstrap is to preserve the data characteristics.

Survivorship Bias

As described in chapter 2, survivorship bias is often present in data bases of product performance series. Because this phenomenon can seriously trouble the multivariate statistical tests and analysis, the simulation will take this into account. To make inference on the effect of survivorship bias, three samples will be generated and tested. First a normal sample of random return series will be generated and tested. After this a second sample will be generated, a survivorship biased version of
the first sample. To bias the first sample, a single year criterion is used, based on Carhart’s (1997) sample. This sample is one of the most comprehensive mutual fund data sets in the literature and without survivorship bias (Carpenter and Lynch, 1999). From this sample on average 3.6% of the funds shut down every year. The single year criterion biases the sample by removing the 3.6% of the worst performing funds every year. A third sample will be created by randomly removing return series of the first sample, until the number of series left is the same as the survivorship biased sample. This last sample is to control for differences in test behavior simply due to differences in size between the first sample and second sample and not to the survivorship bias.

4.3 Size of the samples and the testing parameters

To make inference on the effect of the length of the available data, two simulations are run. The first simulation generates 144 observations of data, the second simulation generates 72 observations. As a proxy for the risk free rate the three month t-bill rate is used. The benchmark is chosen to be the S&P 500 index. The data used in the simulation of 144 observations ranges from July 1997 to June 2009. This is chosen because it involves two market cycles, as illustrated figure 4.1. The simulation of 72 months uses residual data, benchmark returns and risk free returns that range from July 2003 to June 2009.

Figure 4.1

To achieve convergence in the simulation, the number of replications has to be sufficiently large. To balance between appropriate convergence in results and a feasible simulation time, the number of replications of the simulations is chosen to be 10,000. For the stationary bootstrap replications in the different testing procedures a number of 500 is chosen, following White (1999). For appropriate blocking length in the stationary bootstrap a paper written on optimal blocking lengths by Hall, Horowitz and Jing (1995) is consulted. The chosen smoothing parameter:

$$\rho_T = \frac{1}{\sqrt{T}} \text{ for } T = 72,144$$
The chosen significance levels and critical values in the individual tests are chosen to be respectively 5% and 1.96, respectively. Though because all the procedures eventually test the same multiple null hypotheses, the tests should also be comparable with each other. This is not the case when the chosen significance level for the individual tests are used, they lead to different numbers of type 1 errors made. To make the results of the different tests better comparable, also a simulation of 144 observations is run with adjusted significance levels. These significance levels are adjusted in such a way that all the tests lead to a 5% data-snooping bias.

4.4 The Results

The results will be discussed per performance measure. The following results are presented per simulated sample size:

- The average fraction of managers with skill found among the total number of skilled managers, defined as power.
- The average fraction of type 1 errors made among the total number of unskilled managers, defined as data-snooping bias.

Jensen’s Alpha

In table 4.2 the results are given for Jensen’s Alpha. The results are in line with theory, the performance measure is unreliable when used analyzing series that are heteroskedastic and serially correlated. The results illustrate the danger of false inference when using this measure. The result that show that the sample size goes up the power of the test goes down, is a first indication of the non-validity of the test. Also the high data-snooping bias, around 16 to 17 percent of the unskilled managers are identified as skilled, illustrates why this test is very unreliable.

<table>
<thead>
<tr>
<th></th>
<th>Jensen’s Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>72 Months</td>
<td></td>
</tr>
<tr>
<td>Power</td>
<td>54.63%</td>
</tr>
<tr>
<td>DS Bias</td>
<td>17.41%</td>
</tr>
<tr>
<td>144 Months</td>
<td></td>
</tr>
<tr>
<td>Power</td>
<td>52.51%</td>
</tr>
<tr>
<td>DS Bias</td>
<td>16.17%</td>
</tr>
</tbody>
</table>

Table 4.2. The results for Jensen’s Alpha

Mean excess return

The results of the different testing procedures of the mean excess return are shown in Table 4.3. Again the results of the simple t-test show its unreliability. That the power of the test goes down when the sample size increases indicates serious problems. Though the percentages with the
simple t-test in itself seem reasonable, it is a invalid and not to be trusted test procedure on this kind of data.

