Open your mind. LUT.
Lappeenranta University of Technology

Faculty of Technology
Department of Mathematics and Physics
Laboratory of Applied Mathematics

Monte Carlo Hypothesis Testing with The Sharpe Ratio

Senghor Nkuliza

The topic of this Master's thesis was approved
by the faculty council of the Faculty of Technology on 24th May 2012

The examiners of the thesis were: Prof. PhD. Heikki Haario and Prof. PhD. Eero Pätäri. The thesis was supervised by: Prof. PhD Heikki Haario

Lappeenranta, August 20, 2012

Senghor Nkuliza
Liesharjunkatu 9 A 8
53850, Lappeenranta, Finland
+358465959554
nkulizas@gmail.com

Abstract

Lappeenranta University of technology
Department of Technomathematics

Senghor Nkuliza

Monte Carlo Hypothesis Testing with The Sharpe Ratio
Master's Thesis
2012
71 pages, 42 figures, 7 tables, 2 appendices.
Examiners: Prof. PhD. Heikki Haario and Prof. PhD. Eero Pätäri.
Supervisor: Prof. PhD. Heikki Haario.
Keywords: Monte carlo, Hypothesis testing, Bootstrapping, Sharpe ratio, Portfolio Performance.

The purpose of this master thesis was to perform simulations that involve use of random number while testing hypotheses especially on two samples populations being compared weather by their means, variances or Sharpe ratios. Specifically, we simulated some well known distributions by Matlab and check out the accuracy of an hypothesis testing. Furthermore, we went deeper and check what could happen once the bootstrapping method as described by Effrons is applied on the simulated data. In addition to that, one well known RobustSharpe hypothesis testing stated in the paper of Ledoit and Wolf was applied to measure the statistical significance performance between two investment founds basing on testing weather there is a statistically significant difference between their Sharpe Ratios or not.
We collected many literatures about our topic and perform by Matlab many simulated random numbers as possible to put out our purpose; As results we come out with a good understanding that testing are not always accurate; for instance while testing weather two normal distributed random vectors come from the same normal distribution. The Jacque-Berra test for normality showed that for the normal random vector r1 and r2, only 94,7% and 95,7% respectively are coming from normal distribution in contrast 5,3% and 4,3% failed to shown the truth already known; but when we introduce the bootstrapping methods by Effrons while estimating p-values where the hypothesis decision is based, the accuracy of the test was 100% successful.
From the above results the reports showed that bootstrapping methods while testing or estimating some statistics should always considered because at most cases the outcome are accurate and errors are minimized in the computation. Also the RobustSharpe test which is known to use one of the bootstrapping methods, studentised one, were applied first on different simulated data including distribution of many kind and different shape secondly, on real data, Hedge and Mutual funds. The test performed quite well to agree with the existence of statistical significance difference between their Sharpe ratios as described in the paper of Ledoit and Wolf.

Acknowledgements

I am grateful to the Department of Mathematics and Physics of Lappeenranta University of Technology for the financial support during the entire duration of my studies.

I am also grateful to the supervisor of the thesis, Prof. PhD. Heikki Haario and the examiner Prof. PhD. Eero Pätäri, for proposing this interesting topic, comments and guidance; and PhD. Matylda Jablonska for assistance.

Special thanks also go to all my classmates, friends and family for their enthusiastic social relations which fueled my hope for a brighter future and contributed to the creation of a worth living environment for all my endeavors.

The vote of my deep thanks goes to you all.

Murakoze.

Lappeenranta, August 20, 2012

Senghor Nkuliza

Contents

List of Tables

List of Figures

1 INTRODUCTION

In our every day life we are always obliged to make decisions and choosing among two or many alternative and the right decision on every presented alternative affect positively or negatively on our lives; that is the null hypothesis H_0 and the alternative hypothesis H_1.

According to [23] Monte carlo methods refers to simulations that involves the use of random numbers, nowadays the use of computer especially Matlab in our case has simplified several statistical studies based on the fact that monte carlo simulations or experiments are an easy and faster done [10] [19].

In statistics, a hypothesis is claim or statement about a property of a population and a hypothesis testing is a procedure for testing a claim about a property of a population [13] [1].

A Sharpe ratio is one of the adequate instrument used to measure the performance rank investment strategy of a portfolio by looking at historic return and risk [14] [2] [21] [22] [24].

In every hypothesis testing we should be able to understand: (i) an identification of the null hypothesis and alternative hypothesis from a given claim, and how to express them in symbolic from; (ii) how to calculate the value of the test statistics, given a significance level usually known as α; (iii) how to identify the critical value (s), given a value of the test statistic (iv) how to identify the P-values, given of the test statistic, (v) how to state the conclusion about a claim in simple terms understandable by every one [13] [17] [8].

The objective of our work was:

- To check out the accuracy of the hypotheses tests by the use of simulating some known distribution
- What happen while bootstrapping techniques is involved in [23]
- To understand the sharpe ratio approach as performance measure of an investment based decision

Many investors do not understand how to determine the level of risk their individual portfolios. [22] This work contributes to current financial literature by studying methods that can extend the applicability of the statistical tests based on the asymptotic variance to many such performance comparisons for which the other known statistical methods are either too complicate to implement or can not be reliably employed. The first of these adjustments is made to enable statistical inference on performance difference in cases, when the excess returns are negative for both portfolios being compared. The other adjustment procedure is appropriate in cases when excess returns of portfolios are of different sign. The third adjustment is made in order to reduce biases in test statistics stemming from the violations of normality and I.I.D. assumptions [2]

Our current work is subdivided into five parts the first one is the introductory part which introduce the report, the second part is the mathematical background on monte carlo methods hypothesis testing where the theory is discussed and some example of the simulation results is shown, the third part is made of portfolio performance measurement especial with the Sharpe ratio approach where the robust Sharpe hypothesis testing is discussed the fourth part is some results the application of testing the existence of statistical significance difference between two Sharpe ratio or not when using simulated data as well as the mutual and hedge funds data, the last and fifth part is made of conclusion including some recommendation.

2 MONTE CARLO INFERENCE STATISTICS

2.1 Hypothesis testing

Hypothesis testing is a common method of drawing inferences about a population based on statistical evidence from a sample.

Inferential statistics involves techniques such as estimating population parameters using point estimates, calculating confidence interval estimates for parameters, hypothesis testing, and modeling based on the sample has been observed or using managerial judgement [11].

There are two kind of hypothesis testing, parametric one and non parametric one, The parametric hypothesis testing concern parameters of distributions generally assumed to be normal, some conditions about the distribution must be imposed or known while testing; Non parametric hypothesis does not impose conditions about the distribution of the data variables [8].

Since no assumption are imposed here, the non parametric test can be adequate to small sample of variable data furthermore the non parametric hypothesis can test more different hypothesis than the parametric hypothesis.

However, in [8], proved that the non parametric tests are generally not powerful as the parametric tests due to the use of fewer condition imposed on the distributions. in order to compare the power of a test A and a test B, we can determine the power efficiency measure of test B compared with test A, η_{BA} defined as:

$\eta_{BA} = \frac{\eta_A}{\eta_B}$

Where η_A is the sample size needed by A and η_B is the one needed by B.

Concerning our current work thesis we will focus on the testing while inferencing on two population sample hedge funds and mutual funds.

How can you decided about the choice between two investment funds? a variety of decision-making technics are established in several Finance books and published papers but in each hypothesis testing we will follow the same five steps procedure as follow [19], [1], [3], [8] [23] and [13]:

1. Analyze the problem - identify the hypothesis, the alternative hypotheses of interest, and the potential risks associated with a decision.
2. Choose a test statistics.

3. Compute the test statistics.
4. Determine the frequency distribution of the test statistic under the hypothesis.
5. Make a decision using this distribution as a guide.

2.2 How to carry out an hypothesis testing?

Hypothesis testing is carried out using confidence intervals and test of significance.

In hypothesis testing, our goal is to make a decision about not rejecting or rejecting some statement about the population based on data from a random sample; to understand and use statistical hypothesis testing, one needs knowledge of the sampling distribution of the test statistic

Parametric hypothesis testing using different methods is stated hereunder [1]:

1. Carrying out a hypothesis testing using the test of significance approach [3]:
   - Estimate the model parameters and their standards errors in the usual way.
   - Calculate the test statistic by the formular

     $$teststatistic = \frac{\hat{\beta} - \beta*}{SE(\beta)} \tag{1}$$

     Where $\beta*$ is the value of β under the null hypothesis. The null hypotheisi is $H_0 : \beta = \beta*$ and the alternative hypothesis is $H_1 : \beta \neq \beta*$ (for two-sided test).
   - To compare the estimated test statistics a tabulated distribution is required; in this way t-statistics follows distribution with $T-2$ degree of freedom
   - Choose a significance level α, conventionally is 5% or 1% rarely.
   - Given α a rejection region and non rejection can be determined as shown here under in figure 1, 2 and 3

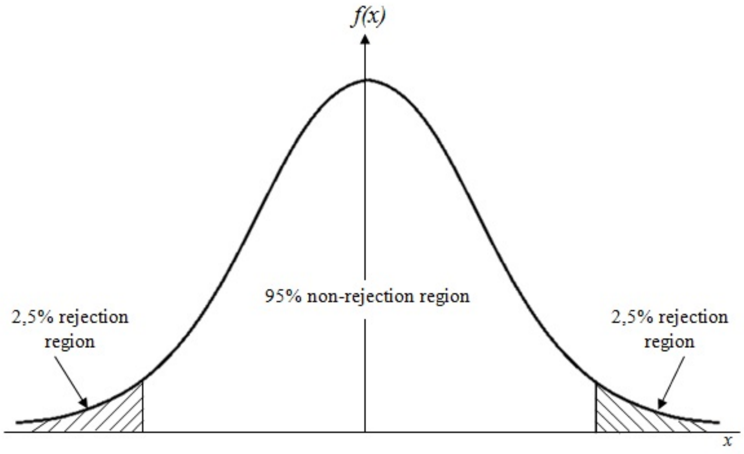

Figure 1: Rejection regions for a two sided hypothesis test

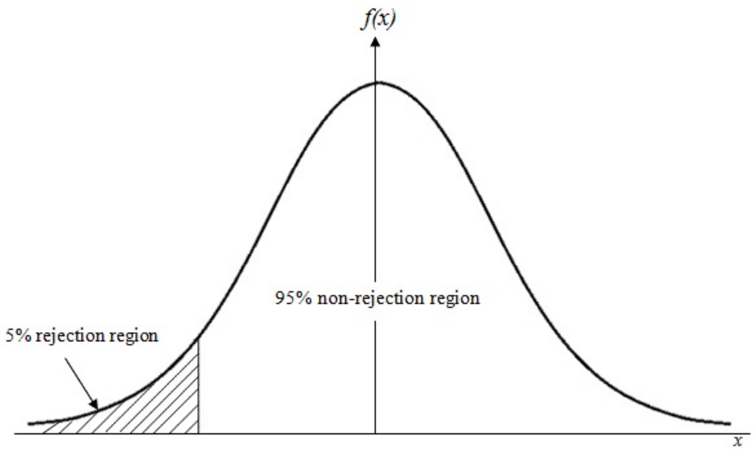

Figure 2: Rejection regions for a one sided hypothesis test of the form $H_0 : \beta = \beta^*$ and $H_1 : \beta < \beta^*$

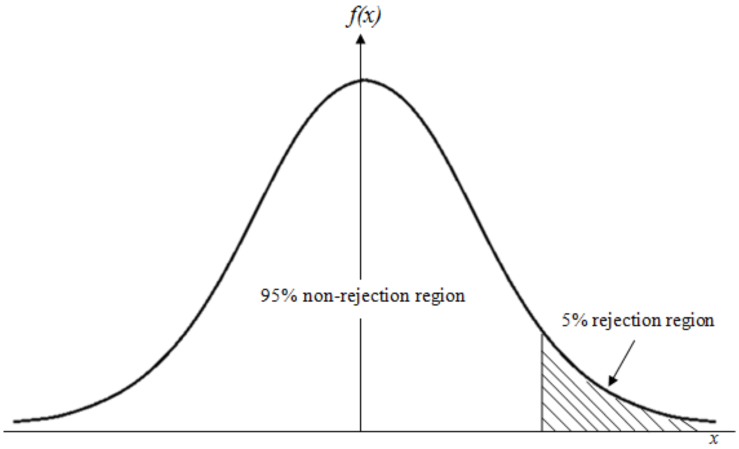

Figure 3: Rejection regions for a one sided hypothesis test of the form $H_0 : \beta = \beta^*$ and $H_1 : \beta > \beta^*$

- Use t-table to find critical value to compare the t-statistics, the critical value will be that value of x that puts 5% into the rejection region.
- Perform the test: if t-statistics lies in the rejection region then reject H_0, else do not reject H_0

2. Carrying out a hypothesis test using confidence intervals
   - Estimate the model parameters and their standards errors as usual
   - Choose a significance level α, conventionally is 5%
   - Use the t-tables to find the appropriate critical value, which will again have $T-2$ degrees of freedom.
   - The confidence interval for the parameter β is given by:

$$(\hat{\beta} - t_{crit}.SE\hat{\beta}, \hat{\beta} + t_{crit}.SE\hat{\beta}) \quad (2)$$

   Where (.) stands for multiplication of two quantities.
   - Perform the test: if the hypothesized value of value β lies outside the confidence interval, C.I, then reject H_0, otherwise do not reject H_0

2.3 Parametric hypothesis testing

Parametric hypothesis test make assumptions about the underlying distribution of the population from which the sample is being drawn, and which is being investigated. Parametric hypothesis tests include, ANOVA applied while comparing the means of several samples, Chi-Square Test, while testing 'goodness of fit' to an assumed distribution, contingency tables applied when a variation of the chi-square test, F-test while comparing variances, Proportion test, for differences between large or small proportions, t-test, while comparing the mean to a value, or the means of two samples, z-test known as t-test but for large samples [8].

If the distribution of the studied population is not known then a nonparametric test is suggested but this one is not powerful because it can not use predictable properties of the distribution.

2.4 Non-parametric hypothesis testing

As it says in [3] Nonparametric tests, known also as distribution free-tests, are valid for any distribution, it can be used either when the distribution is unknown or known, there are based on "order statistics" and are very simple.

The non-parametric tests are various and distinguished according to the inference population, thus we can cite among (i) the inference on one population, the runs test, The Binomial Test, The Chi-Square Goodness of Fit Test, The Kolmogorov-Smirnov Goodness of Fit Test, The Lilliefors Test for Normality, The Shapiro-Wilk Test for Normality (ii)contingency table, the $2x2$ contingency table, the rxc contingency table, the chi-square test of Independence, the measure of Association Revisited (iii) inference on two Population the Tests for Two Independent Samples, the tests for Two Paired Samples and (iv) inference on more than two populations, The Kruskal-Wallis Test for Independent Samples, The Friedmann Test for Paired Samples, The Cochran Q test [8]:

Exemple [3]: *Sign test for the median*

A median of the population is a solution $x = \tilde{\mu}$ of the equation $F(x) = 0.5$, where F is the distribution function. Suppose that eight radio operators were tested, first in rooms without air conditioning and then in air-conditioned rooms over the same period of time, and the difference of errors (unconditioned minus conditioned) were:

9 4 0 6 4 0 7 11

Test the hypothesis $\tilde{\mu} = 0$ (that is, air conditioning has no effect) against the alternative $\bar{\mu} > 0$ (that is, inferior performance in unconditioned rooms).

Solution. We choose arbitrary the significance level $\alpha = 5\%$. If the hypothesis is true, the probability p of a positive difference is the same as that of a negative difference. Hence in this case, $p = 0.5$, and the random variable *number of positive values among n values* has a binomial distribution with $p = 0.5$. our sample has eight values. We omit the value 0, which do not contributes to the decision. then six values are left, all of which are positive. since

$P(X = 6) = (Probability \quad of \quad 6 \quad out \quad of \quad 6 \quad events \quad to \quad occur)$
$= (0.5)^6(0.5)^0$
$= 0.0156 = 1.56\% < 0.5\%,$

We reject the null hypothesis and assert that the number of errors made in unconditioned rooms is significantly higher, so that installing air-conditioning should be considered.

2.5 Type I and Type II errors in Hypothesis testing

Every testing always involve risks of making false decisions, therefore we define [3]:

- Type I error: It is an error made while rejecting a true hypothesis, α is designed as the probability of making a type I error.
- Type II error: It is an error made while accepting a false hypothesis, β is designed the probability of making a type II error.

It is obvious that we can not avoid these errors because uncertainties in sample data drawn from the population, but there are ways and means of choosing suitable levels of risks, that is, of values α and β. the choice of α depends on the nature of the problem (e.g: a small risk $\alpha = 1\%$ is used if it is a matter of life or death).

2.6 Some Available Hypothesis Tests in Matlab

2.6.1 Description of the Tests

There exist several Hypothesis tests functions in Matlab according to what kind of test is needed, In our current work we focus on tests about comparing two random samples [4].

Ansari-Bradley Test of hypothesis:

Ansari-Bradley test, Tests if two independent samples come from the same distribution, against the alternative that they come from distributions that have the same median and shape but different variances. The result is h = 0 if the null hypothesis of identical distributions cannot be rejected at the 5% significance level, or h = 1 if the null hypothesis can be rejected at the 5% level. the two vectors can have different lengths.

The Ansari-Bradley test is a nonparametric alternative to the two-sample F test of equal variances. It does not require the assumption that the two vector come from normal distributions. The dispersion of a distribution is generally measured by its variance or standard deviation, but the Ansari-Bradley test can be used with samples from distributions that do not have finite variances.

The theory behind the Ansari-Bradley test requires that the groups have equal medians. Under that assumption and if the distributions in each group are continuous and identical, the test does not depend on the distributions in each group. If the groups do not have the same medians, the results may be misleading. Ansari and Bradley recommend subtracting the median in that case, but the distribution of the resulting test, under the null hypothesis, is no longer independent of the common distribution of the two vector. If you want to perform the tests with medians subtracted, you should subtract the medians from the two vector before calling ansaribradley.

