02403 Introduction to Mathematical Statistics

Project – Trading with ETF

Mark Andrawes s214654

(s214654@student.dtu.dk)

Number of characters: 16673

Introduction

An exchange traded fund (ETF) is a blend of mutual funds and shares. An example of an ETF is the SP100 index, where purchasing this ETF would mean that you own a part of all the stocks that are included in the index.

This project will investigate the returns of some ETFs over a particular period. First, a descriptive analysis will be conducted – meaning the data will be examined. The second part of the project is the statistical analysis, where hypotheses will be tested.

Descriptive Analysis

a) A Description of the data & its quality *(R: line 5-27)*

The data contains a total of 454 observations (rows) and 96 random variables (columns), where the first column contains the date and the other 95 are the different ETFs. Additionally, the data covers the period between 5/5/2006 and 8/5/2015, where the first observation is recorded on the date 5/5/2006, and the last one is recorded on the data 8/5/2015.

When observing the dataset, we see that the data is overall of very good quality. The total number of cells that contained no data in the dataset was zero, and so there were no missing observations.

b) Examining the distribution of 4 particular ETFs *(R: line 31-56)*

Having described the dataset, we will now select 4 ETFs and analyse their distributions by determining their empirical densities, boxplots, and summary statistics. The selected ETFs are AGG, VAW, IWN, and SPY.

The empirical density plot provides us with an overview of the spread of the data and where most of the data lies. The empirical density plots for the 4 ETFs were found using R, and are shown below:

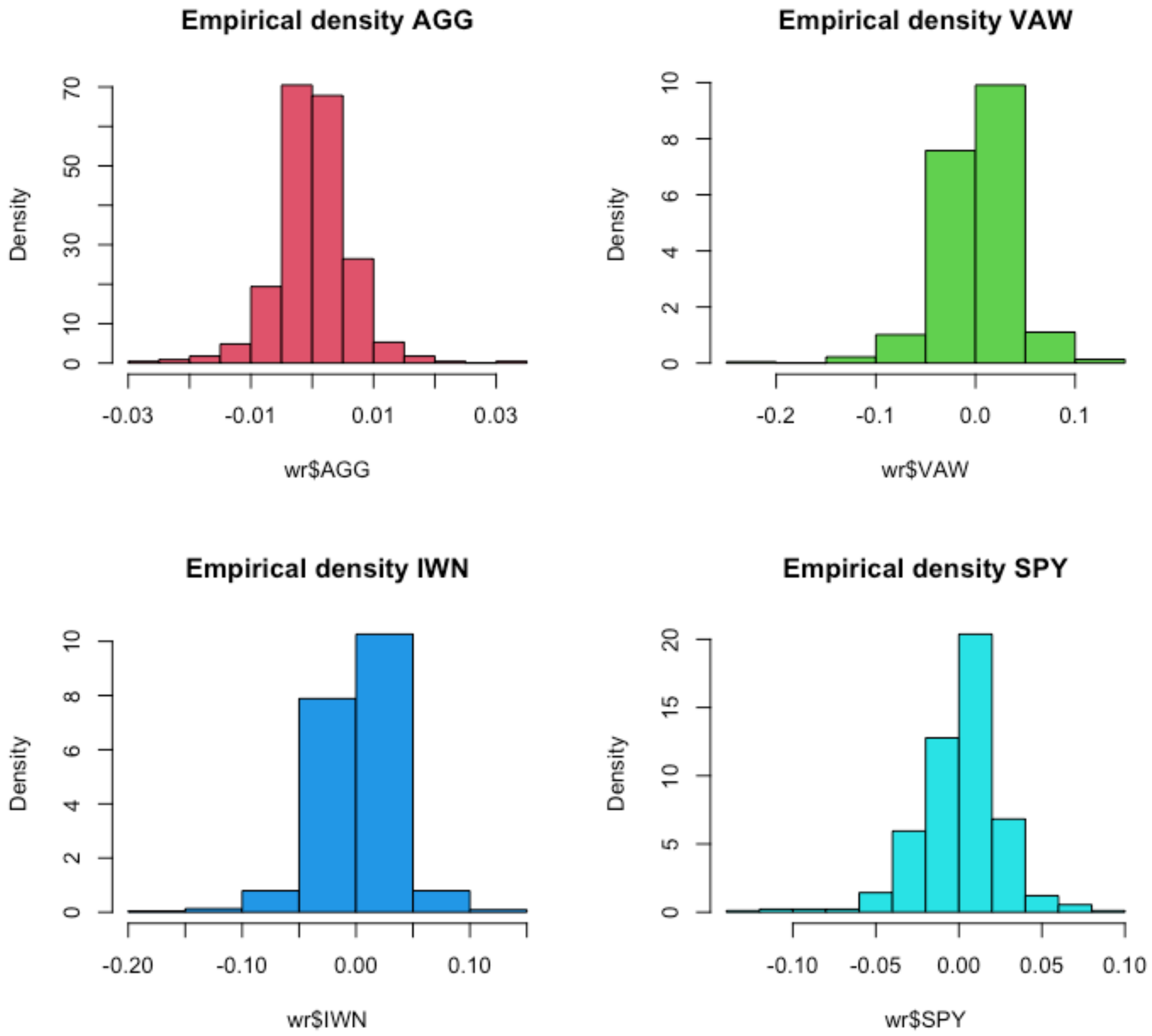


Figure 1: Empirical Density plots for the 4 ETFs.