<table>
<thead>
<tr>
<th>Table 4.3. Testing Results of the Mean Excess Return</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Excess Return</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>72 Months</td>
</tr>
<tr>
<td>Power</td>
</tr>
<tr>
<td>Simple T-test</td>
</tr>
<tr>
<td>Stationary Bootstrap Test</td>
</tr>
<tr>
<td>White RC Statistics</td>
</tr>
<tr>
<td>Summation test</td>
</tr>
<tr>
<td>Order test</td>
</tr>
<tr>
<td>Removal test</td>
</tr>
<tr>
<td>Hansen SPA Statistics</td>
</tr>
<tr>
<td>Summation test</td>
</tr>
<tr>
<td>Order test</td>
</tr>
<tr>
<td>Removal test</td>
</tr>
</tbody>
</table>

The stationary bootstrap test does not show strange behavior. The power is reasonable and it seems that the test is quite conservative according the data snooping bias.

The summation test that follows the RC method does show strange behavior. When the sample size increases the data-snooping bias goes up. Another result of the summation method is a very high value of data-snooping bias and power, for both the RC method and the SPA method. These results are due to the fact that the test looks at combinations of managers. The null hypotheses tested are that combinations of managers are unskilled, not that the individual managers in the combinations are unskilled. This can lead to the situation that a combination of managers is distinguished as skilled, the combination shows significant outperformance, while in the combination there are skilled and unskilled managers present. It is false to conclude that all managers in the combination are skilled. In the context of distinguishing skilled managers, and not skilled combinations of managers, this leads to a very high acceptance of unskilled managers as skilled. Due to this property, the test is not suited to make inference on individual managers. This is backed by the results from the simulation.

The order test shows the behavior that the data-snooping bias goes up when the sample size increases. Another observation is the relative high values of the data-snooping bias when the order test follows the SPA method. This opposite of the much lower data-snooping bias when the RC method is followed. A last result is that the test does not reduce the data-snooping bias, it is higher than the data-snooping bias of the ordinary stationary bootstrap test. The method tests the $j^{th}$ largest performance measure in the vector of performance measures $\tilde{f}$, of length $N$. The statistic is evaluated against its order distribution. According chapter 3 the test statistic is:

$$Q_{order,j} = \left( T^2 \left( f_{[N+1-j]} \right) \right)$$

It is evaluated against the empirical distribution function derived with the stationary bootstrap method. It generates the following realizations:
The results imply that the approximated distribution is too much blurred by bootstrap realizations of the statistic of series other than the series that belongs to the performance measure that is tested. Although the bootstrap method indicates the approximation is consistent, the results imply that the test is not reliable in finite samples of 144 observations or less.

The method that subsequently tests and removes best found managers for the data set show the smallest rate of type 1 errors, both for the RC and SPA methods. The Hansen method clearly shows a significantly higher rate of correctly rejected null hypotheses, while the rate of type errors is not significantly higher. This result is in line with Hansen's findings on the RC method. Clearly the removal test succeeds the best in reducing the data snooping bias. Unfortunately the test has low power.

The results for the adjusted significance levels make the different tests comparable. In general the tests that follow the SPA method show higher power than the tests following the RC method. This is again in line with Hansen’s findings. Both empirical evidence and Hansen’s results imply that the SPA method should be preferred above the RC method. When comparing the different test that follow the Hansen method and the stationary bootstrap method, it seems that the order test has the least power.

When looking at the SPA removal test, the results show that the correction for data-snooping bias that the test should make, is ‘eliminated’ by raising the data-snooping bias to 5%. Compared to the ordinary stationary bootstrap method the results do not significantly differ. This suggests the data-snooping bias can also be simply brought down by choosing a very low significance level for the ordinary stationary bootstrap method. To deduce if this is the case simulations are run where the data-snooping bias of both tests are brought down simultaneously in steps of 1%. Conclusions from this exercise are that the power of the stationary bootstrap goes down faster than with the SPA removal method. From this it can be concluded that the removal test that follows the SPA method really corrects for data-snooping bias when compared to the stationary bootstrap method.

