Danmarks Tekniske Universitet

# Trading with ETFs


**Project**
**by**

**William Peytz (s204145)**


**Introduction to Mathematical Statistics, 02403**


**Technical University of Denmark**

**June 15, 2021**

## Contents

**1 Descriptive Analysis** **1**

**2 Statistical Analysis I** **3**

2.1 Problem 1 - ETF Portfolio . . . . . 3

2.2 Problem 2 - Best investment . . . . . 5

2.3 Problem 3 . . . . . 7

**3 Statistical Analysis II** **8**

3.1 Problem 4 . . . . . 8

# 1 Descriptive Analysis

a) Our dataset finance1 consists of a matrix of 95 different ETFs and for each these we have the weekly returns of each ETF over the course of 454 weeks. This gives us a total of 43.130 observations over then span of approximately 9 years. The first observations were made on the 5th of May 2006 and the last observations were made on 8th of May 2008. In relation to the quality of the data, we can see that no datapoints are missing, which is important for further analysis. It can be noted that there is not always exactly one week between the datapoints, but this is not a huge problem for comparison purposes as all ETFs still have the same amount of datapoints from the same dates and time intervals.

b)

| ETF | number of obs. | sample mean | sample variance | Std. dev. | Lower quartile | Median | Upper Quartile |
|---|---|---|---|---|---|---|---|
| AGG | 454 | 0.0002658 | 0.00003571 | 0.005976 | -0.003048 | 0.000450 | 0.003907 |
| VAW | 454 | 0.001794 | 0.001302 | 0.03608 | -0.016248 | 0.004871 | 0.019737 |
| IWN | 454 | 0.001188 | 0.001025 | 0.03202 | -0.014349 | 0.003141 | 0.01906 |
| SPY | 454 | 0.001360 | 0.0006143 | 0.02479 | -0.011357 | 0.004221 | 0.014523 |