Jacques-Berra test

Jarque-Bera test. Tests if a sample comes from a normal distribution with unknown mean and variance, against the alternative that it does not come from a normal distribution.

T-test

One-sample or paired-sample t-test. Tests if a sample comes from a normal distribution with unknown variance and a specified mean, against the alternative that it does not have that mean.

We performs a t-test of the hypothesis that the data in the vector X come from a distribution with mean zero, and returns the result of the test in H. $H = 0$ indicates that the null hypothesis ("mean is zero") cannot be rejected at a given significance level. H=1 indicates that the null hypothesis can be rejected at a the same level. The data are assumed to come from a normal distribution with unknown variance. We test if $H_0 : x_0 = x_1 \quad against \quad H_1 : x_0 \neq x_1$

Kolmogorov-Smirnov (K-S) test

We distinguish two kind of this test; One-sample Kolmogorov-Smirnov test. Tests if a sample comes from a continuous distribution with specified parameters, against the alternative that it does not come from that distribution. Two-sample Kolmogorov-Smirnov test. Tests if two samples come from the same continuous distribution, against the alternative that they do not come from the same distribution.

2.6.2 Simulation and Accuracy of the Test

In order to check the accuracy of the test of hypothesis, let's generate two vectors $r1$ and $r2$ from "randn" matlab function as shown in the figure 4 and 5, where data are identically, independently distributed (i.i.d) and normally distributed. The interaction between the two random number is observed in the figure **??**.

Knowing that the normal probability density function is given by:

$$y = f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{3}$$

In our case while using the Gaussian distribution $\mu = 0$ and $\sigma = 1$

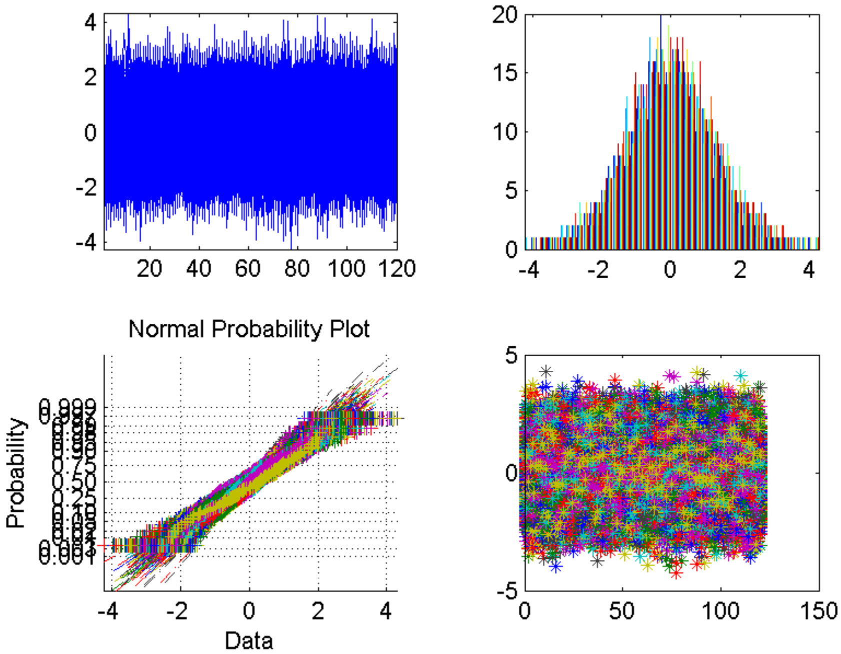

Figure 4: Normal random vectors r1 generated a thousand times

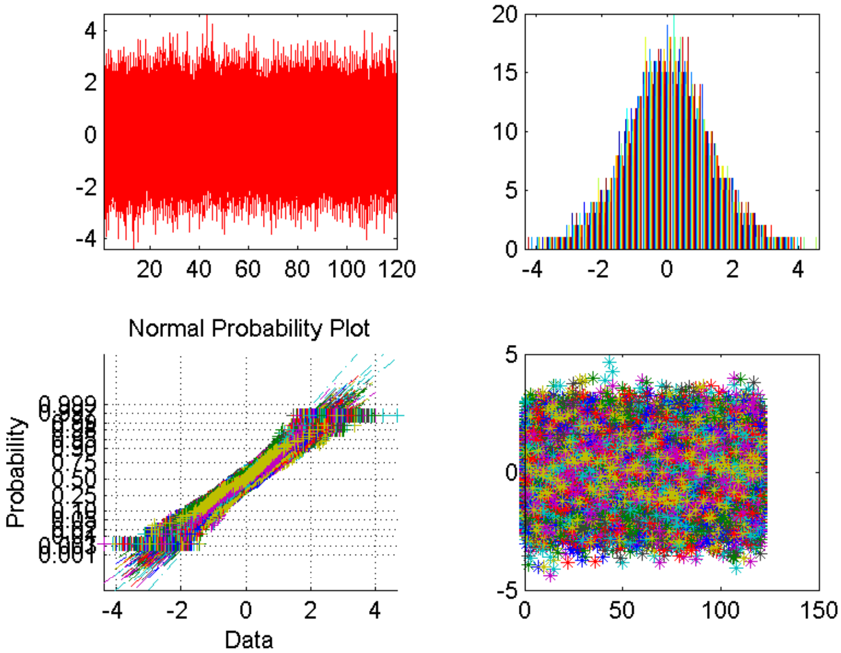

Figure 5: Normal random vectors r2 generated a thousands times

As a first step, you might want to test the assumption that the samples come from normal distributions. A normal probability plot gives a quick idea in the figure 4 and 5. Both scatters approximately follow straight lines, indicating approximate normal distributions.

Performing the tests one thousand times in Matlab the figure 6 below showed how many times in percentage the test itself can fail to give the right answer although we know already the outcome of the test. For instance the Ansari-Bradley showed that two random numbers $N(0,1)$ and $N(0,1)$ are identical distribution only $96,4\%$ times and $3,6\%$ are not identical which is not true in reality.

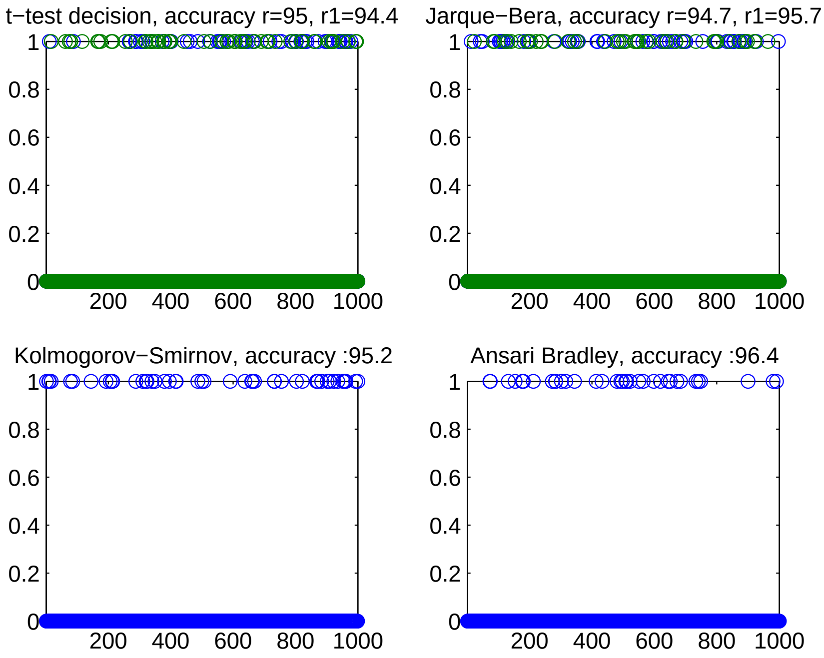

Figure 6: Accuracy of while same distribution

In same manner let's see what happen for the testing of two different distribution where $r1$ is a uniform distribution $U(0,1)$ and $r2$ is from the normal distribution $N(0,1)$ as presented in the figure 7, 8 and 9:

Knowing that the uniform cumulative density function (cdf) is given by:

$$y = f(x|a,b) = \frac{x-a}{b-a} I_{[a,b](x)} \tag{4}$$

In our case while using the standard uniform distribution $a = 0$ and $b = 1$.

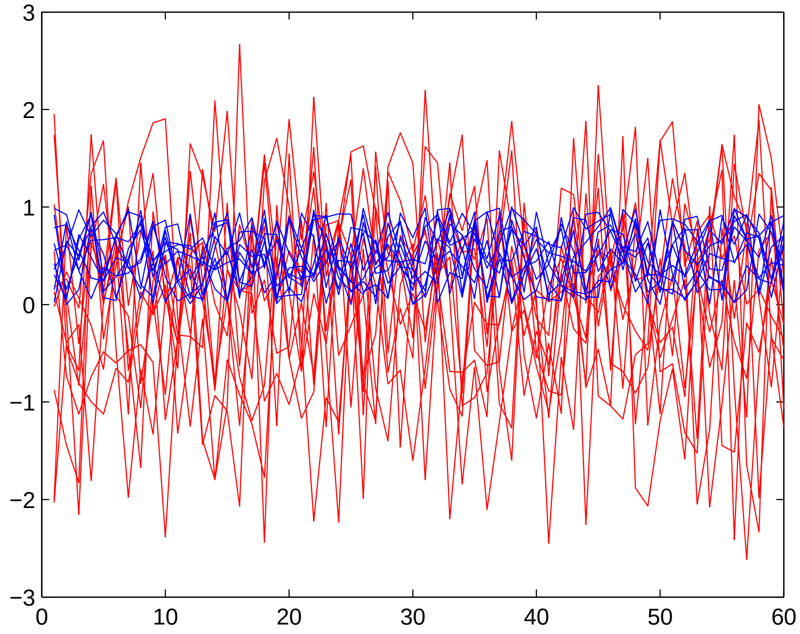

Figure 7: Two random vectors generated a thousand times

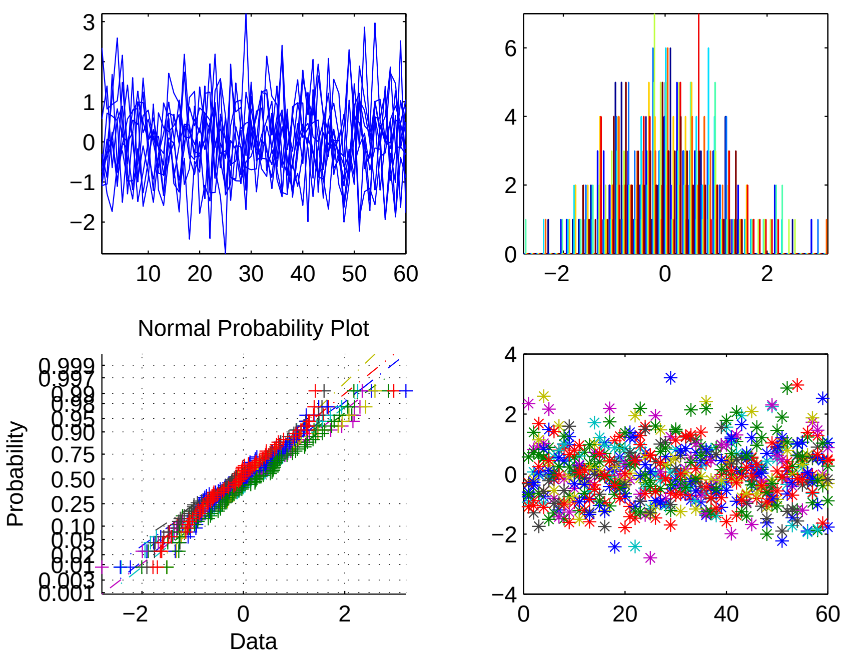

Figure 8: A uniform random vectors $r2 = U(0,1)$ generated a thousand times

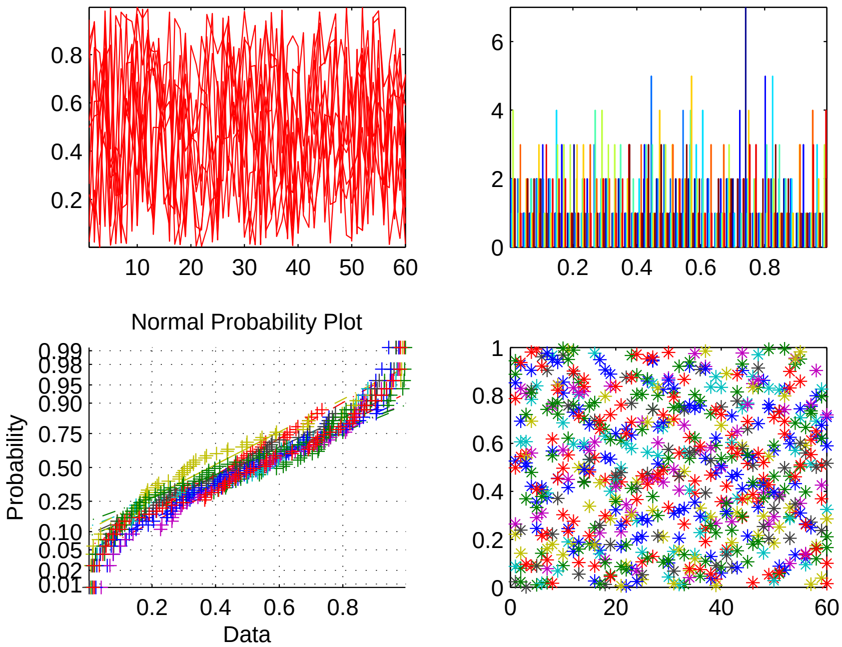

Figure 9: Normal random vectors $r2 = N(0,1)$ generated a thousand times

According to each test we show in the figure 10 how much a null hypothesis were not rejected ($H_0 = 1$) although it should be rejected. Specifically the Ansari-bradley and Kolmogorov-Smirnov perfomed 100% good but Jacque-Berra test showed that only $94,9\%$ times a random number $N(0,1)$ is normal distributed and $5,1\%$ is not.

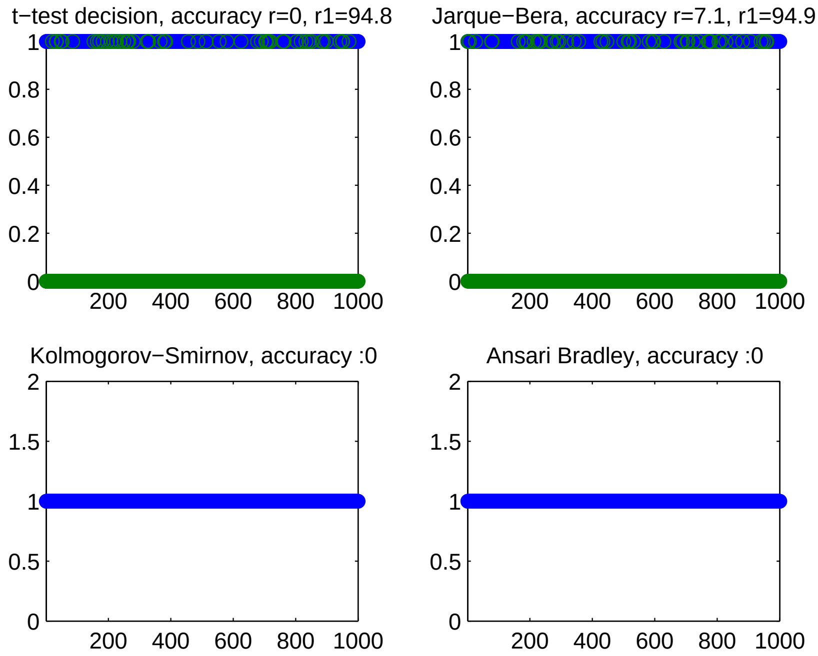

Figure 10: accuracy while different distribution

2.7 Pvalue Approach in Hypothesis Testing

As it says in [23] a p-value is defined as the probability of observing a value of the test statistic as extreme as or more extreme than the one that is observed, when the null hypothesis is true. Instead of comparing the observed value of a test statistic with a critical value, the probability of occurrence of the test statistic, given that the null hypothesis is true, is determined and compared to the level of significance α. The null hypothesis is rejected if the P value is less than the designated α. [11]

The procedure of determining testing an hypothesis with p-value approach is illustrated here under: [23]

1. Determine the null and alternative hypothesis, H_0 and H_1.
2. Find a test statistic T that will provide evidence about H_0
3. Obtain a random sample from the population of interest and compute the value of the statistics t_0 from the sample.
4. Calculate the p-value:

- Lower Tail Test: $P - value = P_H^0(T \leq 0)$
- Upper Tail Test: $P - value = P_H^0(T \geq 0)$

5. If the $p - value \leq \alpha$, then reject the null hypothesis.

For the two-tail test, the p-value is determined similarly.

2.8 Bootstrap methods

2.8.1 Definition

The treatment of the bootstrap methods described here comes from Efron and Tibshirani [1993]. According to [23] the bootstrap methods refer to the resampling techniques. Here, we use bootstrap to refer to Monte Carlo simulations that treat the original sample as the pseudo-population or as an estimate of the population. Thus, in the steps where we randomly sample from the pseudo-population, we now resample from the original sample, No new data is actually produced, but new combinations from the existing data [10]

In the book [19] a potentially more powerful test is provided by the boostrap confidence interval for the variance ratio where the repetition resampling without replacement is done. In this work we have performed the test while bootstrapping the generated random number and see how the results performed, we test the hypothesis using the p-value value approach as described in the Matlab algorithm here under.

2.8.2 Algorithm

According to definition of the bootstrapping methods learnt we have created our bootstrapping Matlab function able to sample with replacement in the original data and calculate each time the p-value and provide a hypothesis decision.