Information Ratio

The results for the simulations of the Information Ratio are given in table 4.4. Note that the Information Ratio tests for the same skill as the mean excess return. The results show that the outcomes of the simulations behave more or less the same as with the mean excess return. The same problems and issues apply. The main difference lays in the results of the RC method. They show that the tests that follow the RC method have significantly higher powers than they do when the mean excess returns is tested.
The results imply that dividing the mean excess return by its standard error, as the Information ratio does, has a comparable effect as the studentizing of the performance measures as the SPA method does. Though the results of the adjust significance simulation show that this effect is not as large as the effect of the studentizing, as the Hansen method still performs better than the White method.

Table 4.4. Testing results of the Information Ratio

<table>
<thead>
<tr>
<th>Information Ratio</th>
<th>72 Months</th>
<th>144 Months</th>
<th>144 Months Adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Power</td>
<td>Power</td>
<td>Power</td>
</tr>
<tr>
<td>White RC Technique</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summation test</td>
<td>29.2%</td>
<td>32.9%</td>
<td>56.0%</td>
</tr>
<tr>
<td>Order test</td>
<td>31.7%</td>
<td>38.4%</td>
<td>39.5%</td>
</tr>
<tr>
<td>Removal test</td>
<td>6.5%</td>
<td>8.8%</td>
<td>50.1%</td>
</tr>
<tr>
<td>Hansen SPA Technique</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summation test</td>
<td>70.8%</td>
<td>77.1%</td>
<td>51.8%</td>
</tr>
<tr>
<td>Order test</td>
<td>44.2%</td>
<td>50.6%</td>
<td>50.9%</td>
</tr>
<tr>
<td>Removal test</td>
<td>8.3%</td>
<td>10.2%</td>
<td>55.5%</td>
</tr>
</tbody>
</table>

Table 4.5. Testing results of the Sharpe Ratio

Sharpe Criterion

The results for the Sharpe criterion are given in table 4.5. Also here results indicate that the simple t-test is not very reliable. 4% of the non-skillful managers are indicated as skillful managers against only around 15% of the skillful managers identified as such. This implies that there is a large risk of accepting unskilled managers as skilled.

<table>
<thead>
<tr>
<th>Sharpe Criterion</th>
<th>72 Months</th>
<th>144 Months</th>
<th>144 Months Adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Power</td>
<td>Power</td>
<td>Power</td>
</tr>
<tr>
<td>White RC Technique</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summation test</td>
<td>8.5%</td>
<td>15.3%</td>
<td>33.8%</td>
</tr>
<tr>
<td>Order test</td>
<td>21.4%</td>
<td>46.1%</td>
<td>5.9%</td>
</tr>
<tr>
<td>Removal test</td>
<td>1.2%</td>
<td>84.3%</td>
<td>9.5%</td>
</tr>
<tr>
<td>Hansen SPA Technique</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summation test</td>
<td>11.9%</td>
<td>79.2%</td>
<td>29.6%</td>
</tr>
<tr>
<td>Order test</td>
<td>4.6%</td>
<td>67.9%</td>
<td>27.8%</td>
</tr>
<tr>
<td>Removal test</td>
<td>1.8%</td>
<td>15.2%</td>
<td>33.2%</td>
</tr>
</tbody>
</table>

When looking at the other procedures the results also indicate problems. The high data-snooping bias in the large sample are worrisome. The fact that the data-snooping bias is much larger in the 144 months simulation than in the 72 months simulation is an indication that the tests experience serious problems. Also the powers of the large sample and small sample seem to lay disproportionately far apart. The test methods do not seem to be asymptotically valid. The problems must be looked for in the stationarity requirements that must apply to make sure the stationary bootstrap method is consistent. Further investigation, using the Augmented Dickey-Fuller test to look for unit roots in the data, pointed out that the risk free rate is not stationary. This issue troubles the approximations made by the
stationary bootstrap method. The results imply that this has a severe influence on the reliability of the testing methods.

Despite this conclusion, the Sharpe Criterion is an important statistic in finding skill in achieving a higher return per unit of risk taken than the benchmark. For this reason it still seems useful to identify the best performing test method.