From Figure 1, we see that the all of the plots are approximately centralized around zero. They each also follow a normal distribution to some extent. From these plots alone it is difficult to spot the outliers in the data, we therefore need to create a boxplot, which provides us with another perspective of the distributions.

The boxplots for each ETF were plotted using R:

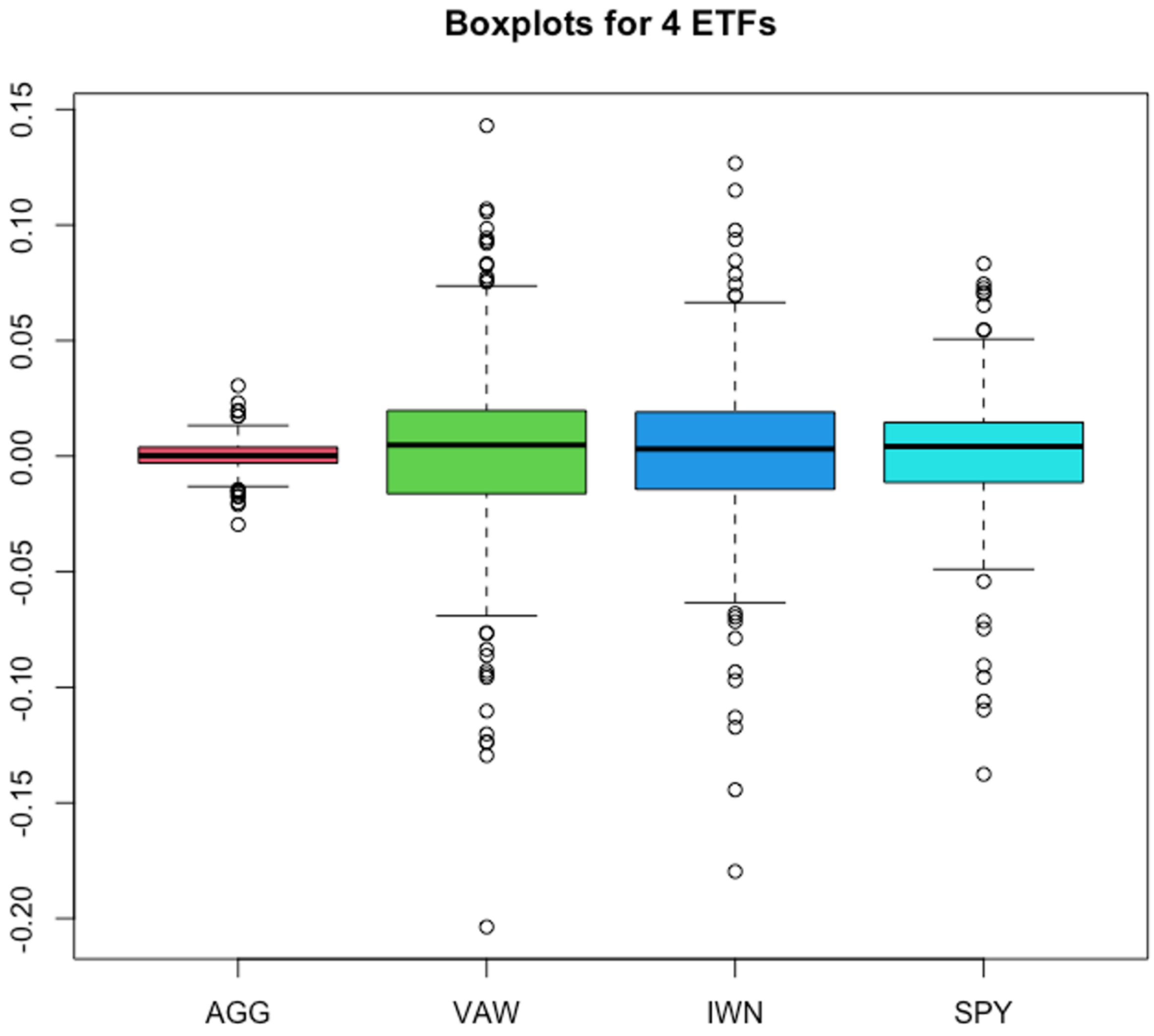


Figure 2: Boxplots for the 4 ETFs.