1. Draw or generate element from an $N(0,1)$ known as normal random numbers
2. Set nRep number of how many times the generation will repeat
3. Compute the P-value of each bootstrap sample by creating the function made of:
   - *Input*:

  - sample - A vector of data to be tested
  - Blocksize - Size of the bootstrap blocks
  - nb - Number of bootstrap iterations
  - alpha - Significance level

- *Output*:
  - H - the function returns hypothesis decision to 0 or 1
  - P-value - based on which the decision about the hypothesis was made and this has made to be the average of all computed p-values among all bootstrap blocks sampled randomly by the Matlab function "randperm"

4. make the hypothesis decision basing the rule: if P-value is more than significance level alpha then do not reject the null hypothesis otherwise reject.

5. Plot the non reject and the rejected cases so as the accuracy percentage of the test.

2.8.3 Simulation and Bootstrapping Method Hypothesis Testing Accuracy

Applying the above algorithm to some Matlab test of hypothesis we have found that when the bootstrapping method is applied on data the accuracy is 100% in many case.

Only two bootstrap function were created while testing with T-test and Jacque-Bera test therefore for two normal random numbers $r1$ and $r2$ the test showed to be 100% accurate as shown in the figure 11:

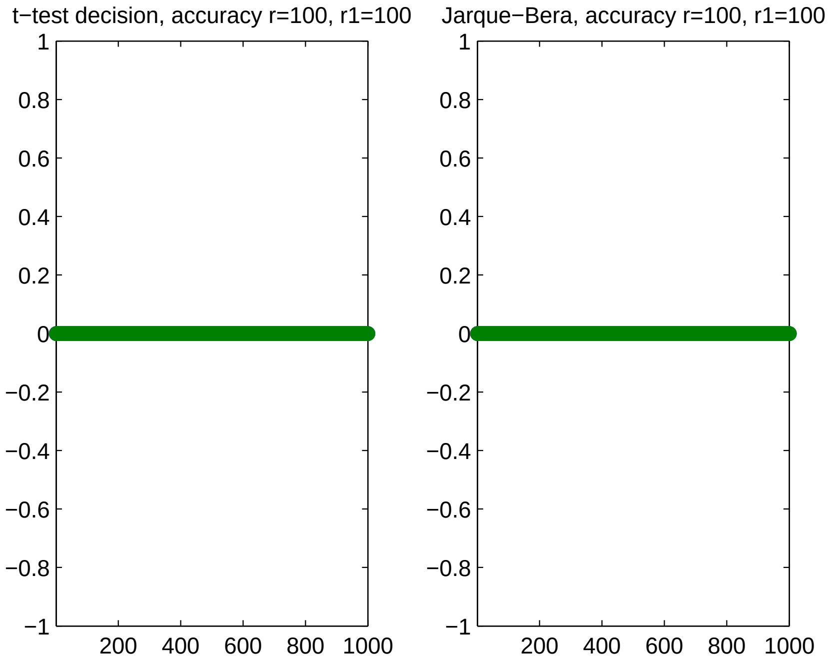

Figure 11: Testing while bootstrapping where the blocksize is 5 and $\alpha = 0.1$

The significance level in this case were set to be $\alpha = 0.1$ the above results are explained by the fact that P-values are as presented in the figure 12:

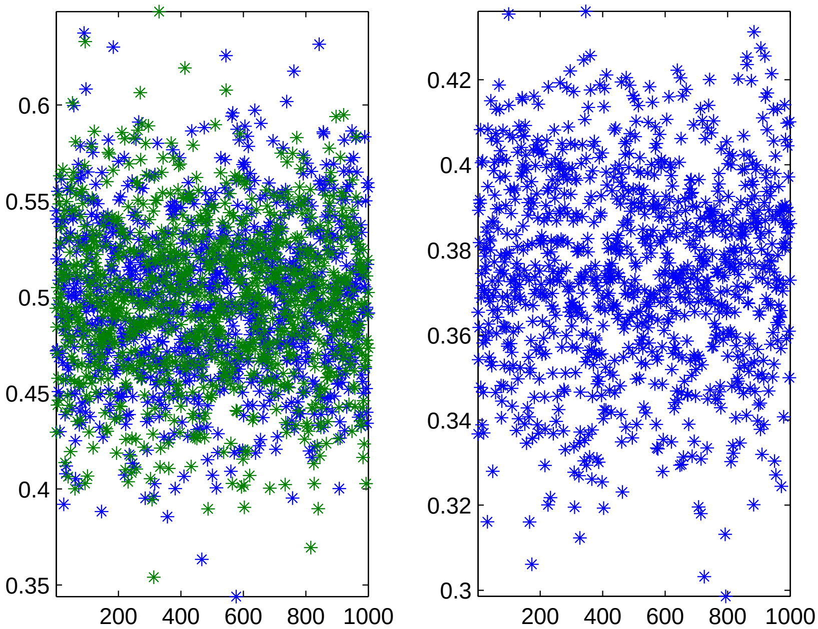

Figure 12: P-value computed at each bootstrap step i.e 1000 times

3 PORTFOLIO PERFORMANCE MEASUREMENT OVERVIEW

3.1 Performance measurement

There exists many Portfolio performance measurement but the most commonly used nowadays are the Sharpe ratio, Treynor Ratio, Jensen Alpha. [2] and Appraisal ratio, while looking at the historic return and risk.

The performance measurement allows to assess and compare the performance (or past returns) of different investment strategies. [13]

For instance once we need to compare a passive and an active investment strategies; where by definition, a passive investment strategy is when an investor holds a portfolio that is an exact copy of the market index and does not rely on superior information in contrast an active investment strategy is when an investor's portfolio differs from the market index by having different weights in some or all of the shares in the market index and the active investor relies on having superior information therefore increment of much greater cost than the passive investor.

3.2 Sharpe Ratio

3.2.1 Definition

Among the portfolio performance investment cited above we will discuss and develop the Sharpe Ratio.

The Sharpe ratio (also known as Reward-to-Volatility-Ratio) is calculated by subtracting the risk-free rate from the rate of return for a portfolio and dividing the result by the standard deviation of the portfolio returns; in other words, the Sharpe Ratio indicates the excess return per unit of risk associated with the excess return. The higher the Sharpe Ratio, the better the performance.

$$SR = \frac{R_i - R_f}{\sigma_i} \tag{5}$$

Where: R_i = the portfolio return during the observation period, R_f = the risk free rate of the return and σ_i = the standard deviation of the return of investment.

3.2.2 Measuring with the Sharpe Ratio

Graphically, the Sharpe Ratio is the slope of a line between the risk free rate of the return and the portfolio in the mean/volatility space. The efficient portfolio in the mean-variance framework with a risk free asset is to maximizing the Sharpe Ratio of the portfolio [2].

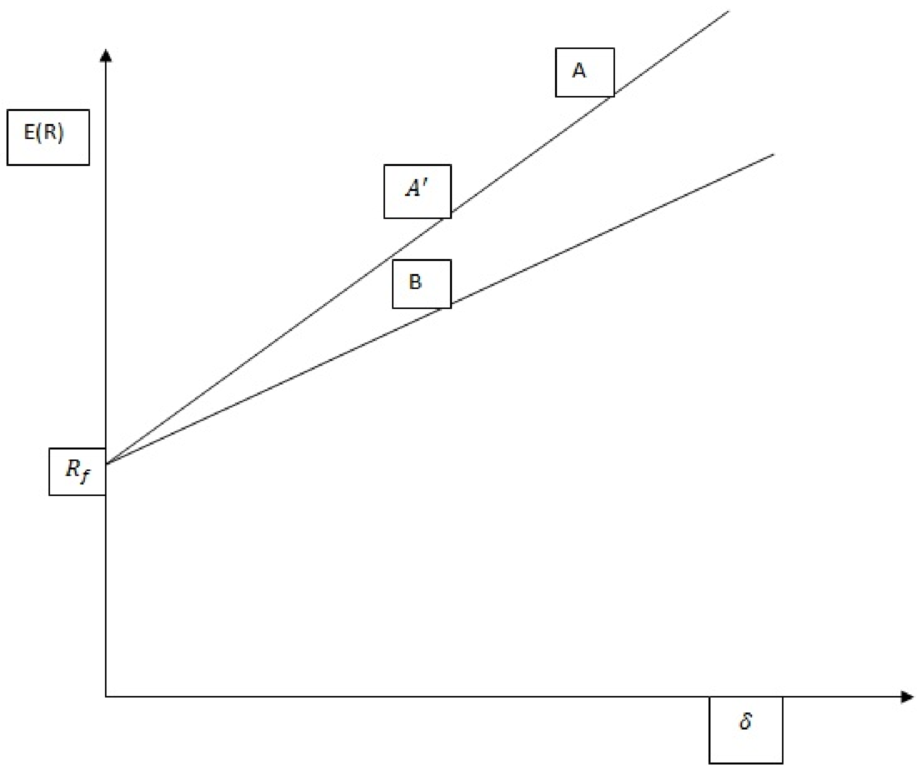

Figure 13: Excess Return Sharpe Ratios for Two Funds

In the above figure, $E(R)$ stands for Expected return, R_f is the Risk free rate of the return and σ the standard deviation of the return; let's consider an investor who plans to put all her money in either fund A or fund B.

Also, assume that the graph plots the best possible predictions of future expected return and future risk, measured by the standard deviation of return. An investor might choose A, based on its higher expected return, despite its greater risk. Or, she might choose B, based on its lower risk, despite its lower expected return.

Her choice should depend on her tolerance for accepting risk in pursuit of higher expected return. Absent some knowledge of her preferences, an outside analyst cannot argue that A is better than B or the converse.

But what if the investor can choose to put some money in one of these funds and the rest in treasury bills which offer the certain return shown at point R_f? Say that she has decided that she would prefer a risk (standard deviation) of for instance 10%. She could get this by putting all her money in fund B, thereby obtaining an expected return of 11%. Alternatively, she could put $\frac{2}{3}$ of her money in fund A and $\frac{1}{3}$ in Risk free (Treasury Bill or T-Bill). This would give her the prospects plotted at point $A^{'}$.

The same risk (10%) and a higher expected return (12%). Thus a Fund/Risk free strategy using fund A would dominate a Fund/Risk free strategy using fund B. This would also be true for an investor who desired, say, a risk of 5%. And, if it were possible to borrow at the same rate of interest, it would be true for an investor who desired, say, a risk of 15%. In the latter case, fund A (by itself) would dominate a strategy in which fund B is levered up to obtain the same level of overall risk.

Prospectively, the excess Return Sharpe Ratio is best suited to an investor who wishes to answer the question:

If I can invest in only one fund and engage in borrowing or lending, if desired, which is the single best fund?

Retrospectively, an historic Excess Return Sharpe Ratio can provide an answer for an investor with the question:

If I had invested in only one fund and engaged in borrowing or lending, as desired, which would have been the single best fund? [2] [24]

In the real world, there are situations in which funds underperform the risk free rate of the return on average and hence have negative average excess returns. In such cases it is often considered that a fund with greater standard deviation and worse average performance may nonetheless have a higher (less negative) excess return Sharpe Ratio and thus be considered to have been better. Let's consider the figure below.

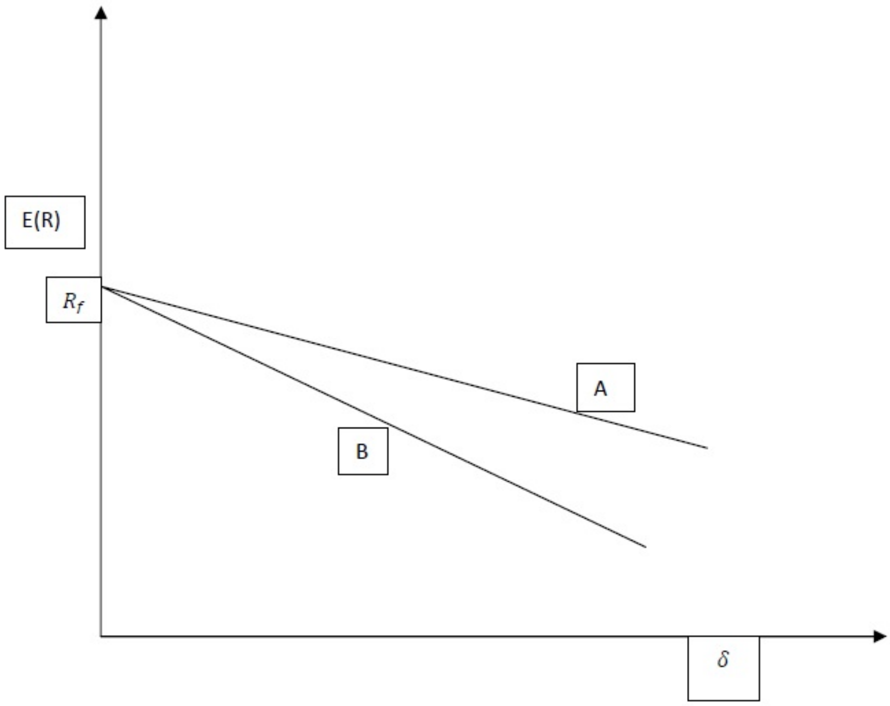

Figure 14: Average negative excess Return Sharpe Ratios for Two Funds

Similarly in the above picture, A is clearly inferior to B (and both were inferior to R_f). But, for an investor who had planned for a standard deviation of 10%, the combination of 2/3 A and 1/3 R_f would have broken even, while investment in fund B would have lost money. Thus a Fund/Risk free strategy using the fund with the higher (or less negative) Excess Return Sharpe Ratio would have been better. Also, one would never invest in funds such as A or B if their prospects involved risk with negative expected excess returns [24]

3.2.3 Sharpe Ratio and T-statistics Relationship

The Historic Sharpe Ratio can be related to the t-statistic or t-ratio for measuring the statistical significance of the mean excess return [21].

$$t - ratio = \frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})} \tag{6}$$

Where: $\hat{\beta}$ is an estimator of the model parameter β, β_0 is zero if the test is H_0 :

$\beta = 0$ and $H_1 : \beta \neq 0$ and $SE(\hat{\beta})$: is the standard error of the $\hat{\beta}$

And the Historic Sharpe Ratio noted as SR_h is equal to:

$$SR_h = \frac{\frac{1}{T}\sum_{t=1}^{T}(R_{it} - R_{ft})}{\sigma_D} \qquad with \qquad \sigma_D = \sqrt{\frac{\sum_{t=1}^{T}(D_t - \bar{D})}{T-1}} \tag{7}$$

Where: R_{it} is the portfolio return in period t, R_{ft} is the risk free rate of the return in period t and σ_D is the standard deviation over the given period, $D_t = R_{it} - R_{ft}$ and $\bar{D} = \frac{1}{T}\sum_{t=1}^{T} D_t$

Therefore, from equations 6 and 7 the t-statistic can be equal the Sharpe Ratio times the square root of T (the number of returns used for the calculation). If historic Sharpe Ratios for a set of funds are computed using the same number of observations, the Sharpe Ratios will thus be proportional to the t-statistics of the means.

The Sharpe Ratio is measured and used without any tests about statistical significance. But a test whether the difference between two Sharpe Ratios is zero can be processed [18]

A Sharpe Ratio can be computed by the mean and standard deviation of the distribution of the final payoff. [21] It can also be measured by the expected return per unit of standard deviation of return for a zero-investment strategy.

3.3 Applicability of Performance Hypothesis Testing with Sharpe Ratio

The determination of statistical significance of the Sharpe ratio difference between portfolios has been widely discussed in the financial literature (e.g. see Jobson and Korkie, 1981 [9]; Vinod and Morey, 1999 [25]; Memmel, 2003 [15]; Ledoit and Wolf, 2008 [18]).

The most popular test for such purpose is still the Jobson-Korkie (1981) [9] test that has been criticized due to its restrictive assumptions related to the characteristics of the return distributions being compared (e.g., see Lo, 2002 [12]). Ledoit and Wolf (2008) [18] prove that the Jobson-Korkie test statistic is not valid if either or both of the return distributions being analyzed are non-normal or if the observations are correlated over time.

In addition, inability of the standard Sharpe ratio to cope with negative excess returns restricts the applicability of the Jobson-Korkie test for such cases. Israelsen

(2005) [6] introduces the adjustment procedure for valid comparison of negative Sharpe ratios. Unfortunately, it can not be applied in the context of the Jobson-Korkie type test without validity loss.

3.4 Comparing Performance Difference between Portfolios with Negative Excess Returns

The dilemma of comparing negative Sharpe ratios is well-recognized in financial literature, but it was not solved until 2003 by Israelsen (2003, 2005) [5] and [6]. The dilemma stems from the fact in some cases the traditional interpretation of the Sharpe ratio (the bigger, the better) may lead to irrational conclusions about performance ranking when excess returns are negative.

For example, let us first consider two real-world portfolios: The average excess monthly return over the 36-month evaluation period for portfolio A is -1.815% and its volatility is 4.736%. For portfolio B, the corresponding numbers are -2.420% and 6.790%. Therefore, the unadjusted Sharpe ratios are -0.383 for portfolio A and -0.354 for portfolio B indicating the slight outperformance of B over A. However, the loss of portfolio A is smaller than that of B, while the risk of B is distinctly higher.

Therefore, very few investors would be willing to prefer B over A. According to Israelsen's refinement method the problem can be solved by powering the denominator of the Sharpe ratio by the ratio of excess return to its absolute value. In this particular case, the refined Sharpe ratios are -0.086% for portfolio A and -0.163% for B, indicating the clear out performance of A over B. However, if the assumptions of both normality and I.I.D. data held, the refined Sharpe ratios would still be inapplicable to the Jobson-Korkie-Memmel (JKM) type performance difference tests, since such statistical tests can not cope with negative excess returns.

on the other hand, we can say that Negative Sharpe Ratios are difficult to interpret. Some people even reject the Sharpe Ratio altogether because of this.