The ordinary bootstrap method does not perform very bad in the small sample simulations, though when the sample size is 144 months the data-snooping bias is very high. This high sensitivity to the sample size and the high data-snooping bias make that the stationary bootstrap method is very unreliable. The summation test and the order test seem to suffer from the same issues as with the other statistics and therefore are also unreliable. The method left is the removal test. This again performs relative well when comparing the power/data-snooping bias proportion to those of the other tests. The sensitivity to size differences in the sample also are less severe in terms of data-snooping bias. The results for the removal test again show that the Hansen method has superior power over the White approach.

**Survivorship Bias**

Survivorship bias does not influence testing methods that individually tests managers. Methods that make use of the whole dataset are influenced by Survivorship bias. This is the case with the tests that follow the RC method and the SPA methods. In table 4.6 the results of the survivorship biased date are given. It must be noted that the effects seen are not due to size differences of the tested data sets, the generated control sample showed no significantly different results with the normal data set.
Table 4.6. Results of the Survivorship biased data

The results in table 4.6 show higher power and higher data-snooping bias for both the RC method and the SPA method than the results of the non-survivorship biased samples. The tests that follow the RC method seem to be most severely affected by the survivorship bias.

For the SPA method this is mainly a result of the fact that the poorest performing managers are removed. The poorest performing managers do not heavily influence the number of managers for which the null hypothesis of no skill is rejected, correctly or falsely. It does influence the total number of skillful managers and the total number of non-skillful managers in a dataset. So the percentages of power and data-snooping bias go up. The real capability of distinguishing skilled managers and the real risk of data-snooping bias therefore are not really affected.

Besides this effect that influences the results of the tests that follow the RC method, these tests are also really influenced by survivorship bias. This effect is reflected by the results, the power and data-snooping bias increase by far larger proportions than with the SPA method. This is in line with the fact that the power of the White procedure is brought down by the inclusion of poor models (Hansen, 2005). The survivorship bias does the opposite, it excludes poor models from the data set, whereby the power goes up. Though it still is not as large as with the Hansen test. The results imply that also in some cases the data-snooping bias increases.

4.4 Conclusions from the simulation

The results indicate that there is no reason to prefer the Information Ratio over the mean excess return when using the tests examined. The Sharpe Criterion is shown to be an unreliable measure to test with the examined methods.

In terms of testing methods the simulations indicate that the ordinary stationary bootstrap does not do a bad job in testing the mean excess return and the information ratio. When simplicity and high power, assumed testing with as significance level of 5%, are of more concern than correction for data-snooping, this method is very useful. When the risk of data-snooping bias is of main concern, the removal test that follows the SPA method is best chosen. The RC method is not preferred, as it has lower power and is also affected by the survivorship bias. When testing the Sharpe Criterion the only useful test is the removal test that follows the SPA method. The other tests showed very unreliable.
Chapter 5. Results of the empirical study

To make inference on the question if there is evidence for skill, the performance measures and test procedures are used in an empirical study. This chapter discusses the results of this study.

The dataset that is tested consists of 108 investment products with a track record of 144 months, benchmarked against the S&P 500. The products are chosen according to the characteristics described in chapter 2. The tests methods used are the ordinary stationary bootstrap test and the removal test that follows the SPA method. The tests are chosen according to the results from the simulation in chapter 4.

The dataset is tested in three ways. First the whole dataset of 144 observations is tested. The whole dataset covers two market cycles. To see if the skill of managers is depended on up-markets or down-markets, the complete set is divided in two sub sets. The up and down market periods can be seen in figure 4.1. The up-market periods are taken from July 1997 to August 2000 and from September 2002 to October 2007. The up-market set consists of 99 observations. The down-market periods are taken from September 2000 to August 2002 and from October 2007 to June 2009, the set consists of 45 observations.

5.1 Results

Results for two market cycles

<table>
<thead>
<tr>
<th></th>
<th>Stationary Bootstrap</th>
<th>Hansen removal test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Excess Return</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>Information Ratio</td>
<td>32</td>
<td>2</td>
</tr>
<tr>
<td>Sharpe Criterion</td>
<td>33</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 5.1. Results of the analysis of two market cycles

In terms of skill in outperforming the benchmark with higher returns, the stationary bootstrap test distinguishes 30 managers with the mean excess return and 32 managers with the Information Ratio. The information ratio identifies 2 more managers as skillful. This is also seen when the performance measures are tested with the Hansen removal tests. This test distinguishes no managers with the mean excess return and 2 managers with the information ratio.