From the above boxplots, we see that each of the ETFs have a number of outliers, as shown by the black circles. However, the outliers of AGG are not as extreme as the ones for VAW and IWN, which are very spread out from the median. This indicates that AGG has the least deviation in its distribution, whereas VAW has the most deviation.

Some summary statistics were then calculated for the ETFs, including sample mean, sample variance, standard deviation, and quartiles. The following values were calculated using R for the 4 ETFs.

EFT	Number of obs.	Sample mean	Sample variance	Std dev	Q1	Median	Q3
AGG	454	0.00027	0.000036	0.0060	-0.0030	0.00024	0.0039
VAW	454	0.0018	0.0013	0.036	-0.016	0.0048	0.020
IWN	454	0.0012	0.0010	0.032	-0.014	0.0031	0.019
SPY	454	0.0014	0.00061	0.025	-0.011	0.0042	0.015

Table 1: Summary statistics for the 4 ETFs

The above table tells us that AGG has the smallest deviation, whereas VAW has the greatest deviation. This aligns with what our boxplot tells us. We also see that the sample mean of AGG is significantly smaller than the other ETFs, meaning it had a smaller average return than the others during the period in question.

c) Describing the distribution of weekly returns for the ETFs

We now want to describe the distributions by determining whether they are symmetrical or skewed. The skewness of a distribution can be found by seeing if the mean of the ETF is located to the left or to the right of its median.

We can find the location of the mean relative to the median by finding their difference – if the difference is negative, it is right skewed. If positive, then it is left skewed. If the difference is approximately zero, then the distribution is considered symmetrical. The skewness of the ETF distributions is found using the values from Table 1:

ETF	Mean – median	Determining skewness
AGG	$3.0 \cdot 10^{-5}$	Approximately symmetric
VAW	-0.0030	Slightly left skewed
IWN	-0.0019	Slightly left skewed
SPY	-0.0028	Slightly left skewed

Table 2: The skewness of the distributions for the 4 ETFs.

We therefore describe that the distributions for VAW, IWN, and SPY are slightly left skewed, meanwhile the distribution for AGG is approximately symmetrical.

This also aligns with the boxplot, which shows that the extreme outliers for VAW, IWN, and SPY are located to the left of the median and are therefore left skewed.

The minimum and maximum values of the ETFs can also be used to investigate the skewness of the distribution:

ETF	**Min**	**Max**	**Max - Min**
AGG	-0.030	0.031	0.061
VAW	-0.204	0.143	0.347
IWN	-0.18	0.13	0.31
SPY	-0.14	0.083	0.223

Table 3: Minimum and maximum values for the 4 ETFs.

We see from this table that AGG has the smallest difference between min and max, meaning it has the smallest deviation. Furthermore, it is also very symmetrical about its mean, and the min and max values are almost equidistant from the mean. Additionally, VAW has the largest difference between the minimum and maximum values, which also aligns with our other observations.

Statistical Analysis I

Problem 1 – ETF Portfolio

When constructing an ETF portfolio, it is important to assess the risks of the portfolio in order to avoid "putting all eggs in one basket". We can begin to measure the risk of a portfolio by computing the covariances between the various ETFs.

d) Computing the covariance between ETFs in a portfolio *(R: line 60-65)*

We will now consider 6 ETFs, namely AGG, VAW, IWN, SPY, EWG, and EWW. We can determine the covariances between all pairs using R:

ETF	AGG	VAW	IWN	SPY	EWG	EWW
AGG	0.000036	-0.000043	-0.000026	-0.000032	-0.000051	-0.000037
VAW	-0.000043	0.001302	0.000984	0.000793	0.001110	0.001185
IWN	-0.000026	0.000984	0.001025	0.000722	0.000950	0.001010
SPY	-0.000032	0.000793	0.000722	0.000614	0.000805	0.000815
EWG	-0.000051	0.001110	0.000950	0.000805	0.001444	0.001180
EWW	-0.000037	0.001185	0.001010	0.000815	0.001180	0.001659

Table 4: A covariance matrix for the portfolio. Notice grey cells are the variances of each ETF.

From the covariance matrix we see that EWW and VAW have the strongest positive covariance of any of the pairs, meanwhile AGG and EWG have the strongest negative covariance.

e) Minimizing the variance of portfolios *(R: line 67-96)*

Let us now only consider portfolios consisting of two ETFs, where we will try to minimize the variance. The following portfolios will be investigated: (EWG, EWW), (AGG, SPY), (VAW, IWN), (VAW, EWG), (VAW, EWW), and (IWN, EWG).