The problem is the following: it is generally assumed that people have a preference for 'more return' and 'less risk'. Risk in the context of the Sharpe Ratio is return volatility. One would therefore expect that when ranking portfolios with equal returns by their Sharpe Ratios, portfolios with lower volatilities are preferred to portfolios with higher volatilities. This is not the case when the returns are negative!

More formally: Given two portfolios X and Y with,

$r(X) = -5\%, r(Y) = -5\%$ $v(X) = 20\%, v(Y) = 25\%$

Calculating the Sharpe Ratios of portfolios X and Y gives,

$SR(X) = r(X)/v(X) = -5/20 = -0.25$ $R(Y) = r(Y)/v(Y) = -5/25 = -0.20$

Since we are dealing with negative number here, -0.25 is a smaller than -0.20 and we get $SR(X) < SR(Y)$. This means that that portfolio Y is preferred to portfolio X because it has a higher Sharpe Ratio, even though portfolio B has the larger volatility.

3.5 Comparing Performance Difference Between Portfolios with Excess Returns of different Sign

Suppose we would like to test the statistical significance of outperformance of Portfolio C with positive excess return against the same two portfolios that we used earlier in our example of comparing portfolios with negative excess returns.

According to previously-done pairwise comparison, Portfolio A is preferable to Portfolio B due to its higher mean return and lower volatility.

As an empirical example, let us compare the performance between portfolio A and portfolio C, whose average monthly return is 1.693% and corresponding volatility is 6.831% for the evaluation period. By subtracting the average excess return of the worse portfolio (ie., that of A) from each of the original time-series returns being compared, the new average excess returns are 0% for portfolio A and 3.508% for portfolio C.

As the above-described subtraction does not affect volatilities, the statistical comparison is now possible without possible bias caused by negative Sharpe ratio of another portfolio.

3.6 Hypothesis testing with the RobustSharpe ratio

3.6.1 Description of the Problem

In this part we stated the problem as it was stated in the paper of Ledoit-Wolf [18] for the well understanding the process of RobustSharpe testing of hypothesis. Using the same notation as in Jobson and Korkie(1981) [9] and Memmel (2003) [15]. suppose that we have two investment strategies i and n whose excess returns over a given benchmark at time t are r_{ti} and r_{tm}, respectively. Typically, the benchmark is the risk-free rate.

A total of T return pairs $(r_{1i}, r_{1n}), \ldots, (r_{Ti}, r_{Tn})$ are observed. It is assumed that these observations constitute a strictly stationary time series so that, in particular, the bivariate return distribution does not change over time. This distribution has mean vector μ and covariance matrix Σ given by:

$$\mu = \begin{pmatrix} \mu_i \\ \mu_n \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \sigma_i^2 & \sigma_{in} \\ \sigma_{in} & \sigma_n^2 \end{pmatrix} \quad (8)$$

The usual sample means and sample variances of the observed returns are denoted by $\widehat{\mu}_i, \widehat{\mu}_n$ and $\widehat{\sigma}_i^2, \widehat{\sigma}_i^2$ respectively. The difference between the two Sharpe ratios is given by

$$\Delta = Sh_i - Sh_n = \frac{\mu_i}{\sigma_i} - \frac{\mu_n}{\sigma_n}$$

And the estimator is

$$\widehat{\Delta} = \widehat{Sh_i} - \widehat{Sh_n} = \frac{\widehat{\mu}_i}{\widehat{\sigma}_i} - \frac{\widehat{\mu}_n}{\widehat{\sigma}_n}$$

Furthermore, let $u = (\mu_i, \mu_n \, \sigma_i^2, \sigma_n^2)'$ and $\widehat{u} = (\widehat{\mu}_i, \widehat{\mu}_n, \widehat{\sigma}_i^2, \widehat{\sigma}_n^2)'$. The standard error for $\widehat{\sigma}$ is computed based on the relation,

$$\sqrt{T}(\widehat{\mu} - \mu) \underset{\rightarrow}{d} N(0; \Omega),$$

where $\underset{\rightarrow}{d}$ denotes convergence in distribution, and an application of the delta method. However, The formula for Ω that crucially relies on i.i.d. return data from a bivariate normal distribution is

$$\Omega = \begin{pmatrix} \sigma_i^2 & \sigma_{in} & 0 & 0 \\ \sigma_{in} & \sigma_n^2 & 0 & 0 \\ 0 & 0 & 2\sigma_i^4 & 2\sigma_{in}^2 \\ 0 & 0 & 2\sigma_{in}^2 & 2\sigma_n^4 \end{pmatrix} \quad (9)$$

This formula is no longer valid if the distribution is non-normal or if the observations are correlated over time. To give just two examples, consider data that are i.i.d. but not necessarily normal.

First, the entry in the lower right corner of Ω is given by $E[(r_{1n} - \mu_n)^4] - \sigma_i^4$ instead of $2\sigma_i^4$.

Secondly, the asymptotic covariance between μ_n and ${\mu^2}_n$ say, is in general not equal to zero.

To give another example, consider data from a stationary time series. Then the entry in the upper left corner is given by $\sigma_i^2 + 2\sum_{t=1}^{\infty} cov(r_{1i}, r_{(1+t)i})$ instead of by simply σ_i^2.

3.6.2 Theoretical Solution

The theoretical solution solution of the above problem has been as well solved in the paper of Ledoit-Wolf [18] and we described it hereunder: Ledoit et al. conveniently worked with the uncentered second moments in the following manner:

Let $\gamma_i = E(r_{1i}^2)$ and $\gamma_i = E(r_{1i}^2)$. Their sample counterparts are denoted by $\widehat{\gamma}_i$ and $\widehat{\gamma}_n$, respectively [18].

Furthermore, let $v = (\mu_i, \mu_n, \gamma_i, \gamma_n)$ and $\widehat{v} = (\widehat{\mu}_i, \widehat{\mu}_n, \widehat{\gamma}_i, \widehat{\gamma}_i)$.

which allowed to write

$$\Delta = f(\upsilon) and \widehat{\Delta} = f(\widehat{\upsilon})$$

With

$$f(a,b,c,d) = \frac{a}{\sqrt{c-a^2}} - \frac{b}{\sqrt{d-b^2}}$$

Assuming that

$$\sqrt{T}(\widehat{\upsilon} - \upsilon) \underset{\rightarrow}{d} N(0;\Psi),$$

where Ψ is an unknown symmetric positive semi-definite matrix. This relation holds under mild regularity conditions. For example, when the data are assumed i.i.d., it is sufficient to have both $E(r_{1i}^4)$ and $E(r_{1i}^4)$ finite. In the time series case it is sufficient to have finite $4+\sigma$ moments, where σ is some small positive constant, together with an appropriate mixing condition, The delta method then implies:

$$\sqrt{T}(\widehat{\Delta} - \Delta) \underset{\rightarrow}{d} N(0;\Delta' f(v)\Psi\Delta f(v))$$

with

$$\nabla' f(a,b,c,d) = \left(\frac{c}{(c-a^2)^{1.5}}, -\frac{d}{(d-b^2)^{1.5}} - \frac{1}{2}\frac{a}{(c-a^2)^{1.5}}, -\frac{1}{2}\frac{b}{(d-b^2)^{1.5}} \right).$$

If the estimator $\hat{\Psi}$ of Ψ exists then a standard error for $\hat{\Delta}$ is given by

$$S(\hat{\Delta}) = \sqrt{\frac{\nabla' f(\hat{\upsilon})\hat{\Psi}\nabla f(\hat{\upsilon})}{T}}. \quad (10)$$

3.6.3 Pseudo Algorithm of the RobustSharpe Ratio Matlab Function

The programming code cited in the paper of Ledoit et al. [18] were downloaded freely from internet [20] the function has been studied and used in our work therefore we generated its pseudo algorithm for the well understanding of this function as follow:

1. Set two column vectors to be tested by their sharpe ratios
2. Set nRep number of how many times the test will repeat
3. Compute the Sharpe ratio to be compared
4. Call the matlab function robustsharpe made of:
   - *Input*:
     - Data - [Tx2] matrix of excess returns
     - Alpha - fixed significance level; default value = 0.05
     - H0 - null hypothesized value for the value of Sharpe ratios difference; default value = 0
     - M - number of bootstrap iterations; default value = 5,000
     - bl - block size in Circular Block Bootstrap. Use routine optimalblrobustSharpe.m to determine optimal block size. If no block size is specified, optimalblrobustSharpe.m is called automatically, with default candidate block sizes 1,3,6,10,15
     - kernel - the Quadratic spectral (QS) is taken by defaults.
     - extsim - 1 if the indices matrix bootMat in the circular block bootstrap is fed in rather than simulated in robustSharpe itself, 0 else useful to achieve comparability of results based on other implementations.

    - bootMat - exogenous indices matrix in circular block bootstrap of size [MxT] or 0 where M is number of CBB iterations, T is time series length

- *Output*:
    - Rejected - 1 if H_0 was rejected at significance level alpha, 0 else.
    - pval - p-value.
    - teststat - test statistic.
- Set the inputs default values if needed.
- Start by calling the data (Tx2).
- Computation of studentized test statistic and generation Circular Block Bootstrap (CBB) Index Matrix.
- Prewhiten (see explanation in appendix) data with VAR(1) model and estimate HAC kernel estimator using AR(1) models as univariate approximating parametric models.
- Studentization of 'raw' test statistic and set the values of μ, the means of two return time series, Difference of Sharpe ratios and HAC std estimate.
- Generate M CBB matrices by X_{T*m} *where*, $1 \leq m \leq M$.
- Call function which determines a matrix with corresponding studentized test statistics for each bootstrap iteration (row), the simulated excess returns of two assets and the HAC std estimate of difference of two Sharpe ratios.
- Call another function that computes critical value and tests H_0

5. Plot the data in different plots to visualize the shape and distribution of data

6. Plot the non reject and the rejected cases so as the accuracy percentage of the test.

4 RESULTS

4.1 Data

Consider two applications to investment funds. In each case, we want to test the null hypothesis of statistical significant equality of the Sharpe ratios of the two funds.

For the first step we used simulated data from known distribution in order to measure the performance of different test.

Secondly we used the same data used in the paper of Ledoit and Wolf [18], where the first application deals with mutual funds, the selected funds are Fidelity (FFIDX), a 'large blend' fund, and Fidelity Aggressive Growth (FDEGX), a 'mid-cap growth' fund. The data were obtained from Yahoo! Finance; the second application deals with hedge funds, The selected funds are Coast Enhanced Income and JMG Capital Partners. The data were obtained from the CISDM database.

In above both applications, we use monthly log returns in excess of the risk-free rate. The return period is 01/1994 until 12/2003, so the period T was equal to 120 (see in the Appendix detailed data set).

4.2 Robust Sharpe Perfomance hypothesis testing

4.2.1 Simulation study and results

Consider two random number $r1$ and $r2$ generated from $N(0,1)$ the period is $T = 60$ and the data has been simulated 10 only times (due to the fact of time taken to run matlab code) both numbers can be presented in matlab plot as shown in the figures 15 and 16.

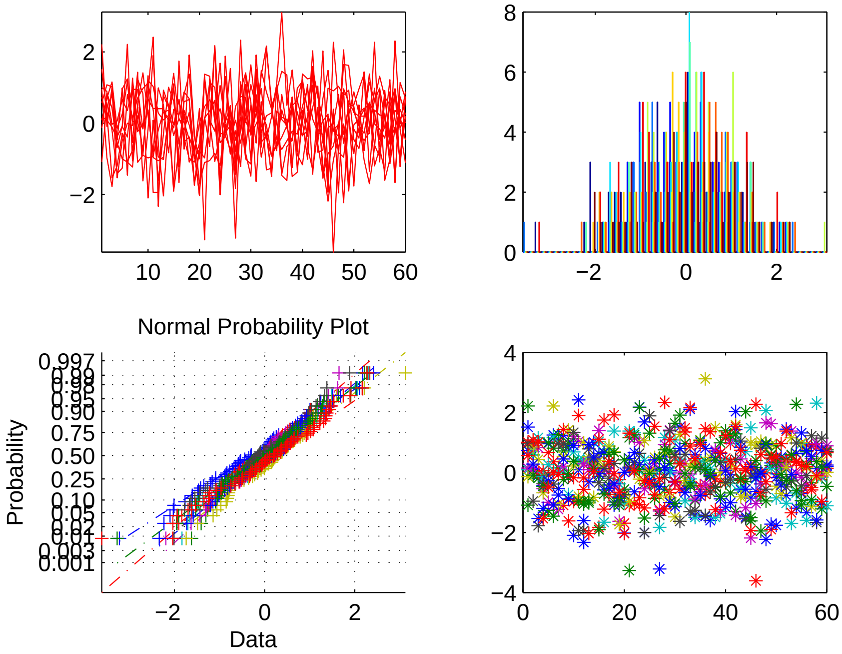

Figure 15: Normal random vectors r1 generated ten times

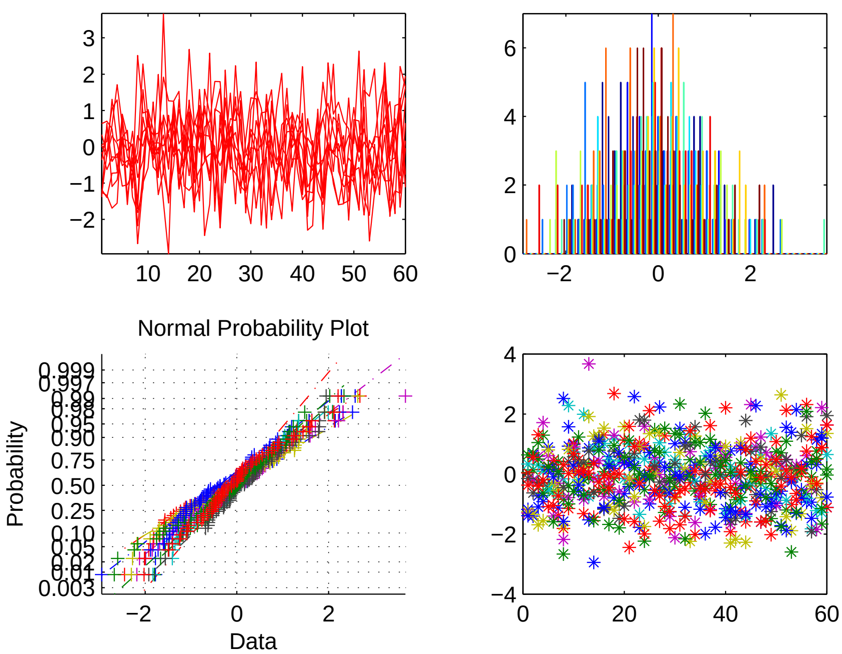

Figure 16: Normal random vectors r2 generated a ten times

The Generated 10 pair of random columns has been created and their Sharpe ratio has been calculated and reported in table 1:

Table 1: Simulated data from the same distribution

	Mean(r1)	**Mean(r2)**	**Std(r1)**	**Std(r2)**	**SR(1)**	**SR(2)**	**Decision H**
1	-0.0222	-0.1466	0.8623	1.0410	-0.0258	-0.1408	0
2	-0.0854	-0.0709	1.0074	1.0620	-0.0847	-0.0668	0
3	0.0935	-0.1785	0.9355	0.9858	0.0999	-0.1811	0
4	-0.0196	-0.1069	0.9204	1.0002	-0.021	-0.1069	0
5	-0.0394	0.1582	1.0465	1.0092	-0.0376	0.1568	0
6	-0.2314	0.1118	1.0020	1.0480	-0.2310	0.1067	0
7	0.1784	0.0893	0.8546	1.1728	0.2088	0.0761	0
8	-0.1603	0.2499	1.1180	0.8099	-0.1434	0.3086	1
9	0.0511	0.0025	1.0019	0.9811	0.0510	0.0026	0
10	0.1144	-0.0856	0.8704	0.9113	0.1315	-0.0939	0

Graphically we have the result as presented in the figure 17:

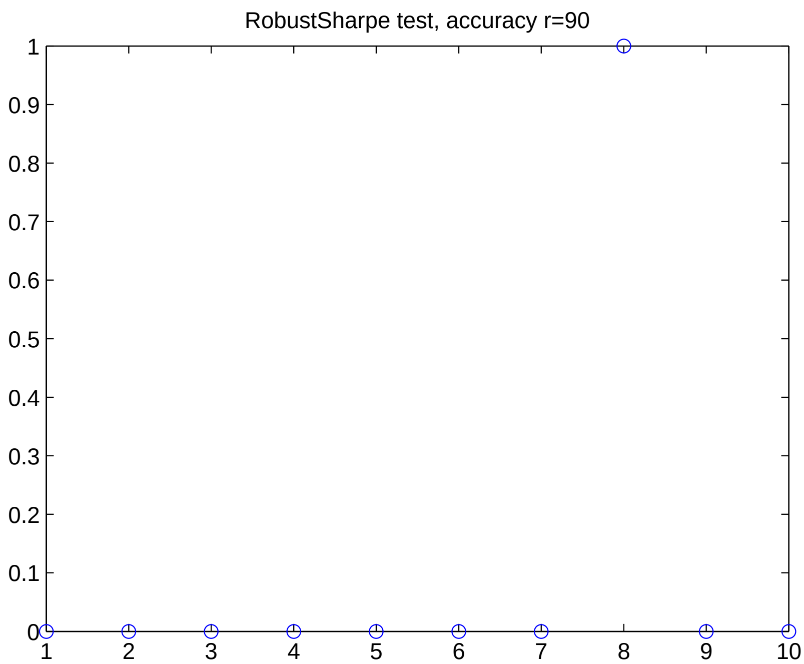

Figure 17: RobustSharpe Test on two simulated normal random vectors

We see that in 90% of our cases the difference of Sharpe Ratios is not statistically significant and 10% the two Sharpe Ratio is statistically significant different. hence from table 1 you can easily compare the Sharpe Ratios of the 8th case where $SR(2) > SR(1)$.