According the results there is evidence in favor of skill in outperforming the benchmark. The behavior of the tests in the simulation indicate it is very reasonable to assume that the two managers, distinguished with the removal test on the information ratio, do have skill. It is also not unlikely that there are more managers with skill present. The removal test has low power with the used 5% significance level and the stationary
bootstrap method showed also is conservative when used testing this kind of return data.

The same goes for skill in achieving a higher return per unit of risk taken, also here the results indicate there is evidence for this kind of skill. The low type 1 error rate of the removal tests indicated by the simulation does makes it reasonable the 6 found managers do have skill. Again the low power of the removal test and the fact that the stationary bootstrap method distinguishes 33 managers, makes it very likely there are more managers with this kind of skill present in the data set.

Results for up-markets

<table>
<thead>
<tr>
<th></th>
<th>Stationary Bootstrap</th>
<th>Hansen removal test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Excess Return</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>Information Ratio</td>
<td>26</td>
<td>0</td>
</tr>
<tr>
<td>Sharpe Criterion</td>
<td>21</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5.2, The number of managers identified as skillful, up-market

The results are different from the tests on two market cycles. No evidence for any sort of skill can be found with the removal test. The results for the stationary bootstrap method on outperforming the benchmark with excess returns does indicate evidence in favor of skill. Though it must be noted that the data set with 99 observations is smaller than the data set of two market cycles, with 144 observations, what makes the evidence weaker. Also the fact that only the stationary bootstrap method distinguishes skill also make the evidence less reliable then with the previous test on the large dataset.

The stationary bootstrap method does also distinguishes managers that have skill in outperforming the benchmark per unit of risk. Though taken the results of the simulations in mind, this evidence is unreliable.

Results for down-markets

<table>
<thead>
<tr>
<th></th>
<th>Stationary Bootstrap</th>
<th>Hansen removal test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Excess Return</td>
<td>54</td>
<td>15</td>
</tr>
<tr>
<td>Information Ratio</td>
<td>39</td>
<td>14</td>
</tr>
<tr>
<td>Sharpe Criterion</td>
<td>51</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 5.3, The number of managers identified as skillful, down-market

The dataset for the down market is only 44 observations long. This is a small data set that makes inference not as reliable as with the previous tests. Though taken this into account, the number of managers distinguished
as skilled is large. Especially the number of rejections of no skill in terms of outperforming the benchmark is so high it is unlikely there is no skill present in the down market. The stationary bootstrap and the removal test results show high numbers of distinguished skillful managers. In terms of outperforming the benchmark, the most convincing evidence is found in favor of skill.

The removal test shows 4 managers with the skill in achieving a higher return per unit of risk taken, the stationary bootstrap tests indicates that there are 51 managers with this skill. So also here there is evidence in favor of skill.

5.2 Conclusions

The results indicate there is evidence for successful active asset management. The tests show there are performance series that show significant skill, both in terms of outperforming the benchmark in pure returns and in achieving higher returns per unit of risk taken. The tests of the different market regimes indicate that there are skillful managers, though they mainly seem to manifest in down markets.
Chapter 6. Summary

In the field of financial asset management there are differences in the performance of asset managers. Especially managers that actively try to beat a benchmark, in general a market index, show big discrepancies in terms of returns. These differences are mainly the result of differences in skill and luck. Skilled managers outperform the benchmark on the long run, while unskilled managers will not be capable to add significant value on top of the index returns. The trouble in identifying skilled managers from unskilled managers comes from the fact that also luck is an important factor in investing. Skilled managers can experience bad luck and show underperformance, while unskilled managers can show outperformance due to luck. Although on the long term this luck should cancel, on the shorter term the element of luck can really trouble identifying skilled from unskilled managers. Due to this it is not possible to simply conclude that managers that have historically outperformed are skilled and managers that have underperformed are unskilled.