We will now define a random variable P_i for each portfolio that describes the proportion of the portfolio invested in each ETF:

$P_1 = \alpha \cdot X_{EWG} + (1-\alpha) \cdot X_{EWW}$

$P_2 = \alpha \cdot X_{AGG} + (1-\alpha) \cdot X_{SPY}$

$P_3 = \alpha \cdot X_{VAW} + (1-\alpha) \cdot X_{IWN}$

$P_4 = \alpha \cdot X_{VAW} + (1-\alpha) \cdot X_{EWG}$

$P_5 = \alpha \cdot X_{VAW} + (1-\alpha) \cdot X_{EWW}$

$P_6 = \alpha \cdot X_{IWN} + (1-\alpha) \cdot X_{EWG}$

Where X_{EWG} represents the weekly returns for EWG. The process of minimizing variance for a portfolio will only be shown for the first portfolio. We start by finding an expression for the variance of P_1:

$$\begin{aligned}\mathrm{Var}(P_1) &= \mathrm{Cov}(P_1, P_1)\\ &= \alpha^2 \cdot \mathrm{Var}(X_{EWG}) + (1-\alpha)^2 \cdot \mathrm{Var}(X_{EWW}) + 2 \cdot \alpha(1-\alpha) \cdot \mathrm{Cov}(X_{EWG}, X_{EWW})\end{aligned}$$

We now insert the variance and covariance values found in part **d)** to obtain an expression in terms of α, $V(\alpha)$:

$$V(\alpha) = 0.0014 \cdot \alpha^2 + 0.0017 \cdot (1-\alpha)^2 + 0.0012 \cdot 2 \cdot \alpha(1-\alpha)$$

We can now plot this function in R to see where the variance in minimized:

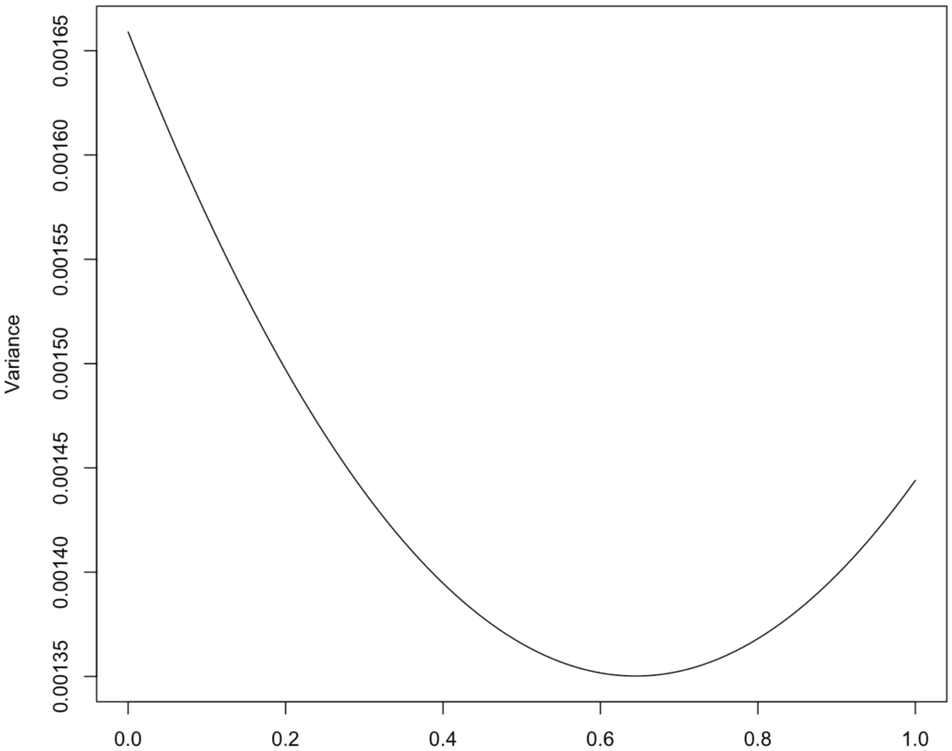


Figure 3: Variance as a function of alpha.

We can find the value of alpha that minimizes the variance by writing $V(\alpha)$ in quadratic form, as shown below:

$$V(\alpha) = \alpha^2 \cdot \left(\mathrm{Var}(X_{EWG}) + \mathrm{Var}(X_{EWW}) - 2 \cdot \mathrm{Cov}(X_{EWG}, X_{EWW})\right) + \alpha \cdot \left(-2 \cdot \mathrm{Var}(X_{EWW}) + 2 \cdot \mathrm{Cov}(X_{EWG}, X_{EWW})\right) + \mathrm{Var}(X_{EWW})$$