Next let's consider two different distribution where $r1$ is the normal distribution $N(0,1)$ and $r2$ from the uniform distribution $U(0,1)$ as shown in the figure 18, 19 and 20:

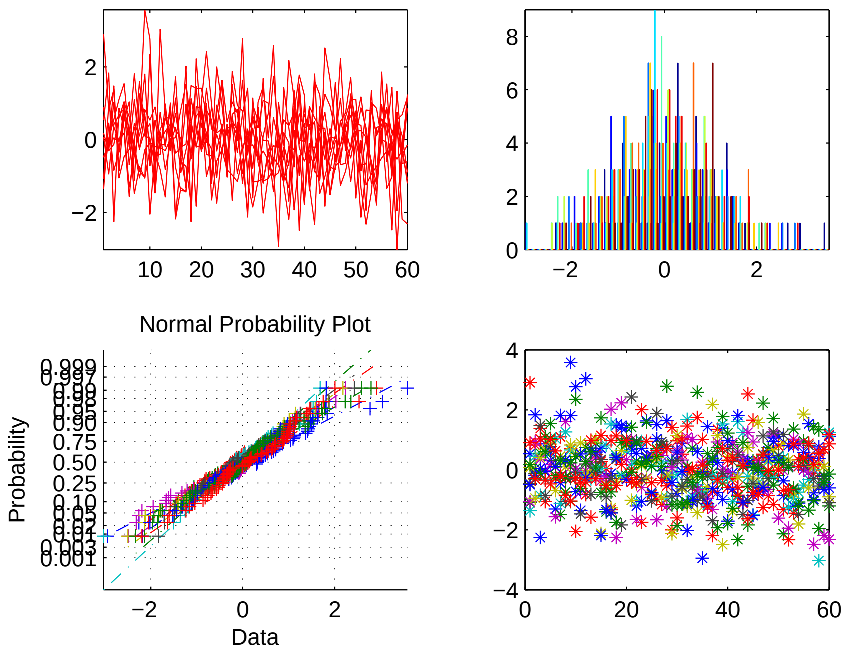

Figure 18: Normal random vectors r1 generated ten times

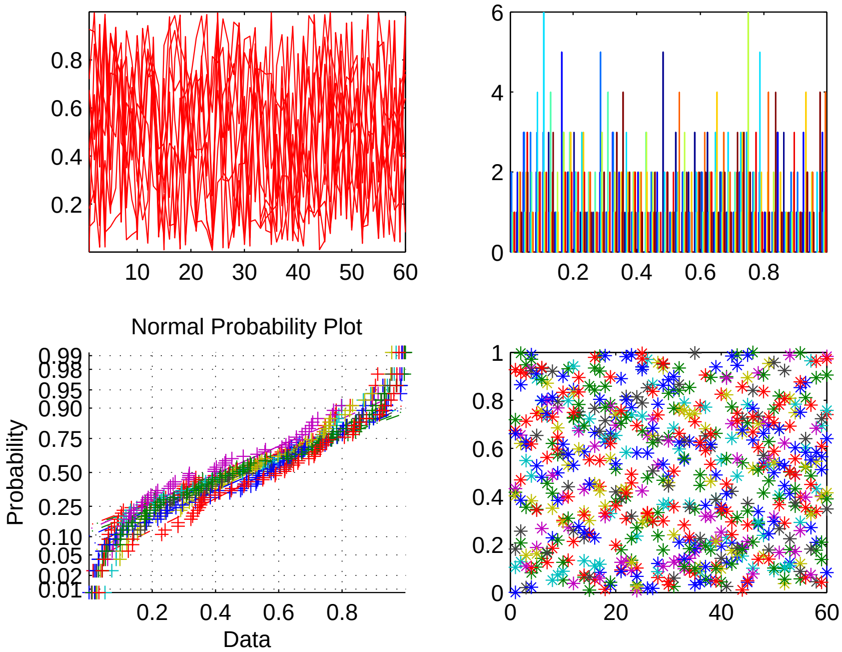

Figure 19: Uniform random vectors r2 generated ten times

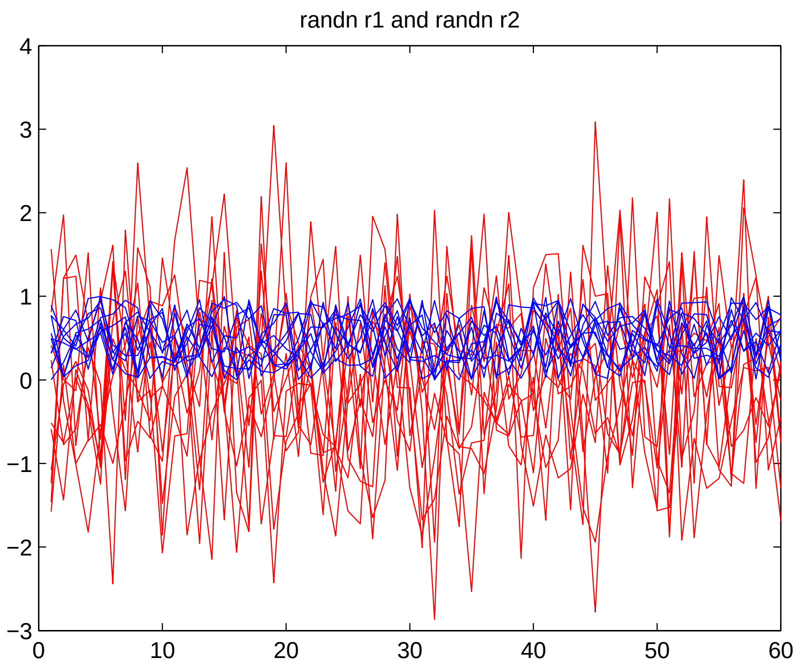

Figure 20: Random vectors r1 and r2 generated ten times

The null hypothesis has been rejected all the 10 times as shown in the table 2:

Table 2: Simulated data from different distribution

	Mean(r1)	**Mean(r2)**	**Std(r1)**	**Std(r2)**	**SR(1)**	**SR(2)**	**Decision H**
1	-0,0065	0,4744	1,0752	0,2871	-0,0060	1,6521	1
2	-0,1035	0,5126	0,9046	0,2985	-0,1144	1,7171	1
3	0,0126	0,5089	0,9498	0,2993	0,0133	1,7005	1
4	-0,1146	0,4964	0,8423	0,2707	-0,1361	1,8336	1
5	0,0081	0,4126	0,9664	0,2603	0,0084	1,5852	1
6	-0,1145	0,4250	1,1147	0,3058	-0,1028	1,3896	1
7	0,0324	0,5234	1,0959	0,2807	0,0295	1,8646	1
8	-0,1159	0,5040	1,0593	0,3026	-0,1095	1,6656	1
9	-0,0864	0,5378	0,9146	0,2831	-0,0945	1,8997	1
10	0,0196	0,5151	1,1258	0,2925	0,0174	1,7610	1

Graphically the null hypothesis is 1 as in figure 21:

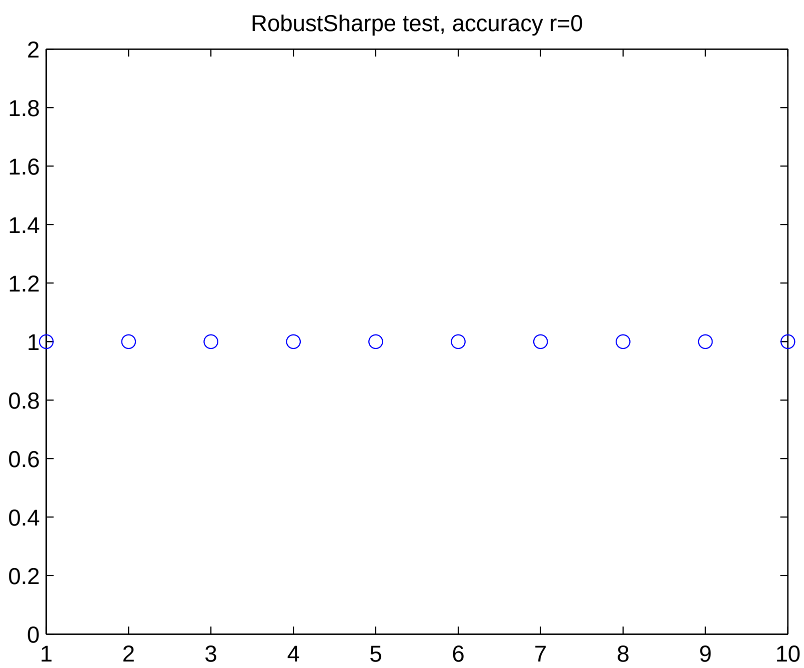

Figure 21: RobustSharpe test on two different simulated distribution

In the above two different kind of distribution we see that the test has detected some statistically significant difference between the two SR all the ten times, hence one can easily make comparison of each two distributions according to their Sharpe Ratios.

Also let's consider two different distribution normal ditribution where $r1$ is the normal distribution $N(0.5,4.7)$ and $r2$ from the normal distribution $N(0.1,9)$ (in order to illustrate the case of two mutual funds data as were presented in the paper of Ledoit and Wolf [18]). Data are presented in figures 22, 23 and 24:

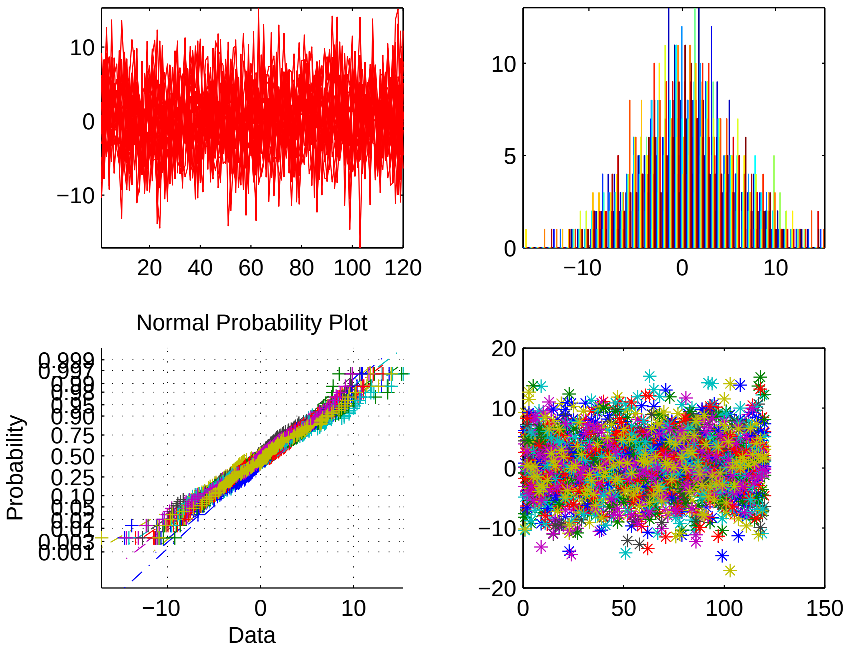

Figure 22: Random vectors $r1 = N(0.5,4.7)$ generated 20 times

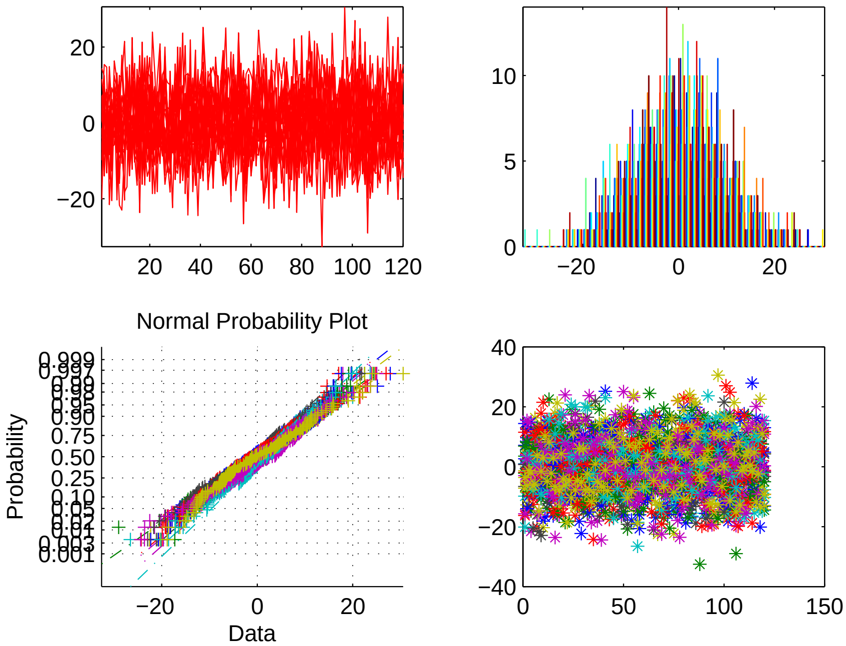

Figure 23: Random vectors $r2 = N(0.1, 9)$ generated 20 times

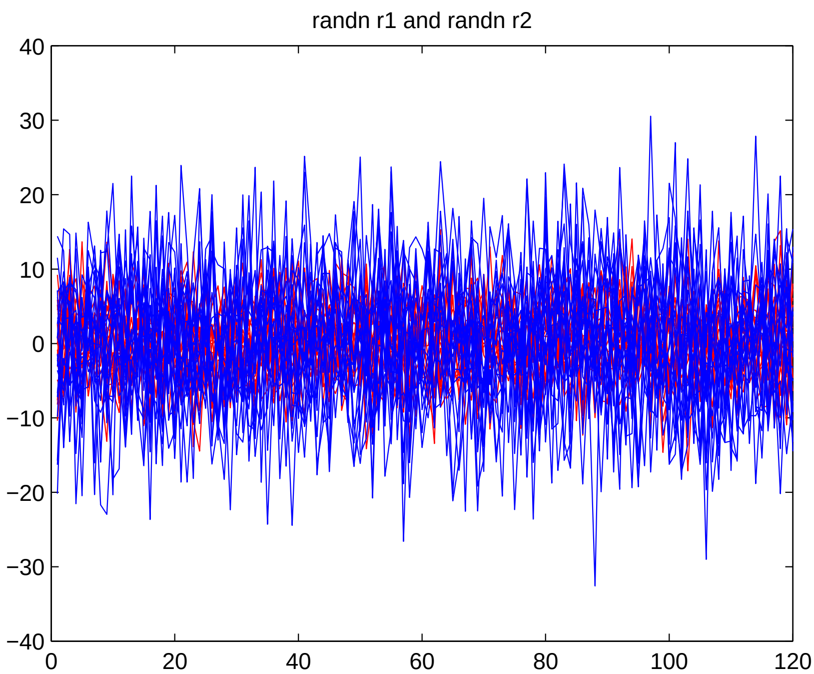

Figure 24: Random vectors r1 and r2 generated 20 times

Computing the Sharpe ratios for the 20 random numbers and their respective testing we report the results in table 4

Table 3: Simulated data from different distributions $N(0.5,4.7)$ and $N(0.1,9)$

	Mean(r1)	**Mean(r2)**	**Std(r1)**	**Std(r2)**	**SR(1)**	**SR(2)**	**Decision H**
1	0.8248	-0.3939	4.708	8.868	0.175	-0.0444	0
2	0.2976	0.7631	4.454	9.302	0.0668	0.0820	0
3	0.4646	-0.2644	4.59051	9.322	0.1012	-0.0284	0
4	0.1052	-0.3803	4.929	7.979	0.0213	-0.0477	0
5	0.3844	1.247	4.446	8.694	0.0864	0.1434	0
6	0.4293	0.3833	4.427	9.362	0.0970	0.0409	0
7	-0.2686	-0.5416	4.397	8.489	-0.0611	-0.0638	0
8	0.8955	-0.8264	4.699	8.802	0.1905	-0.0939	1
9	0.5748	-0.5041	4.598	9.702	0.1250	-0.0520	0
10	0.6368	-0.6939	4.869	8.597	0.1308	-0.0807	0
11	1.216	1.4979	5.057	8.719	0.2405	0.1718	0
12	0.0391	1.1683	4.975	9.138	0.0079	0.1278	0
13	-0.1587	0.3259	5.23061	9.116	-0.0303	0.0358	0
14	0.179	-0.7356	4.977	8.966	0.0360	-0.0820	0
15	0.7695	0.6712	4.543	8.487	0.1694	0.0791	0
16	0.6333	0.3709	5.19021	8.508	0.1220	0.0436	0
17	0.3274	-0.0287	4.593	9.084	0.0713	-0.0032	0
18	0.3182	0.978	5.1306	7.771	0.0620	0.1258	0
19	0.0483	0.3913	4.701	9.236	0.0103	0.0424	0
20	0.6444	0.4519	4.848	9.344	0.1329	0.0484	0

In the table 4 the test showed the non statistical significance difference 95% times.