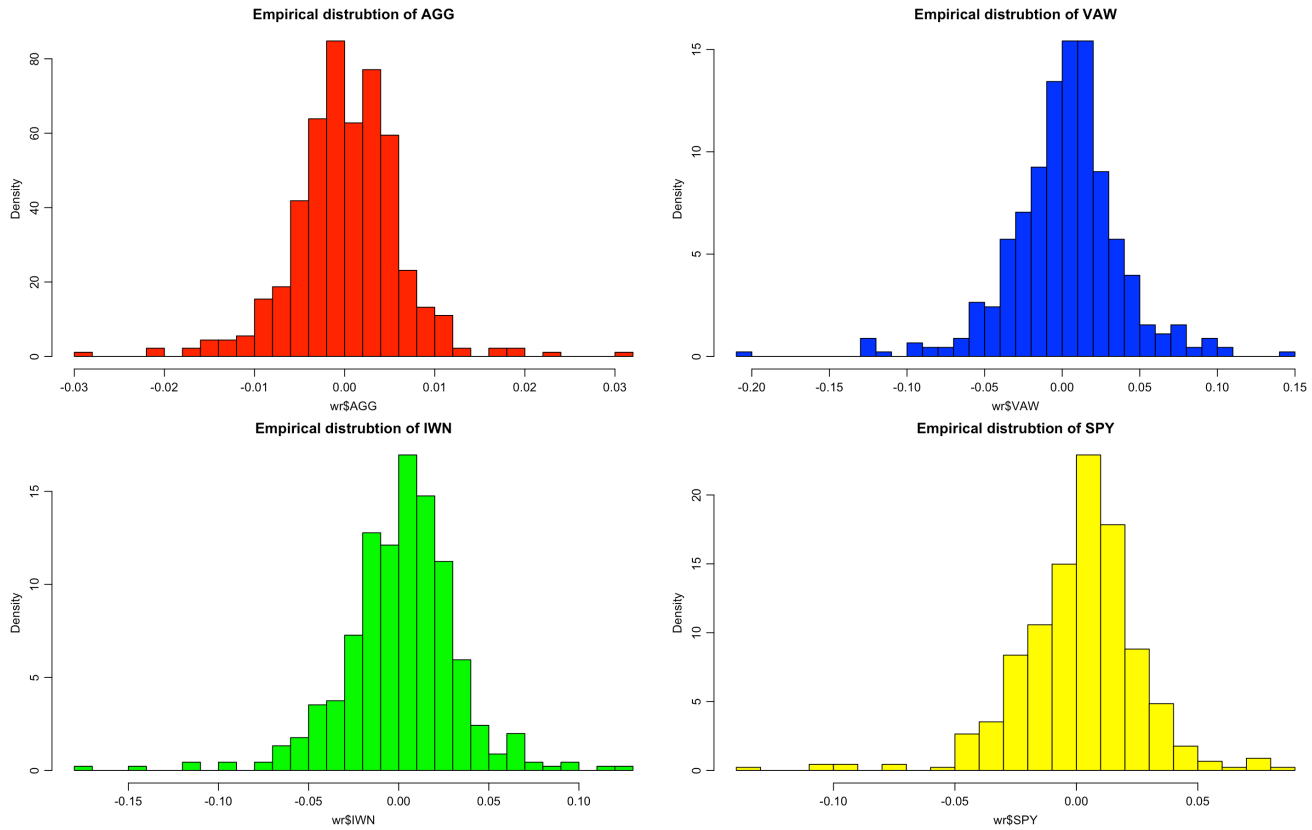


Figure 1: The empircal density of AGG

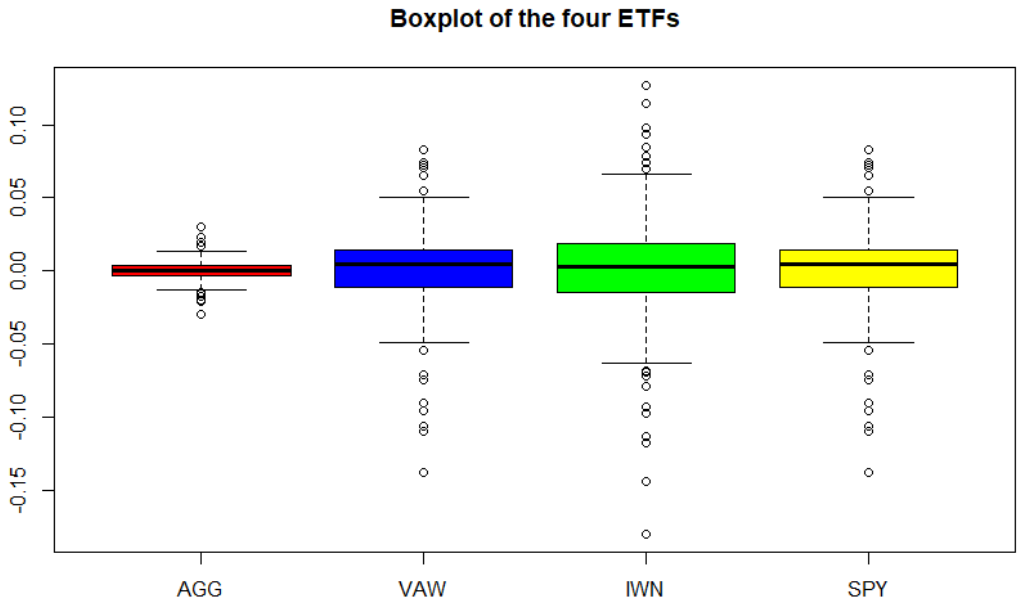


Figure 2: Box plots of all 4 ETFs

c) Descriptions of the weeekly returns of the four ETFs.

If we look at the weekly returns of the AGG-ETF we can see that this ETF is the clear ourlier in the comparison. It is seems to be symmetrically distrubed around the mean. It has a sample mean that is signicantly lower than the other three ETFs, which means that on average this ETF will give the lowest expected returns. If we want to compare the sample means of the different ETFs it can make sense to convert it to yearly returns as these figures are easier to comprehend. We can do this with the exponential formula:

$$ExpectedYearlyReturn = (1 + WeeklySampleMean)^{52}. \tag{1}$$

This gives us a expected yearly return of

$$(1 + 0.000271)^{52} = 1.01419 \tag{2}$$

This is roughly equal to a return of 1.42% on average pr year. When we compare this return to the other ETFs it becomes really clear how poor this return is compared to the others. The variance is 17 to 37 times lower than the others which means that our returns won't fluctuate nearly as much as it will with the other three. This can also be seen when we look at the minimum and maximum values of our sample from the ETF. The sample maximum is 0.03051 and the sample minimum is -0.02960, for a total range of 0.06011, which is low compared to the other three. The ETF has a total of 15 outliers which is few compared to the others, which makes sense due to the lower variance. We can conclude that this ETF is very conservative, since it has a lower variance and lower expected return compared to the others.

The VAW-ETF is very different compared to the AGG-ETF. It also seems symmetrical arounds its mean, but as the VAW sample mean is larger, it has more weeks with a positive weekly return compared to the others. It has a weekly sample mean of 0.001798. When calculated to a yearly return it gives us the following

$$(1 + 0.001798)^{52} = 1.0979 \tag{3}$$

This is equal to a return of 9.79% on average pr year, which is almost a seven times better return than the yearly AGG returns. This ETF does come with more risk compared to AGG, as its variance is 0.00130197, 37 times higher than AGG. This risk can also be seen on the sample max and min 0.14298 and -0.20366 for a total range of 0.34664, more than five times greater than AGG. It also has more outliers at 25. We can conclude that this is a

way more aggressive ETF, as it has a higher mean, but also a higher variance.