We can then find the vertex (minimum point) by using the formula $-\frac{b}{2a}$. The value of alpha that minimizes the variance, α_m, is found by:

$$\alpha_m = -\frac{b}{2a} = \frac{-\left(-2 \cdot \mathrm{Var}(X_{EWW}) + 2 \cdot \mathrm{Cov}(X_{EWG}, X_{EWW})\right)}{2 \cdot \left(\mathrm{Var}(X_{EWG}) + \mathrm{Var}(X_{EWW}) - 2 \cdot \mathrm{Cov}(X_{EWG}, X_{EWW})\right)} = 0.64$$

The minimum variance is found by inserting α_m into $V(\alpha)$, and the expected weekly returns for the portfolios with minimum variance is found by α_m into P_i. These values were found for each of the portfolios and are shown below Table 5:

ETF	α_m	Minimum variance	Expected weekly returns
(EWG, EWW)	0.64	0.0014	0.0014
(AGG, SPY)	0.90	0.000029	0.00037
(VAW, IWN)	0.11	0.0010	0.0013
(VAW, EWG)	0.64	0.0012	0.0016
(VAW, EWW)	0.80	0.0013	0.0018
(IWN, EWG)	0.87	0.0010	0.0012

Table 5: The alpha values, minimum variances, and returns of all portfolio combinations.

From Table 5, we see that the portfolio (AGG, SPY) has the smallest variance as well as the smallest expected weekly returns. On the other hand, (VAW, EWW) has one of the largest variances and has the greatest expected weekly returns.

The best portfolios that have a reasonably low variance, as well as a relatively high return are (VAW, EWG) and (VAW, EWW).

If $\alpha_m < 0$ or $\alpha_m > 1$, then there would be a negative proportion for one of the ETFs.

Problem 2 – Best Investment

f) **Modelling the 4 ETFs** *(R: line 101-117)*

We now wish to model the ETFs: AGG, IWN, VAW, and SPY. From the empirical density plots in Figure 1, we can assume that the ETFs are normally distributed. We also must assume that the data is independently and identically distributed (i.i.d) in order to model the distributions, even though this may not be the case.

The normal distribution parameters (mean and variance/standard deviation) were computed for each ETF. These estimated parameters are shown in the table below:

ETF	**Mean**	**Variance**	**Standard deviation**
AGG	0.00027	0.000036	0.0060
VAW	0.0018	0.0013	0.036
IWN	0.0012	0.0010	0.032
SPY	0.0014	0.00061	0.025

Table 6: The normal distribution parameters for the ETFs.

Now that we have made model assumptions and calculated the estimated model parameters, we can now carry out a model validation for each ETF by plotting the QQ normal function for each ETF:

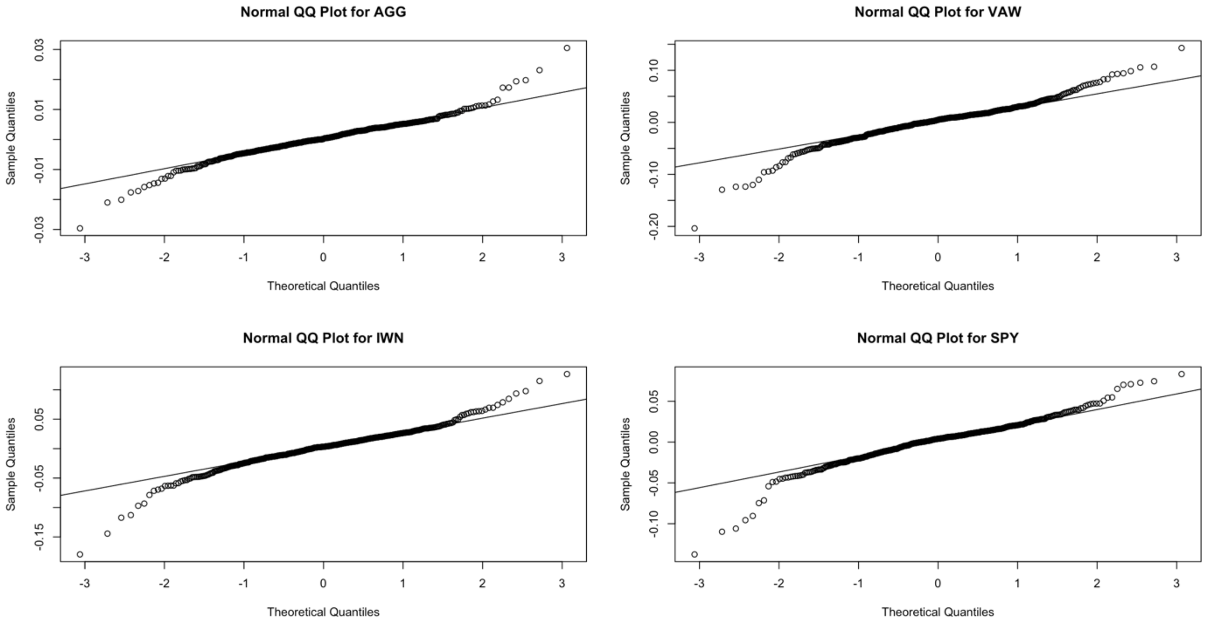


Figure 4: The QQ Normal plots for the 4 ETFs. Notice that they approximately fit the straight line.