Also let's consider two different distribution normal ditribution where $r1$ is the normal distribution $N(0,1)$ and $r2$ from the normal distribution $N(10,1)$ to visualize what could happen while having a remarkable difference in mean returns and same volatility as shown in the figure 25, 26 and 27:

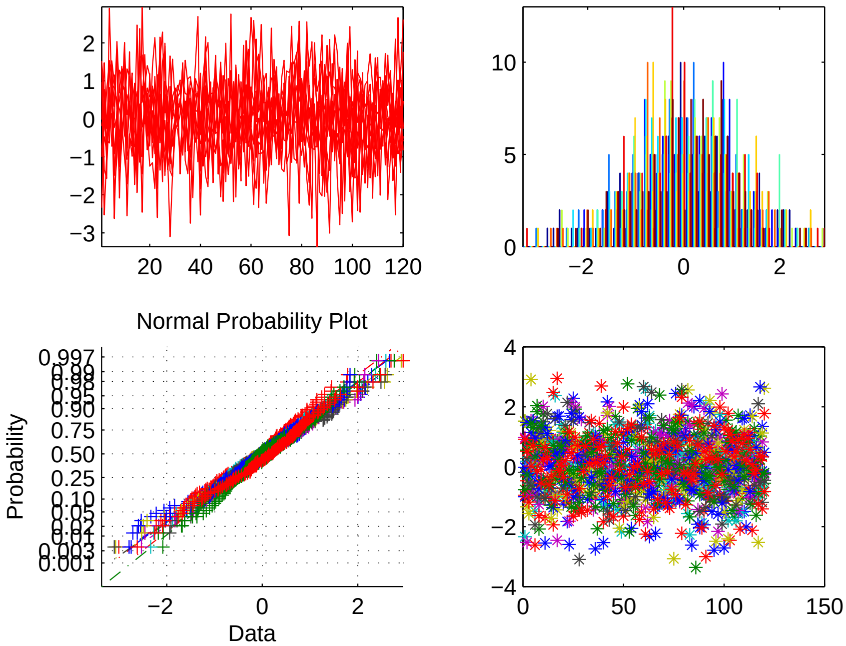

Figure 25: Random vectors $r1 = N(0,1)$ generated 10 times

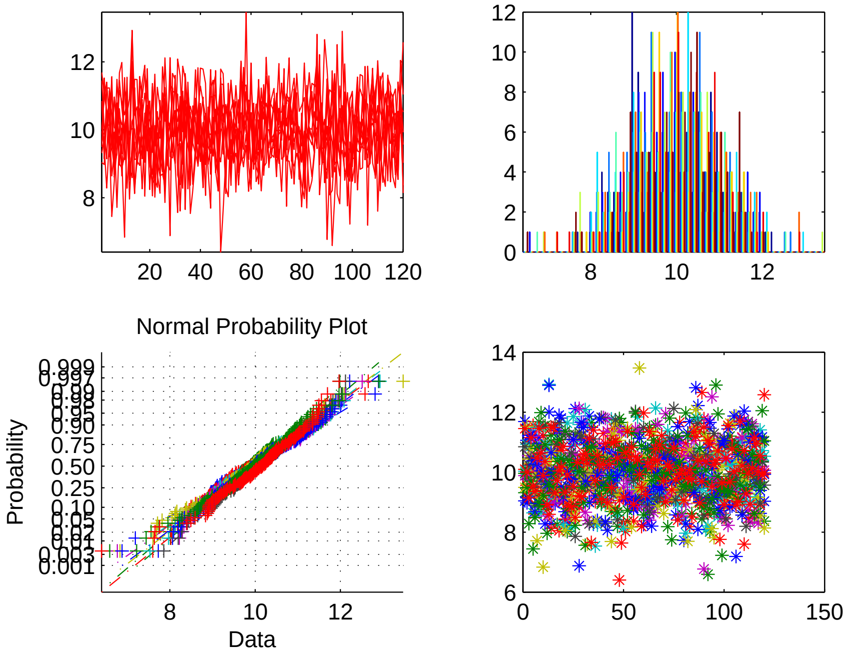

Figure 26: Random vectors $r2 = N(10,1)$ generated 20 times

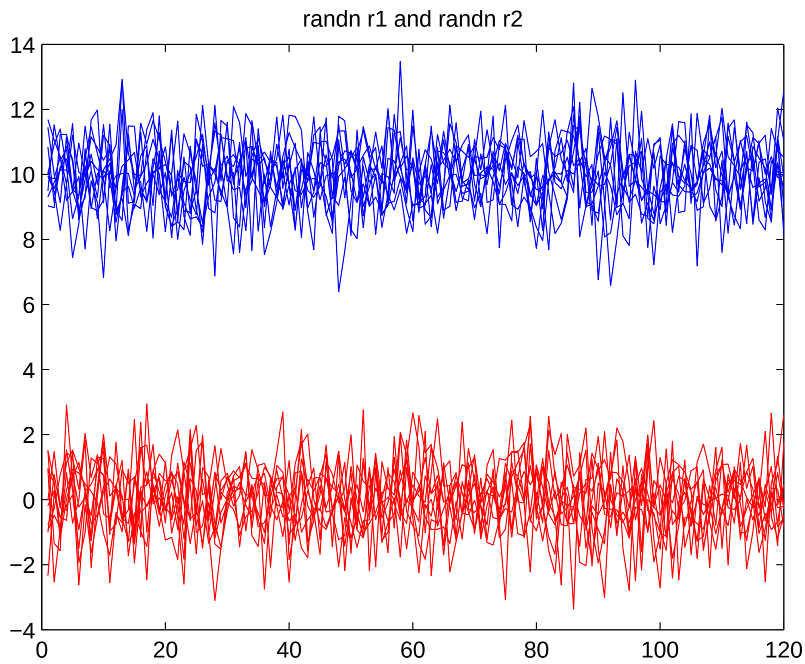

Figure 27: Random vectors r1 and r2 generated 10 times

Testing the significance of difference between the two Sharpe ratios for the 10 random numbers the Hypothesis have been accepted all the times as shown in the figure 28 we see the total existence of statistical significant difference between each SR.

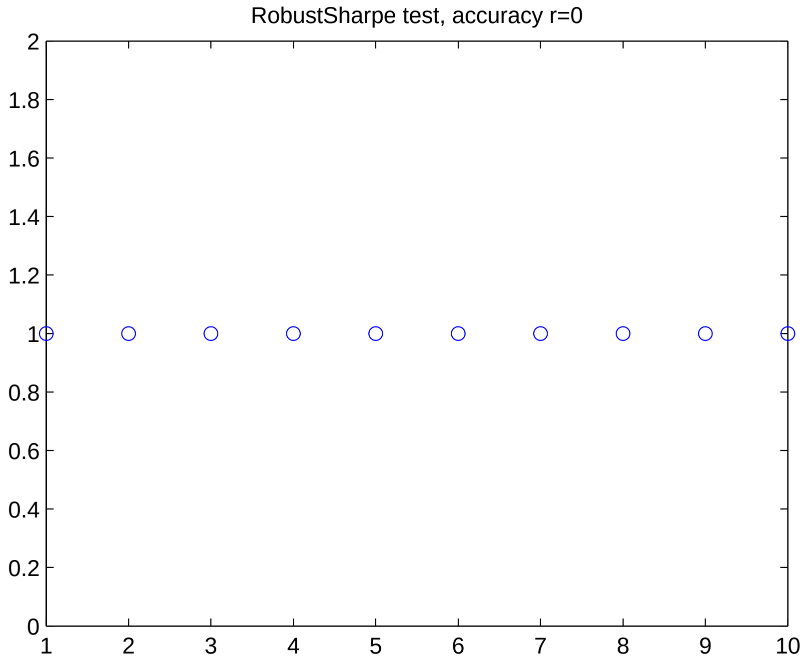

Figure 28: RobustSharpe test accuracy for $N(0,1)$ and $N(10,1)$

One more case which might be an interested area of discussion is to compare Sharpe Ratio of a Gamma distribution and normal distribution which reflect most cases of the Hedge funds time series data; for instance let's compare a gamma distribution $r1 = \gamma(2.5,2)$ and a normal distribution of the same mean and variance $r2 = N(2.5,2)$ that may be represented in the figure 29, 30 and 31:

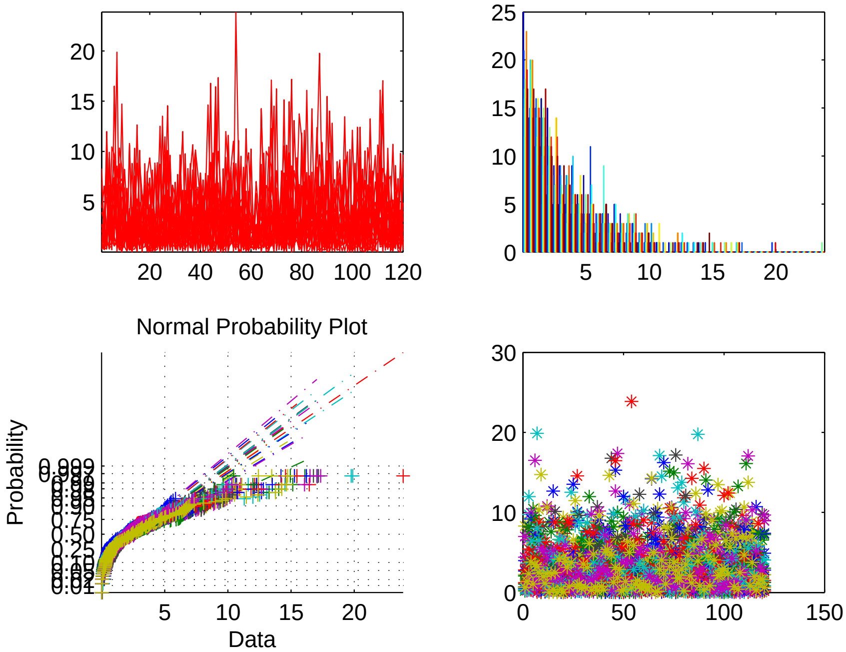

Figure 29: Random vectors $r1 = \gamma(2.5, 2)$ generated 20 times

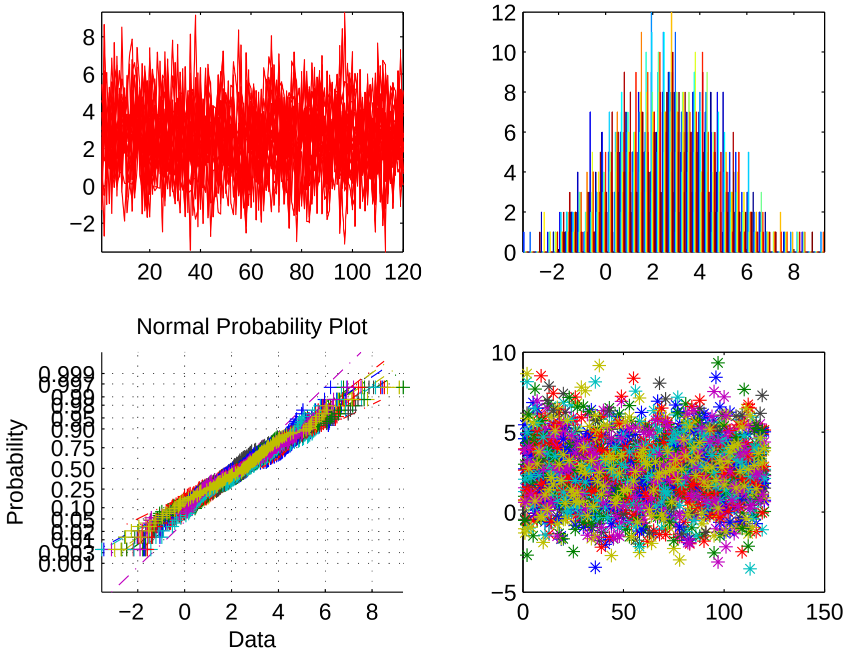

Figure 30: Random vectors $r2 = N(2.5, 2)$ generated 20 times

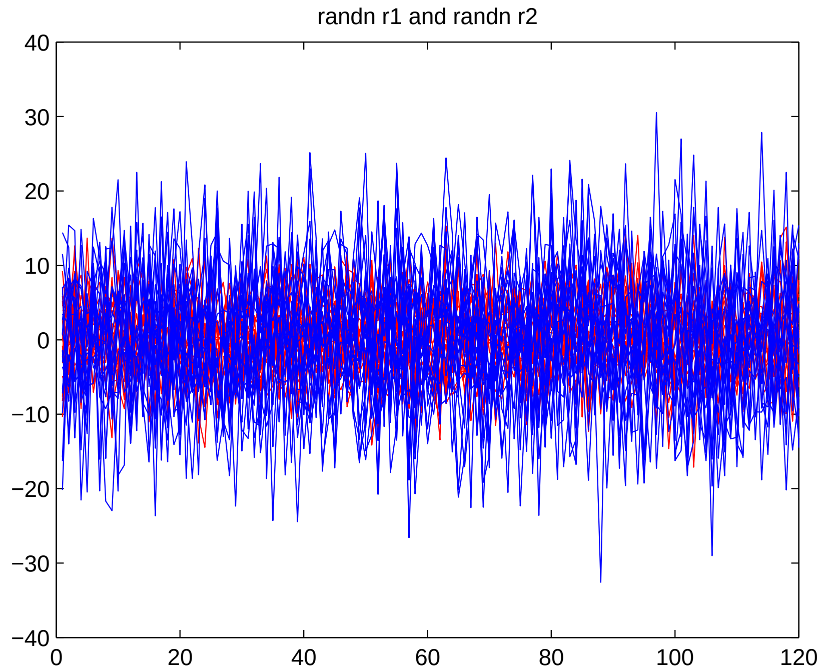

Figure 31: Comparison of r1 and r2 generated 20 times

While testing the two distribution with the RobustSharpe test 20 times, we report in table 4 the results including the mean, the standard deviation, the Sharpe Ratio and Hypothesis test decision in each case.

Table 4: Simulated data from distribution $r1 = \gamma(2.5, 2)$ and $r2 = N(2.5, 2)$

	Mean(r1)	**Mean(r2)**	**Std(r1)**	**Std(r2)**	**SR(1)**	**SR(2)**	**Decision H**
1	2.586	2.747	2.391	2.165	1.081	1.268	0
2	2.946	2.458	2.752	2.042	1.0708	1.203	0
3	2.892	2.377	2.753	2.187	1.0502	1.086	0
4	**3.145**	**2.638**	**3.442**	**2.079**	**0.914**	**1.269**	**1**
5	3.084	2.456	2.986	1.862	1.032	1.318	0
6	**3.087**	**2.747**	**2.964**	**2.002**	**1.041**	**1.371**	**1**
7	2.761	2.159	2.566	1.935	1.075	1.116	0
8	**3.219**	**2.541**	**3.213**	**1.883**	**1.002**	**1.349**	**1**
9	3.479	2.352	3.351	1.813	1.038	1.313	0
10	3.133	2.631	3.288	2.003	0.953	1.348	0
11	**2.871**	**2.594**	**2.949**	**1.923**	**0.973**	**1.296**	**1**
12	3.413	2.428	3.362	1.877	1.015	1.293	0
13	3.025	2.436	2.844	2.0703	1.063	1.176	0
14	**3.348**	**2.699**	**3.186**	**1.995**	**1.0508**	**1.352**	**1**
15	3.027	2.221	3.085	1.831	0.9294	1.212	1
16	2.964	2.599	2.871	2.103	1.0541	1.235	0
17	**2.868**	**2.535**	**2.995**	**1.845**	**0.9896**	**1.374**	**1**
18	2.829	2.606	2.819	1.789	1.0036	1.456	0
19	2.799	2.398	2.701	2.019	1.0364	1.187	0
20	3.321	2.361	3.305	2.049	1.0043	1.152	0

Graphically we see in figure 32 the accuracy of test along the twenty times of testing the same distribution of the means and variance shown in the table 4

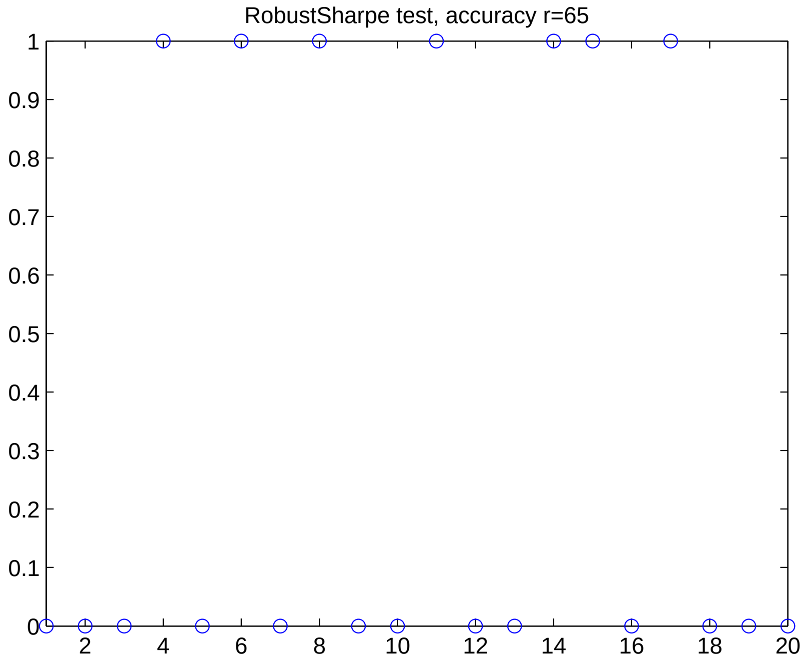

Figure 32: RobustSharpe test accuracy for $\gamma(2.5,2)$ and $N(2.5,2)$

In the above case we have not good accurate of the RobustSharpe test (65%). Therefore we can deduce that while testing the statistical significance performance of two Sharpe Ratio in such case the RobustSharpe Test is critical.

4.2.2 Hedge and Mutual funds results

Hedge funds

While testing the Sharpe ratio of the hedge funds data fifty times the hypothesis has been rejected fifty times. recalling that while calling this function in Matlab all input were taken by default except the block size which varied randomly along fifty testing and equal 1, 3, 6, 10, 15.