The IWN-ETF is is similar to the VAW, but is is less extreme. It seems symmetrical around its mean. It has a weekly sample mean of 0.001190. The expected yearly return would therefore be

$$(1 + 0.001190)^{52} = 1.06380 \tag{4}$$

This is a return of 6.38% pr year, which is around 35% lower than the VAW return, but still way above the AGG return. It has a variance of 0.00102499, a max and min of 0.1267016 and -0.179655 for a total range of 0.3063565. It has 19 outliers. This is also a more aggressive ETF, but not as aggressive as VAW.
The final ETF is SPY. It also seems symmetrical around its mean. It has a sample mean of 0.001363. For a yealy return it would be:

$$(1 + 0.001363)^{52} = 1.07340 \tag{5}$$

This is a return of 7.34% per year, which makes it fall in between the VAW and IWN ETFs. It has a variance of 0.00061435 which is actually lower than the IWN-ETF. This is interesting as this means that the SPY-ETF, is therefore objectivily better than the IWN-ETF as it both has a higher mean and and a lower variance. This means that you can expect higher returns with less risk. It does still though have a lower average return than VAW. SPY has a max and min of 0.0832756 and -0.1375981 for a range of 0.2208737. It has 16 ourliers. This ETF is more aggressive than AGG, more conservative than VAW and objectivily better than IWN.

## 2 Statistical Analysis I

### 2.1 Problem 1 - ETF Portfolio

d) This is the covariance matrix between the following six ETFs: AGG, VAW, IWN, SPY, EWG and EWW:

| | AGG | VAW | IWN | SPY | EWG | EWW |
|---|---|---|---|---|---|---|
| AGG | 0.00003571 | -0.00004261 | -0.00002588 | -0.00003240 | -0.0005084 | -0.00003711 |
| VAW | -0.00004261 | 0.001302 | 0.0009838 | 0.0007927 | 0.001110 | 0.001185 |
| IWN | -0.00002588 | 0.0009838 | 0.001025 | 0.0007222 | 0.0009502 | 0.001010 |
| SPY | -0.00003240 | 0.0007927 | 0.0007222 | 0.0006143 | 0.0008046 | 0.0008153 |
| EWG | -0.0005084 | 0.001110 | 0.0009502 | 0.0008046 | 0.001444 | 0.001180 |
| EWW | -0.00003711 | 0.001185 | 0.001010 | 0.0008153 | 0.001180 | 0.001659 |

An important feature to notice is that we get the variance for the diffrent ETFs in the diagonal.

e) Creating a portfolio of two ETFs so that the variance of the portfolio is minimized.

These are the equations for the corresponding random variables of six different ETF combinations.

$$P_1 = \alpha * X_{EWG} + (1 - \alpha) * X_{EWW} \tag{6}$$

$$P_2 = \alpha * X_{AGG} + (1 - \alpha) * X_{SPY} \tag{7}$$

$$P_3 = \alpha * X_{VAW} + (1 - \alpha) * X_{IWN} \tag{8}$$

$$P_4 = \alpha * X_{VAW} + (1 - \alpha) * X_{EWG} \tag{9}$$

$$P_5 = \alpha * X_{VAW} + (1 - \alpha) * X_{EWW} \tag{10}$$

$$P_6 = \alpha * X_{IWN} + (1 - \alpha) * X_{EWG} \tag{11}$$

Using P1 as an example, this is how the variance of P1 can be expressed:

$$Var(P1) = \alpha^2 * Var(X_{EWG}) + (\alpha - 1)^2 * Var(X_{EWW}) + 2 * (\alpha - \alpha^2) * Cov(X_{EWG}, X_{EWW}) \tag{12}$$

We notice that this variance can be written as a function of $\alpha$ when inserting the different values of our variances and covariance from the covariance matrix from above.

$$V(\alpha) = \alpha^2 * 0.001444 + (\alpha - 1)^2 * 0.001659 + 2 * (\alpha - \alpha^2) * 0.001180 \tag{13}$$

V($\alpha$) can be reduced, so that its quadric nature becomes obvious.

$$V(\alpha) = \alpha^2 * 0.001444 + \alpha^2 * 0.001659 + 0.001659 - \alpha * 0.003318 + \alpha * 0.00236 - \alpha^2 * 0.00236 \tag{14}$$

$$V(\alpha) = 0.000743 * \alpha^2 - 0.000958 * \alpha + 0.001659 \tag{15}$$

We can plot this functinon V($\alpha$)