As shown in the QQ normal plots above, the data points approximately follow a straight line, meaning that the normal distribution holds and the assumption is considered valid. If we take a deeper look, the points that are close to the median fall neatly on the straight line, whereas points near the tails of the distributions deviate much more from the line. We also see that assumption fits AGG the best, as the points deviate the least. On the other hand, VAW has the largest deviation and the assumption does not fit its distribution quite as well as the others.

The central limit theorem applies to the ETF distributions, as the number of observations is greater than the thumb rule of 30, and so the average weekly returns approximately follow a normal distribution.

g) Confidence intervals for the ETFs *(R: line 122-151)*

We wish to find the confidence intervals for the average weekly returns for the ETFs as well as the variance. The intervals will be computed for each of the four ETFs, but the method will only be shown for AGG.

The formula for finding the confidence interval for the average weekly return is:

$$\bar{x} \pm t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

We want to find the 95% confidence interval, so α is 0.05, and s is the standard deviation. We insert the mean and number of observations found in Table 1:

$$0.00027 \pm 1.965 \cdot \frac{0.0060}{\sqrt{454}} = [-0.00029, 0.00082]$$

We can find the confidence interval for the variance using the following formula:

$$\sigma^2 \colon \left[\frac{(n-1)\cdot s^2}{\chi^2_{1-\frac{\alpha}{2}}}, \frac{(n-1)\cdot s^2}{\chi^2_{\frac{\alpha}{2}}}\right]$$

We now use R to calculate the denominators and insert the variance from Table 1:

$$\sigma^2 \colon \left[\frac{(454-1)\cdot 0.000036}{\chi^2_{0.975}}, \frac{(454-1)\cdot 0.000036}{\chi^2_{0.025}}\right] = \left[\frac{453\cdot 0.000036}{513.9}, \frac{453\cdot 0.000036}{395.9}\right]$$

$$= [0.000031, 0.000041]$$

We have now found the 95% confidence intervals for the mean and variance of AGG. The intervals for the other ETFs are shown in the table below:

ETF	Lower bound (mean)	Upper bound (mean)	Lower bound (variance)	Upper bound (variance)
AGG	−0.0029	0.00082	0.000031	0.000041
VAW	−0.0015	0.0051	0.0011	0.0015
IWN	−0.0018	0.0041	0.00090	0.0012
SPY	−0.00093	0.0036	0.00054	0.00070

Table 7: Confidence intervals for mean and variance for the 4 ETFs.

The table shows that, for both the mean and variance confidence intervals, VAW has the widest interval whereas AGG has the most narrow. This information matches the observations made for the normal QQ plots in Figure 3.

h) Non parametric bootstrap *(R: line 156-191)*

Having found these confidence intervals, we now want to determine the confidence intervals for the mean and variance with non-parametric bootstrap. The simulations are done in R, and the intervals for the 4 ETFs are shown in the table below:

ETF	Lower bound (mean)	Upper bound (mean)	Lower bound (variance)	Upper bound (variance)
AGG	−0.0029	0.00081	0.000028	0.000044
VAW	−0.0016	0.0051	0.0010	0.0016
IWN	−0.0018	0.0042	0.00081	0.0013
SPY	−0.00097	0.0036	0.00048	0.00077

Table 8: Confidence intervals with non-parametric bootstrap.

If we compare these intervals with the ones in table 7, we see that there are fairly small differences between them. For the mean confidence intervals, there is a very small difference, however, there is a more significant difference for the variance intervals.

i) Saving money under pillow vs. average weekly return *(R: line 195-213)*

We now want to investigate the hypothesis that saving your money under the pillow (i.e. its value remains constant) does not differ much from the average weekly returns from investing in the ETFs. This hypothesis can be expressed as:

$$H_0\colon\ \mu_{ETF} = 0$$

We can investigate the hypothesis using the built in t-test function, which computes the probability of obtaining a test statistic is at least as extreme as the test statistic that was actually observed, known as the p-value. The t-test was run for the 4 ETFs:

ETF	Test statistic	Degrees of freedom	p-value
AGG	0.947	453	0.344
VAW	1.06	453	0.291
IWN	0.790	453	0.430
SPY	1.17	453	0.243

Table 9: t-test values for the 4 ETFs.