The two Hedge fund data can be presented in the plots as shown in the figure 33, 34 and 35:

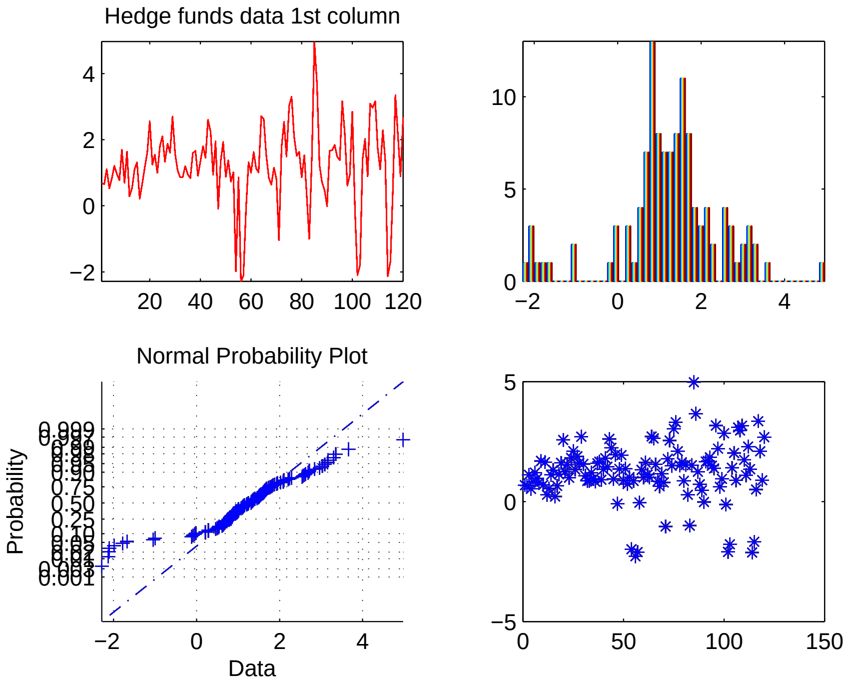

Figure 33: The coast Enhanced Income fund data

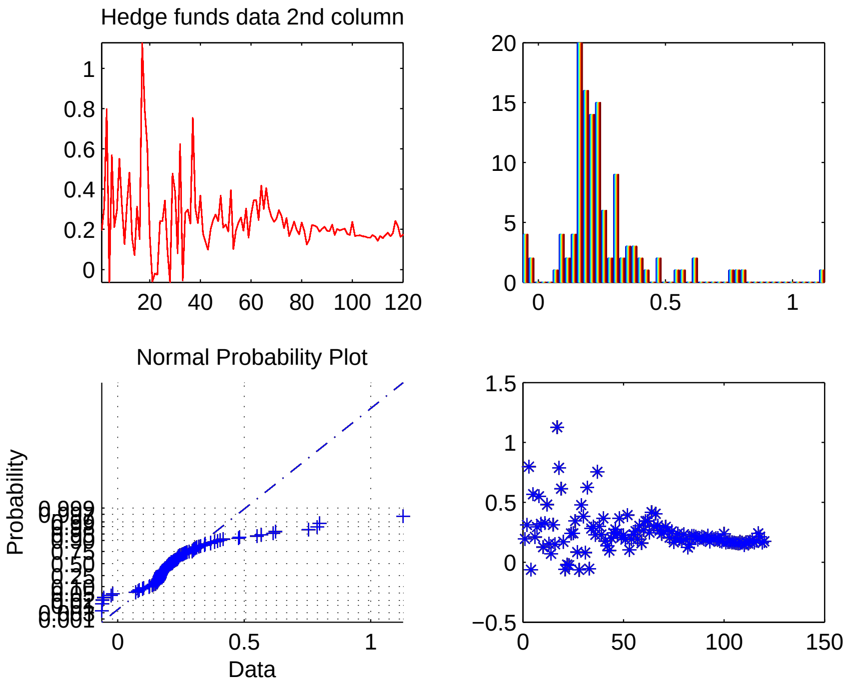

Figure 34: The JMG capital partners fund data

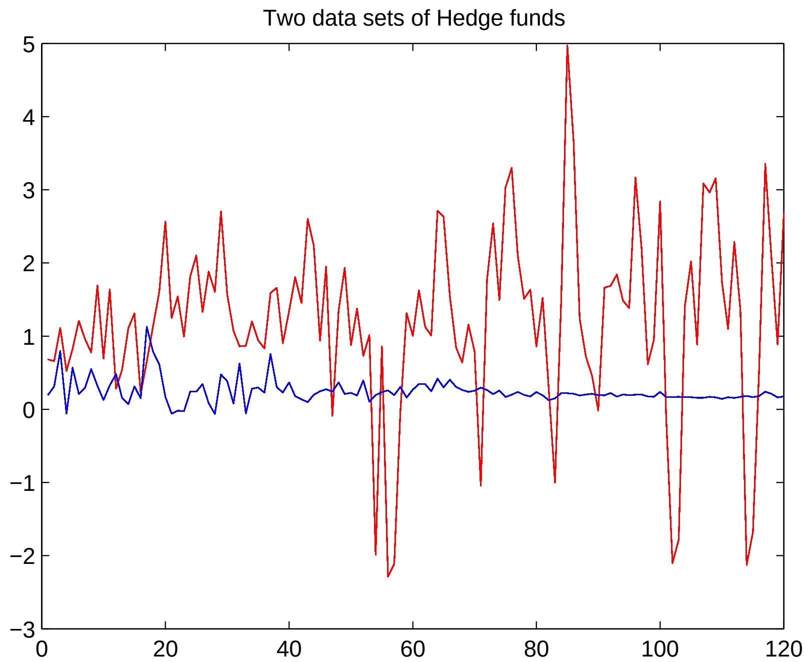

Figure 35: Comparison of the 2 Hedge funds data

The null hypothesis were not rejected all the fifty times as shown hereunder in figure 36 recalling that:

$H_0 = 0$ means that there is no significant performance difference between the two Sharpe Ratio being compared.

$H_0 = 1$ means that statistical significance perfomance difference between the two Sharpe Ratio exist.

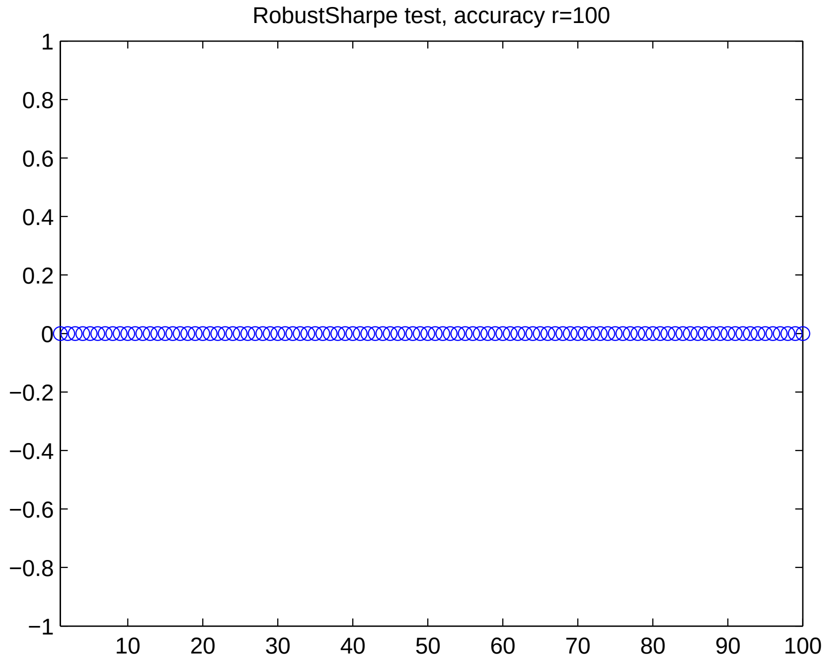

Figure 36: RobustSharpe Test on Hedge funds fifty times

The decision of the non rejection of the null hypothesis were based on the P-values calculation with studentized bootstrap methods and the this values were close to the one found in Ledoit-Wolf paper [18] equal to 0.294, our P-values obtained all the fifty times are shown in the figure 42 recalling that decision rule is to reject the null hypothesis if $Pvalue > \alpha$ with $\alpha = 0.05$ in our case.

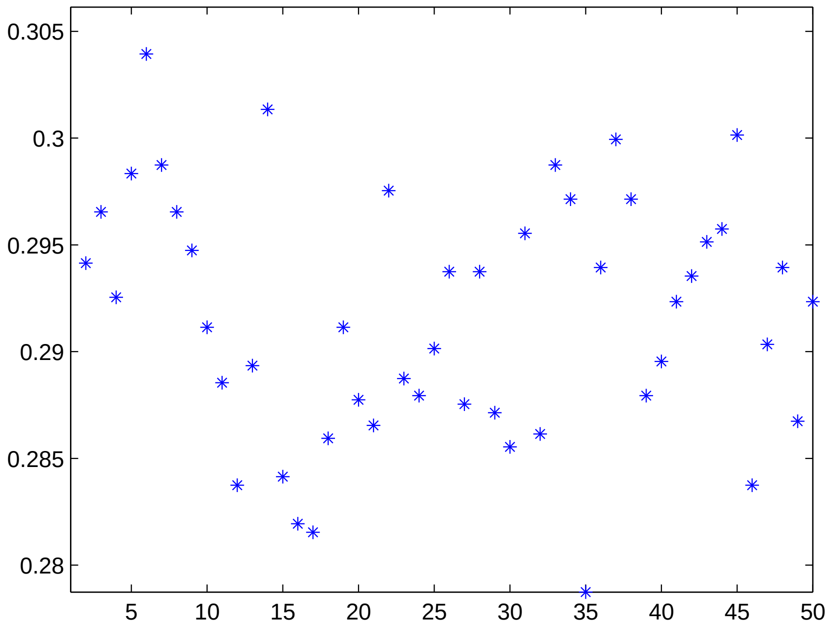

Figure 37: All the P-values are between 0.27 and 0.305 hence the rejection of H_0 at 0.05 of significance level

We performed the computation of the Sharpe ration being compared and we got results in the table 5 where Mean stands for the average return μ of each hedge funds, Std stands for the standard deviation or volatility σ of each hedge funds, SR stands for the Sharpe Ratio being compared of the two Hedge funds and it has be calculated as $SR = \frac{\mu}{\sigma}$ as defined in [2] while Zero investment Strategy (defined in Appendix).

Table 5: Hedge Funds data results

	Hedge1	**Hedge2**
Mean	1,228	0,2448
Std	1,2107	0,1676
SR	1,0142	1,4605

Basing on the test of significance difference performance between two Sharpe Ratio done we can deduce that the difference of two Sharpe Ratios is statistically insignificant hence, the first hedge fund, Coast Enhanced Income is statistically equal to the second hegde fund, The JMG capital Partners.

Mutual funds

The same manner we had a look on mutual funds and presented the mutual funds data as shown here under in figures 38, 39, 40:

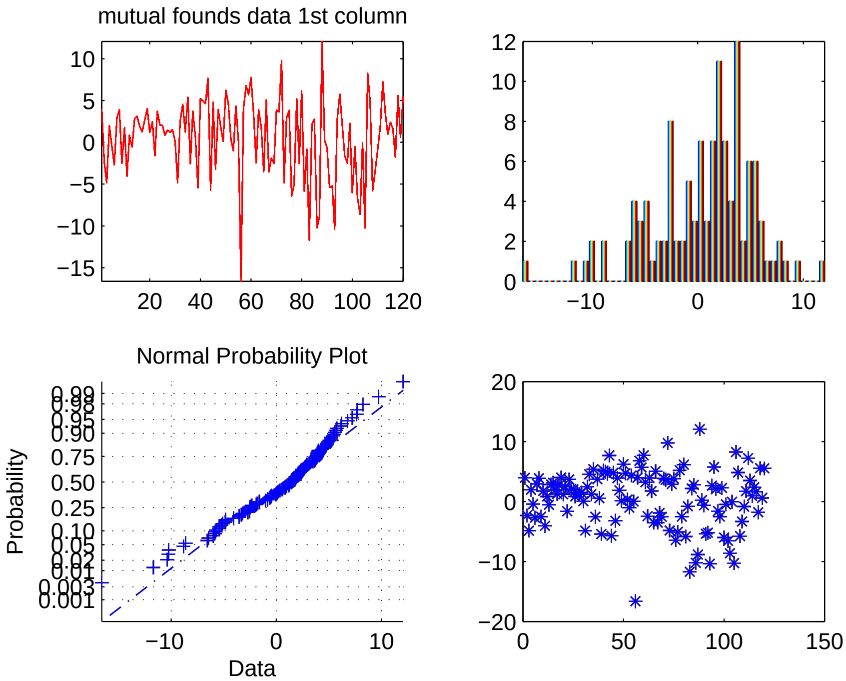

Figure 38: The fidelity fund data

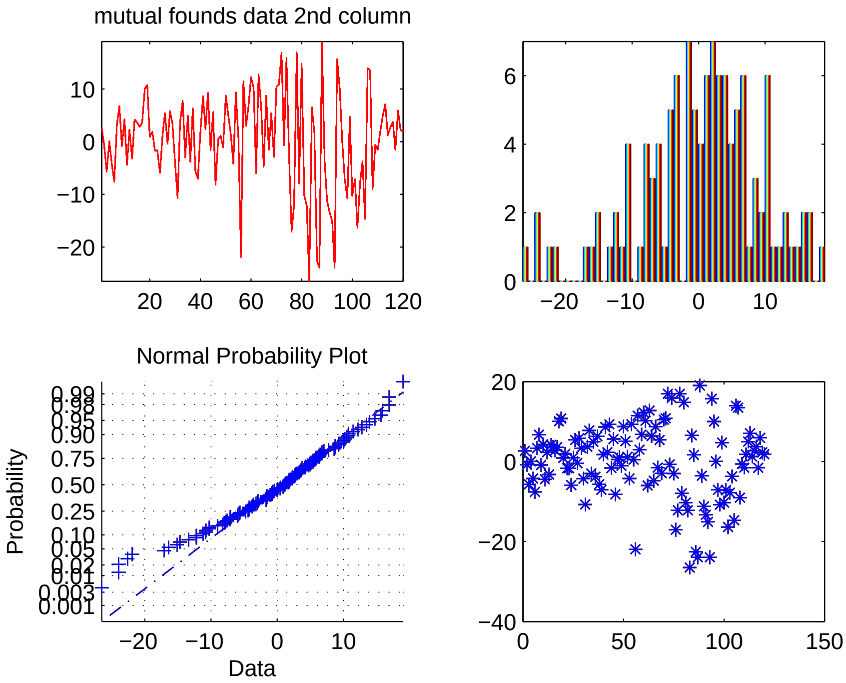

Figure 39: The fidelity Aggressive Growth fund data

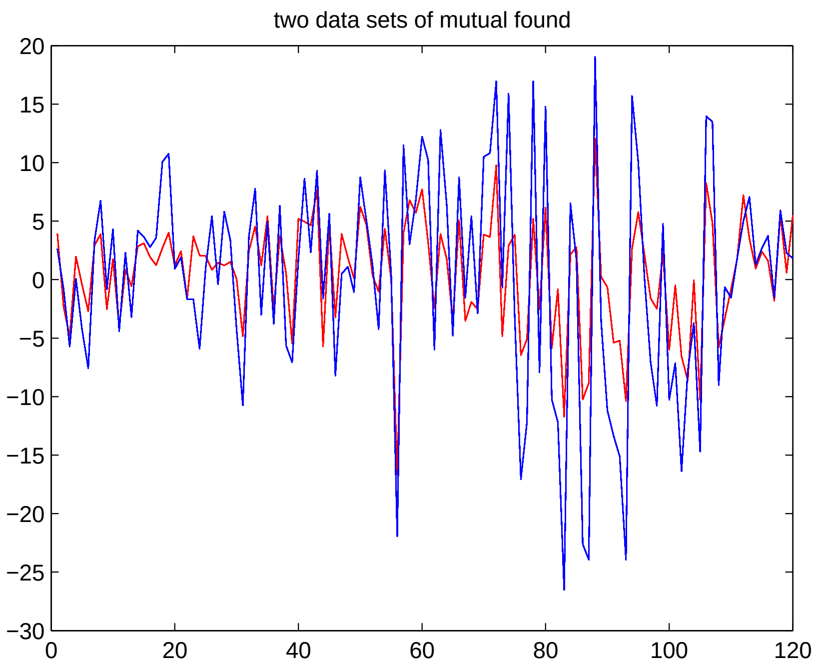

Figure 40: Comparison of two mutual fund data

Also we have got the null hypothesis rejected all the fifty times as shown in Figure **??**.

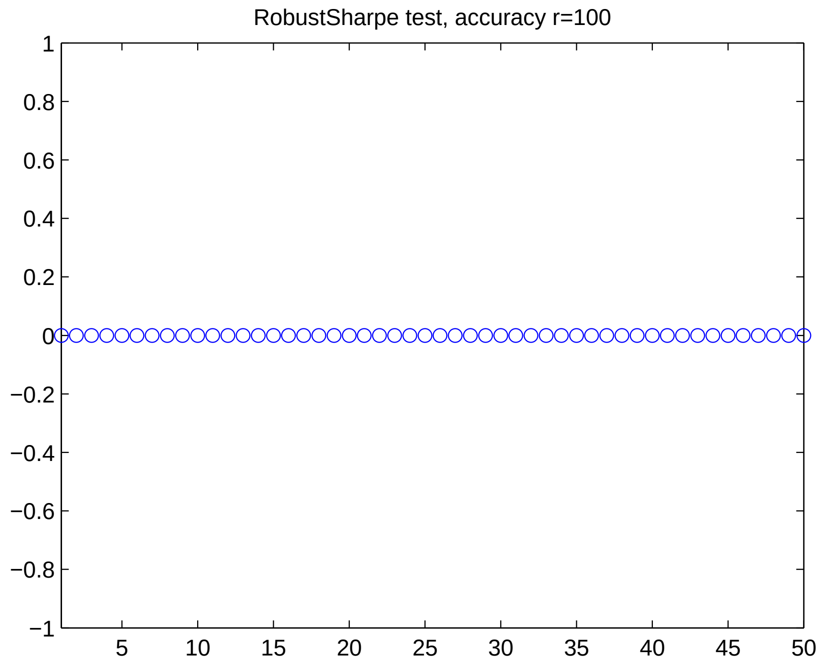

Figure 41: RobustSharpe Test on Mutual funds fifty times

The above non rejection of the null hypothesis decision were based on the P-values also close to the one found in Ledoit-Wolf paper [18] equal to 0.092, our P-values obtained all the fifty times are shown in the figure 42 and all P-values are greater than the $\alpha = 0.05$ hence the non rejection of H_0 since the decision rule is to reject the null hypothesis if $Pvalue > \alpha$.