In statistics, analysis methods have been developed to distinguish skilled managers. An important category of these methods is based on performance measures on historical return series. An important problem of these methods is that they involve data-snooping bias. Data-snooping bias refers to the number of falsely distinguished managers as skillful. In other words it is about the number of type 1 errors made when testing a whole set of managers, assumed the null hypotheses implies no skill. This thesis tries to identify the reliability in distinguishing skillful managers and how big the data-snooping bias is. The behavior of the methods is analyzed through a simulation. The performance measures examined are the mean excess return, the Sharpe criterion, the information ratio and Jensen’s alpha. The test methods that have been analyzed are standard t-tests, the stationary bootstrap method, the Reality Check method and the Superior Predictive Ability method.

The first set of tests, the standard t-test, is simulated on the mean excess return, on the Sharpe Criterion and on Jensen's alpha. The tests show to be unreliable. A very high data-snooping bias can be observed. Also the tests behave very strange when the sample size increases. The reason lies in the fact that due to the statistical properties of the performance series the standard t-tests are invalid. The second method examined is the stationary bootstrap. This method is theoretically valid as long as the performance series are stationary and do not show to much dependence. The method shows reliable and comparable results when used to test the mean excess return and the Information ratio. The power of distinguishing skilled managers is very acceptable and the test show conservative data-snooping bias. The results imply that the stationary bootstrap is useful to test for skill. The Sharpe ratio tests though, show very unreliable and invalid results. A closer analysis reveals that the inclusion of the risk free rate in the statistic, which is a non stationary series, causes the
problems. The measure can be concluded not being reliable when tested with the stationary bootstrap.

The Reality Check method, developed by White (2000), is a method based on the stationary bootstrap. It is developed to correct for the data-snooping bias. The original method tests if the best model performs better than the benchmark. To make the method suitable to distinguish the skilled managers from the unskilled managers in general, three alternatives are proposed. The alternatives are referred to as the summation method, the order method and the removal method. One alternative, the removal method, showed useful in the results from the simulations. The power of the test is low, but the data-snooping bias is almost eliminated. For testing the Sharpe criterion the test also experienced problems, though not as severe as with the ordinary stationary bootstrap method.

The last alternative that is analyzed is the Superior Predictive Ability method. This method is developed by Hansen (2005) and corrected for some problems the Reality Check experiences. Also for this method the three alternatives are proposed, the same as with the Reality Check and for the same reason. Also here the removal method showed the best result. The power of the removal test when following the SPA method is even significantly higher than with the RC method, while the data-snooping bias is as low. Still, the power is not as high as with the ordinary bootstrap method.

The simulation results show that when testing the mean excess return or the information ratio two tests are the most useful. The ordinary bootstrap method is the best alternative when power of the test is of main concern. When a small data-snooping bias is of more importance, the SPA removal test is the best testing method to use. The Sharpe criterion showed not very reliable results. Though it can be important to test the Sharpe criterion. In this case it is best to use the SPA removal method, as it behaved the least unreliable in terms of data-snooping bias when compared to the other methods.

With the results from the simulation an empirical study is done on a dataset of 108 managers that contains two market cycles. The main conclusion is that over two market cycles there is evidence of skill. Also the up-market and down-market periods are analyzed in isolation. The conclusions from these analysis are that there is evidence for skill in the up and down market. The down-market showed the most evidence, therefore skill of managers seem to be higher in down-markets than in up-markets.

Though the results in the thesis leads to conclusions of which method best used, further research on the subject is needed. Recommendations for future research on the topic are model assumptions other than the one used for the data generation process, the results are dependent on the validity of model chosen. Also using different simulating methods will be a valuable addition to the research done. Another point is the dataset used. The simulation and the empirical study only involve equity managers that benchmark against the
S&P 500. Other markets and different asset classes with different dynamics and risks could lead to different results. Also of interest is the effect of including managers in the analysis that have a inception date later than the starting point. The managers involved in this thesis all had a track record of twelve years or longer. Within this twelve years many new managers started to manage new products. These managers are ignored in this thesis. Also an interesting extension to the analysis can be the effect of track records different than 144 observations and 72 observations. Lastly, analyzing different performance measures is also a valuable addition.
References


Hall, P., Horowitz, J.L., Jing, B. (1995), On Blocking Rules for the Bootstrap with Dependent Data, Biometrika 82, 561-574


Horowitz, J.L. (2000), The Bootstrap, Department of Economics, University of Iowa