As the p-value for each ETF is greater than 0.1, the t-test suggests that there is little to no evidence against the null hypothesis for each ETF. Therefore, we cannot reject the hypothesis that saving money under the pillow does not differ much from the average weekly return.

Problem 3 – Similarity between weekly returns

j) **Similarity hypothesis testing** *(R: line 219-224)*

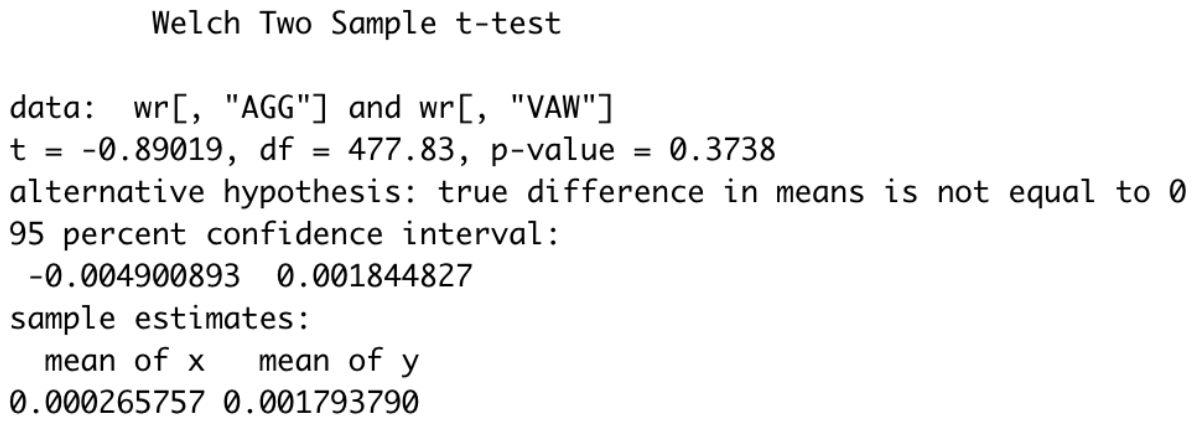
We now want to examine whether there is a similarity between the weekly returns for the 4 ETFs. We can do this by statistically analysing the following hypothesis:

$$H_0:\ \mu_{ETF_{low}} = \mu_{ETF_{high}}$$

By looking at Table 1, The ETF with the lowest average weekly return is AGG and the ETF with the highest average weekly return is VAW:

$$H_0:\ \mu_{ETF_{low}} = \mu_{ETF_{high}} \Longrightarrow \mu_{AGG} = \mu_{VAW}$$

In order to use the built-in t-test function for testing the hypothesis, we again must assume that the data is normally distributed and independent. As we are assuming that the two ETFs weekly returns are independent, then the data is not considered paired. The significance level was chosen to be $\alpha = 0.05$, and the results for the Welch two sample t-test is shown below:



```
        Welch Two Sample t-test

data:  wr[, "AGG"] and wr[, "VAW"]
t = -0.89019, df = 477.83, p-value = 0.3738
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.004900893  0.001844827
sample estimates:
  mean of x   mean of y 
0.000265757 0.001793790 
```

Figure 5: Welch two sample t-test for highest and lowest weekly return.

As reported in the figure, the p-value is 0.3738, which suggests that there is little to no evidence against the hypothesis, meaning that we must accept it.

The critical values are $t_{1-\alpha/2}$ and $t_{\alpha/2}$, whereas the p-value is found using the formula:

$$Pvalue = 2 \cdot P(T > |t_{obs}|)$$

Statistical Analysis II

Problem 3 – Similarity between weekly returns

Volatility is the standard deviation of the ratio between the exchange rate of the ETF in question at the beginning and the end of the week, and can be used to measure the risks of the ETF.

The conditional value at risk (CVaR) is another risk measure, which describes the expected loss of the 5 percent worst cases, additionally, the maximum drawdown (maxDD) describes the largest possible loss in a given period and maximum time under water (maxTuW) measures the time needed to return to the historical peak. These are illustrated below:

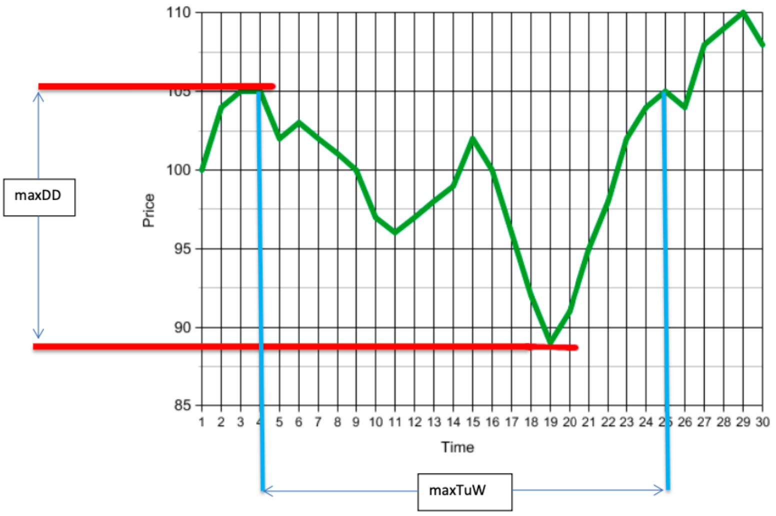


Figure 6: Visualisation of maxDD and maxTuW

We are now given data which includes observations of the 8 variables below for the 95 ETFs:

Variable	Meaning	Unit
X	Name of the ETF	
Geo.mean	Geometric mean relative weekly return r_{week}	Pct.
Volatility	The weekly volatility	Pct.
maxDD	Maximum Draw Down	Pct.
maxTuW	Maximum Time under Water	Weeks
VaR	Weekly Value-at-Risk	Pct.
CVaR	Weekly Conditional Value at Risk	Pct.

Our goal is to use empirical correlations in order to assess the relations between the numerical variables.

k) Scatter plots & empirical correlations *(R: line 234-266)*

We will now use the given data to create scatter plots for the following relations: Volatility vs. CVaR, Geo.mean vs. maxTuW, Volatility vs. maxDD, maxTuW vs. Volatility. These plots are shown in the figure below:

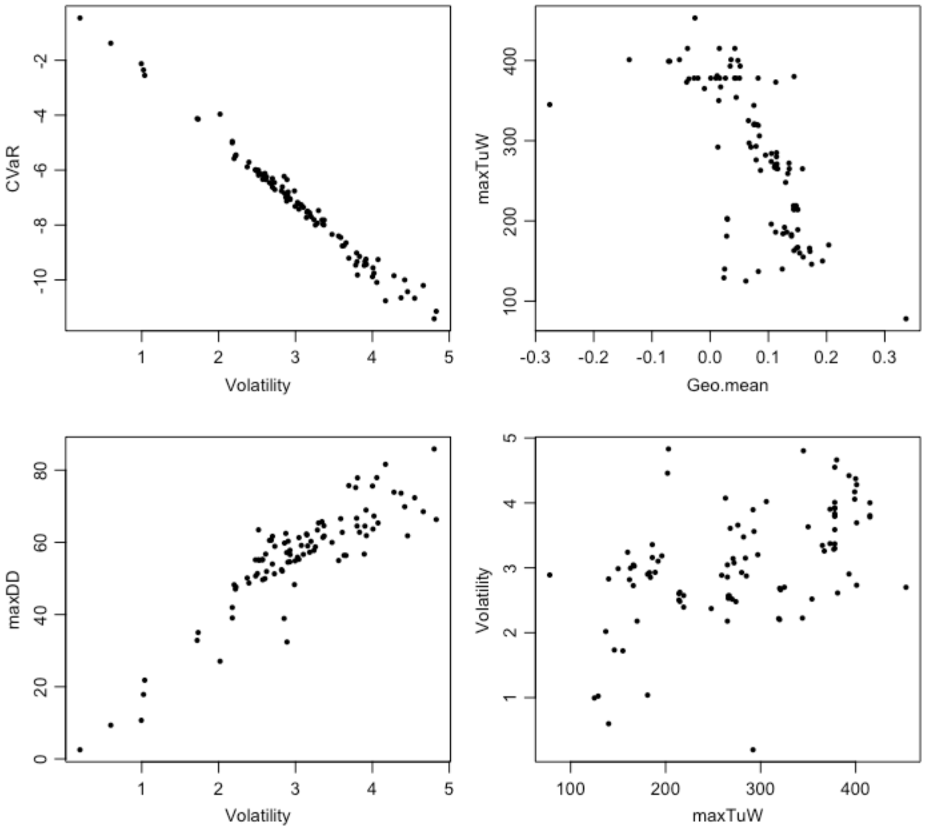


Figure 7: Scatter plots.

From looking at these scatter plots, it is evident that CVaR vs. Volatility has a clear negative correlation, and maxDD vs. Volatility has a clear positive correlation. Geo.mean vs. maxTuW seems to have a light negative correlation, whereas maxTuW vs. Volatility seems to have no correlation.

Let us now compute the empirical correlation for Geo.mean vs. maxTuW using the formula below:

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\mathrm{cov(Geo.mean, maxTuW)}}{\mathrm{sd(Geo.mean)} \cdot \mathrm{sd(maxTuW)}} = \frac{-4.9635}{0.08087 \cdot 91.8358} = -0.668$$

This value indicates indeed that there is a fairly strong negative correlation between the two variables, which was slightly difficult to see from the scatterplots. This value is also identical to the correlation computed using R.