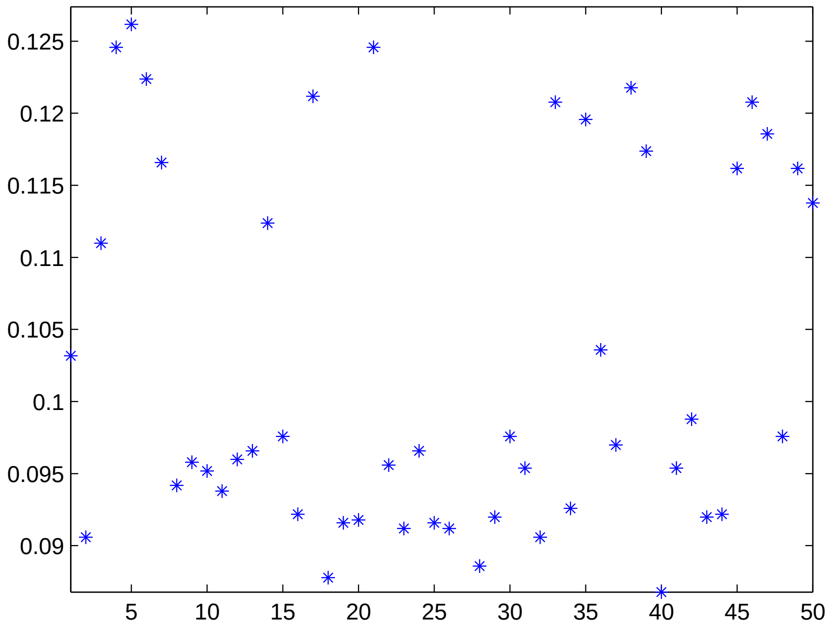

Figure 42: All the P-values are between 0.08 and 0.13 hence the rejection of H_0 at 0.05 of significance level

The Sharpe ratio of the two mutual funds has been computed and presented in table 6:

Table 6: Mutual Fund data results

	Mutual1	**Mutual2**
Mean	0,5115	0,0985
Std	4,7597	9,1606
SR	0,1075	0,0107

According to the results of the RobustSharpe test we do not reject the null hypothesis that the difference in the Sharpe ratio is statistically significant and the first mutual fund, the Coast Enhanced Income is statistically equal the second mutual fund, the JMG Capital Partners.

5 CONCLUSION

This master thesis was aimed to gain a practical and theoretical understanding about measuring the ranks investment strategies of a portfolios by the Sharpe ratio while applying Matlab as a computation software tool. To this, we have studied the work of Ledoit and Wolf including the Matlab codes created by his team able to test statistically significant difference between two Sharpe ratios.

The Programming code were using a studentised bootstrapping methods and we showed that bootstrapping method while testing or estimating some statistics should always considered because at most cases for the outcome are almost always accurate and errors are minimized in the computation.

In this work we narrowed our objectives and studied only few cases of performance measurement. We recommend in future works the use of other measures like the Treynor's measure, Jensen's measure or Appraisal ratio as discussed in [2] and [14] Also the RobustSharpe Test should be emphasized on the portfolios that have negative Excess Returns and portfolios that have Excess Returns of different signs.

References

[1] Chris Brooks, Introductory Econometrics for Finance, second edition, Cambridge University Press, New York, USA, 2008

[2] Eero Pätäri, Essay on Portfolio Perfomance Measurement, Thesis of the degree of Doctor of Science at LUT, December 2000

[3] Erwin Kreyszig, Advanced Engineering Mathematics, 8th edition, Phoenix Color Corp., 1999

[4] Hypothesis testing by Matlab, available at http://www.mathworks.se/help/toolbox/stats/f15360dfi3.html

[5] Israelsen, C.L., Sharpening the Sharpe ratio. Financial Planning 33, 2003.

[6] Israelsen, C.L., A refinement to the Sharpe ratio and Information ratio. Journal of Asset Management 5, 2005.

[7] Jeremy M. Shaver, Barry M. Wise, Using Pre-Whitening Techniques for techniques for process model stabilization, Eigenvector Research, Inc. Seattle/Wenatchee, WA, 2007.

[8] Joaquim P. Marques, Applied Statistics Using SPSS, STATISTICA, MATLAB and R, second edition, Springer-Verlag Berlin Heidelberg, 2007

[9] Jobson, J.D., Korkie, B.M., Performance hypothesis testing with the Sharpe and Treynor measures. Journal of Finance 36, 1981.

[10] Hoang Pham, Springer Handbook of Engineering Satistics, Springer-Verlag London Limited, New Jersey, USA, 2006.

[11] Leonard J. Kazmier, Schaum's outline series theory and problems of Business Statistics with computer applications, 2nd edition, McGraw-Hill, Inc., USA, 1988.

[12] Lo, A.W., 2002. The statistics of Sharpe ratios. Financial Analysts Journal 58 (4), 36–52.

[13] Mario F. Triola, elementary statistics, 11th edition, Addison-Wesley, New York, USA, 2008

[14] Mary Jackson and Mike Staunton, Advanced Modelling in Finance using Excel and VBA, John Wiley and Sons, Ltd, England, 2001

[15] Memmel, C., Performance hypothesis testing with the Sharpe ratio. Finance Letters 1, 2003.

[16] Mohammad Shahid, Measurement portfolio Performance, Master thesis, university of Uppasala, june 2007. see at http://www2.math.uu.se/research/pub/Shahid1.pdf

[17] Murray R.Spiegel., John Shiller, and R. Alu Srinivasan, Schaum's Easy Outlines of Probability and Statistics, The McGraw-Hill Companies, Inc., USA, 2001.

[18] Oliver Ledoit and Michael Wolf, Robust performance hypothesis testing with the Sharpe ratio, Journal of Empirical Finance, March 2008

[19] Phillip Good, Permitation, Parametric and Bootstrap Tests of Hypotheses, 3rd edition, Springer Science+Business Media, Inc., New York, NY 10013, USA, 2005.

[20] Programing code, University of Zurich, link: http://www.econ.uzh.ch/faculty/wolf/publications.html

[21] William F. Sharpe, The Sharpe Ratio, the journal of Portfolio Management, Stanfold University, Fall 1994, see at http://www.stanford.edu/ wfsharpe/art/sr/sr.htm

[22] Sharpe William F., Mutual Fund Performance, The Journal of Business, Vol. 39, No. 1, Part 2: Supplement on Security Prices, Jan 1966.

[23] Wendy L. Martinez Angel R. Martinez, Computational Statistics Handbook with MATLAB, 2002, USA

[24] William F. Sharpe, Morningstar's Perfomance measures, available at http://www.stanford.edu/ wfsharpe/art/stars/stars0.ht

[25] Vinod,H.D.,Morey,M.R., Confidence intervals and hypothesis testing for the Sharpe and Treynor performancemeasures: a bootstrap approach. In: Abu-Mostafa, Y.S., LeBaron, B., Lo, A., Weigend, A.S. (Eds.), Computational Finance 1999. The MIT Press, Cambridge, 1999.

Appendices

A Some Definitions

Mutual fund: is a type professionally managed collective investment plans that collect money from many investors to buy securities a financial value.[wiki]

Hedge fund: is an investment find that can undertake a wider range of investment and trading activities than other funds, but which is only open for investment from particular types if investors specified by regulators.[wiki]

Return:is the The gain or loss of a security in a particular period. The return consists of the income and the capital gains relative on an investment. It is usually quoted as a percentage.

The general rule is that the more risk you take, the greater the potential for higher return - and loss. [investopedia]

Volatility:1. A statistical measure of the dispersion of returns for a given security or market index. Volatility can either be measured by using the standard deviation or variance between returns from that same security or market index. Commonly, the higher the volatility, the riskier the security.

2. A variable in option pricing formulas showing the extent to which the return of the underlying asset will fluctuate between now and the option's expiration. Volatility, as expressed as a percentage coefficient within option-pricing formulas, arises from daily trading activities. How volatility is measured will affect the value of the coefficient used. [investopedia]

Zero investment Portfolio: A group of investment which, when combined, create a zero net value, the zero investment strategy can be achieved by simultaneously purchasing securities and selling equivalent securities; thus achieving a lower risk or gain compared to only purchasing or selling the same securities. This application has many advantages like

- Reduce taxes, Because of little or no interest income
- Reduce risk by protecting against unexpected shifts in the value of the held securities.
- Protect the overall value of portfolio so that investment can be made at a later data

- Determine if the average portfolio returns are statistically different from zero.

Prewhitening: A signal is "white" if each sample is independent statistically from every other sample, once you knew the data up until sample n-1, you would have no information about sample n.

A pre-whitening filter takes a signal that is not white and produces a white signal. If you subtract this prediction from the actual sample n, you will be left with the portion of sample n that is not related to the rest of the samples.

To explain where the stochastic gradient algorithm comes into play, here is a quick explanation of the system used to whiten the signal.

Say you're original signal is x[n] (discrete samples) Delay signal by single samples to get x[n-1] Run signal through whitening filter to get x'[n] (estimate of x[n]) Subtract x'[n] from x[n] to get e[n] (error signal) Feed error signal to whitening filter to update filter coefficients using stochastic gradient algorithm (also look under LMS filter) The signal e[n] is your whitened signal. If the system is perfect, the predictor error will be a white signal, as that is the best performance the predictor can hope to have. Depending on how correlated the samples are in your original x[n] are, you may need a very large filter to whiten the signal. The usefulness of whitening data is in performing data compression.

For example, digital cellular phones do not transmit your digitized voice over the air. Instead, they whiten the voice signal, and transmit the whitened signal power (or impulse period) and the filter coefficients.

shortly, Prewhitening aims to make the signal contain equal-strength components at every possible frequency. The assumption is that all frequencies present in the signal contain some information that is useful. If the signal is NOT white, then some information at some frequencies might swamp other, equally valid, useful information.

Treasury Bill or T-Bill is A short-term debt obligation backed by the U.S. government with a maturity of less than one year. T-bills are sold in denominations of USD 1,000 up to a maximum purchase of USD 5 million and commonly have maturities of one month (four weeks), three months (13 weeks) or six months (26 weeks).

T-bills are issued through a competitive bidding process at a discount from par, which means that rather than paying fixed interest payments like conventional bonds, the appreciation of the bond provides the return to the holder. [investopedia]

B Mutual and Hedge Funds Data

Table 7: Mutual and Hedge Funds time series data

	Mutual Fund data		**Hedge Fund data**	
	Mutual1	**Mutual2**	**Hedge1**	**Hedge2**
1	3,9267	2,6476	0,6813	0,1944
2	-2,3305	-0,8163	0,6591	0,3118
3	-4,8583	-5,724	1,1132	0,7974
4	1,96	0,0475	0,5253	-0,0614
5	-0,4397	-4,2416	0,8251	0,5679
6	-2,7056	-7,5912	1,2098	0,2104
7	2,9257	3,3391	0,9581	0,2947
8	3,8781	6,7505	0,778	0,5506
9	-2,5249	-0,8431	1,6924	0,3212
10	1,7459	4,2813	0,6913	0,126
11	-4,0699	-4,4142	1,6391	0,3275
12	0,8401	2,2952	0,2833	0,4815
13	-0,5617	-3,2121	0,5422	0,1553
14	2,8434	4,1845	1,1103	0,0711
15	3,1108	3,6622	1,3098	0,3127
16	1,9165	2,7866	0,2097	0,1502
17	1,2407	3,5773	0,6543	1,1278
18	2,7006	10,063	1,1338	0,7886
19	4,0165	10,753	1,6093	0,6146
20	1,1931	0,9062	2,566	0,1699
21	2,4422	1,9005	1,2471	-0,0595
22	-1,615	-1,6643	1,5419	-0,0196
23	3,6948	-1,6861	0,9945	-0,0259
24	2,0702	-5,9162	1,8171	0,2401
25	2,0209	0,9967	2,1019	0,2413
26	0,8348	5,4183	1,3318	0,3442
27	1,4459	-0,4034	1,8805	0,0854
28	1,2052	5,8189	1,6073	-0,0632
29	1,515	3,3712	2,703	0,4779
30	0,0613	-4,2658	1,5664	0,3831
31	-4,8483	-10,753	1,0703	0,0802
32	2,4498	3,6975	0,8612	0,6241
33	4,5504	7,7571	0,868	-0,0543
34	1,2171	-3,0071	1,2012	0,2818
35	5,407	4,9361	0,942	0,2985

	Mutual Found data		Hedge Found data	
	Mutual1	**Mutual2**	**Hedge1**	**Hedge2**
36	-2,5487	-3,7872	0,8329	0,2286
37	3,7254	6,3227	1,5907	0,7542
38	0,5411	-5,6427	1,661	0,3001
39	-5,4384	-7,0698	0,9036	0,2302
40	5,1897	1,749	1,3449	0,3676
41	4,956	8,6352	1,8047	0,1777
42	4,6393	2,3316	1,4517	0,1375
43	7,6587	9,2936	2,6035	0,0981
44	-5,7242	-1,6182	2,2368	0,2004
45	4,8064	5,6314	0,9382	0,2453
46	-3,2255	-8,205	1,948	0,2735
47	3,9052	0,5	-0,0883	0,2402
48	1,8622	1,0918	1,3549	0,3676
49	0,1058	-1,0718	1,9325	0,2083
50	6,2328	8,7543	0,8779	0,2243
51	4,7072	5,1108	1,375	0,1893
52	0,3427	1,2007	0,731	0,3942
53	-1,0273	-4,224	1,0133	0,1021
54	4,3375	9,339	-1,9875	0,1932
55	0,0311	0,4632	0,8585	0,2345
56	-16,625	-21,925	-2,2863	0,2592
57	3,9789	11,495	-2,1105	0,1928
58	6,7855	2,9996	-0,0441	0,3044
59	5,6955	6,7845	1,3163	0,159
60	7,7121	12,225	1,0026	0,27
61	3,2071	10,22	1,6263	0,3435
62	-2,4622	-5,9869	1,1269	0,3455
63	3,8969	12,781	1,0086	0,2461
64	1,7631	6,3812	2,7126	0,417
65	-3,5363	-4,7808	2,6278	0,301
66	5,0614	8,7382	1,5492	0,4046
67	-3,5352	-1,5323	0,8418	0,3069
68	-1,9028	5,4122	0,6403	0,2636
69	-2,5466	-2,8854	1,1569	0,237
70	3,8562	10,479	0,7975	0,2524
71	3,6341	10,826	-1,044	0,2954
72	9,7548	16,936	1,7733	0,2652
73	-4,8312	-0,7	2,5434	0,206
74	2,9038	15,91	1,491	0,2574
75	3,824	-2,9817	3,0273	0,1668
76	-6,4514	-17,061	3,2998	0,1991
77	-5,0871	-12,186	2,0978	0,2385
78	5,1907	16,948	1,5092	0,1967
79	-2,5309	-7,9302	1,635	0,1754
80	6,1504	14,78	0,8581	0,2347

	Mutual Found data		**Hedge Found data**	
	Mutual1	**Mutual2**	**Hedge1**	**Hedge2**
81	-5,8471	-10,254	1,5241	0,1921
82	-0,8205	-12,191	0,2828	0,1239
83	-11,711	-26,517	-1,0001	0,149
84	2,1565	6,5252	1,5128	0,2202
85	2,7675	1,6718	4,9735	0,2195
86	-10,242	-22,616	3,6596	0,2111
87	-8,8416	-23,94	1,2367	0,188
88	12,067	19,028	0,7182	0,2022
89	0,2347	-3,5765	0,4607	0,2124
90	-0,6507	-11,193	-0,0162	0,193
91	-5,3993	-13,367	1,6634	0,1913
92	-5,2209	-15,11	1,6853	0,2233
93	-10,378	-23,946	1,8415	0,1721
94	2,6288	15,686	1,4879	0,2012
95	5,7568	10,007	1,3837	0,195
96	2,072	0,0186	3,1651	0,1998
97	-1,6054	-7,0658	2,206	0,203
98	-2,5013	-10,805	0,6142	0,1766
99	2,2475	4,7581	0,9462	0,1716
100	-6,006	-10,273	2,8428	0,2372
101	-0,4965	-7,1369	-0,123	0,1666
102	-6,5602	-16,376	-2,0995	0,1691
103	-8,6304	-7,7748	-1,7822	0,1707
104	-0,0426	-3,7139	1,4042	0,1656
105	-10,261	-14,697	2,0218	0,1648
106	8,2459	13,963	0,8841	0,1589
107	4,8368	13,486	3,0868	0,1578
108	-5,7832	-9,0086	2,9641	0,1711
109	-3,3307	-0,636	3,1595	0,1627
110	-0,8246	-1,5489	1,7362	0,1428
111	1,7395	1,8077	1,0992	0,166
112	7,2318	4,9774	2,2878	0,156
113	3,4893	7,066	1,3509	0,171
114	0,9472	1,2639	-2,1272	0,1833
115	2,3977	2,7054	-1,6775	0,165
116	1,604	3,7319	0,4995	0,1809
117	-1,8021	-1,5551	3,3526	0,2415
118	5,559	5,9161	2,09	0,2133
119	0,586	2,2775	0,8882	0,1626
120	5,5201	1,8237	2,6868	0,175
Mean	**0,511455**	**0,098456**	**1,22797**	**0,24482**
Std	**4,759708**	**9,160556**	**1,21074**	**0,16762**
SR	**0,107455**	**0,010748**	**1,01423**	**1,46055**