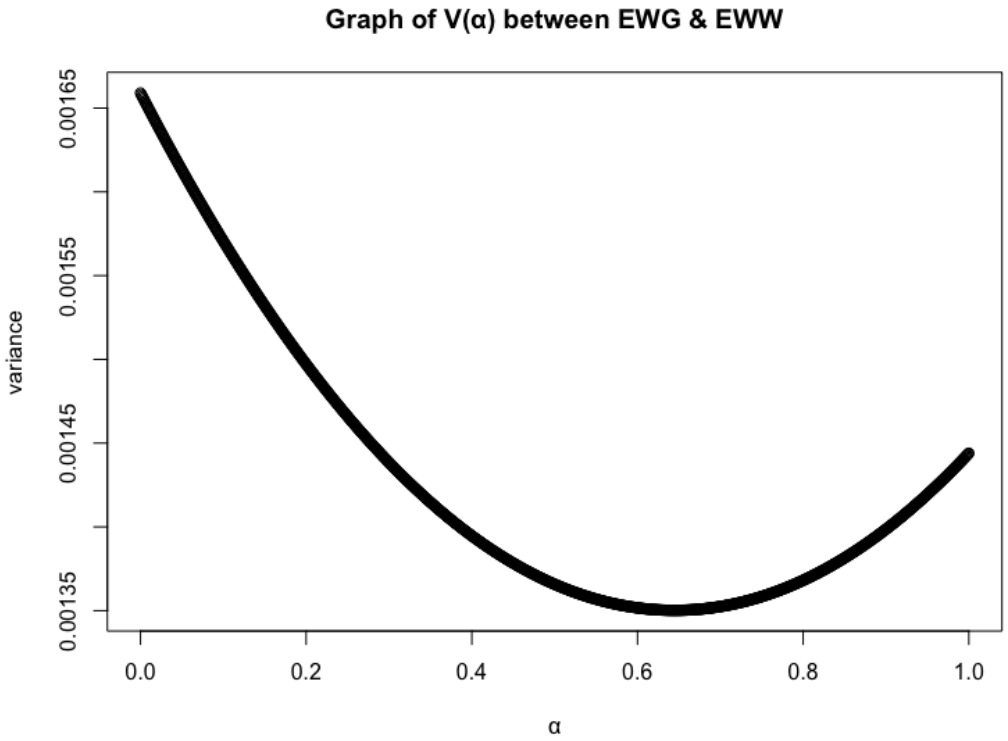


Figure 3: The V($\alpha$) function for P1(EWG, EWW)

Since the quadratic coefficient in the parabola is positive, we know that the vertex of the parabola is a global minimum. Therefore we can differentiate V($\alpha$) and set it equal to zero, to find the value of $\alpha$ that minimises V($\alpha$).

$$V'(\alpha) = 0, \alpha <=> \alpha = 0.6449 \tag{16}$$

This value of $\alpha$ can be inserted to find the values of var($P_i$) and E($P_i$) with the minimized variance in mind. This process can be repeated for all of our six combinations of ETFs. It has been filled out in the following table:

| | P1(EWW,EWG) | P2(AGG,SPY) | P3(VAW,IWN) | P4(VAW,EWG) | P5(VAW,EWW) | P6(IWN,EWG) |
|---|---|---|---|---|---|---|
| $\alpha$ | 0.6449 | 0.9047 | 0.1146 | 0.6351 | 0.8020 | 0.8685 |
| Var($P_i$) | 0.001350 | 0.00002922 | 0.001020 | 0.001232 | 0.001279 | 0.001015 |
| E($P_i$) | 0.001538 | 0.0003700 | 0.001257 | 0.001590 | 0.001776 | 0.001194 |

When picking an ETF portfolio it is in our best interest to pick one that maximises our expected return, while at the same time trying to minimize the variance. When looking at the table it is clear to see that P5, the combination of VAW and EWW is the best option when it comes to expected weekly return. In regards to the variance we have computed that a split of 80.2% VAW and 19.8% EWW is optimal, as this is the split that minimizes the variance for the ETF-combination. One could argue that if you're very risk averse, that you don't want to pick the portofolio, because it has a higher variance than some of the other combinations. Alternatives for this type of investor could be P3(VAW,IWN) or P6(IWN, EWG).

## 2.2 Problem 2 - Best investment

f) One of the most common models used on data in the real world is the normal distrubtion. It can be interesting to analyze whether our four ETFs AGG, VAW, IWN SPY follow a normal distrubution. A way to validate the assumption about them being normally distributed is to create qqplots for the four ETFs and see how well they fit, as we from this can tell whether or not the data is normally distibuted. Before looking at the qqplot it makes sense to first think about the Central Limit Theorm. The Central Limit Theorm says that

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{17}$$

where Z is a random variable which distrubution function approaches that of the standard normal distrubution, N(0,$1^2$) when n approches $\infty$. What this means is that when we get a large sample size for n, typically above 30, we will be able to tell whether or not the data is normally distributed. Therefore, since we have more than 30 datapoints for our four ETFs, it does make sense to look at our qqplots since we have a sufficient amount of data.

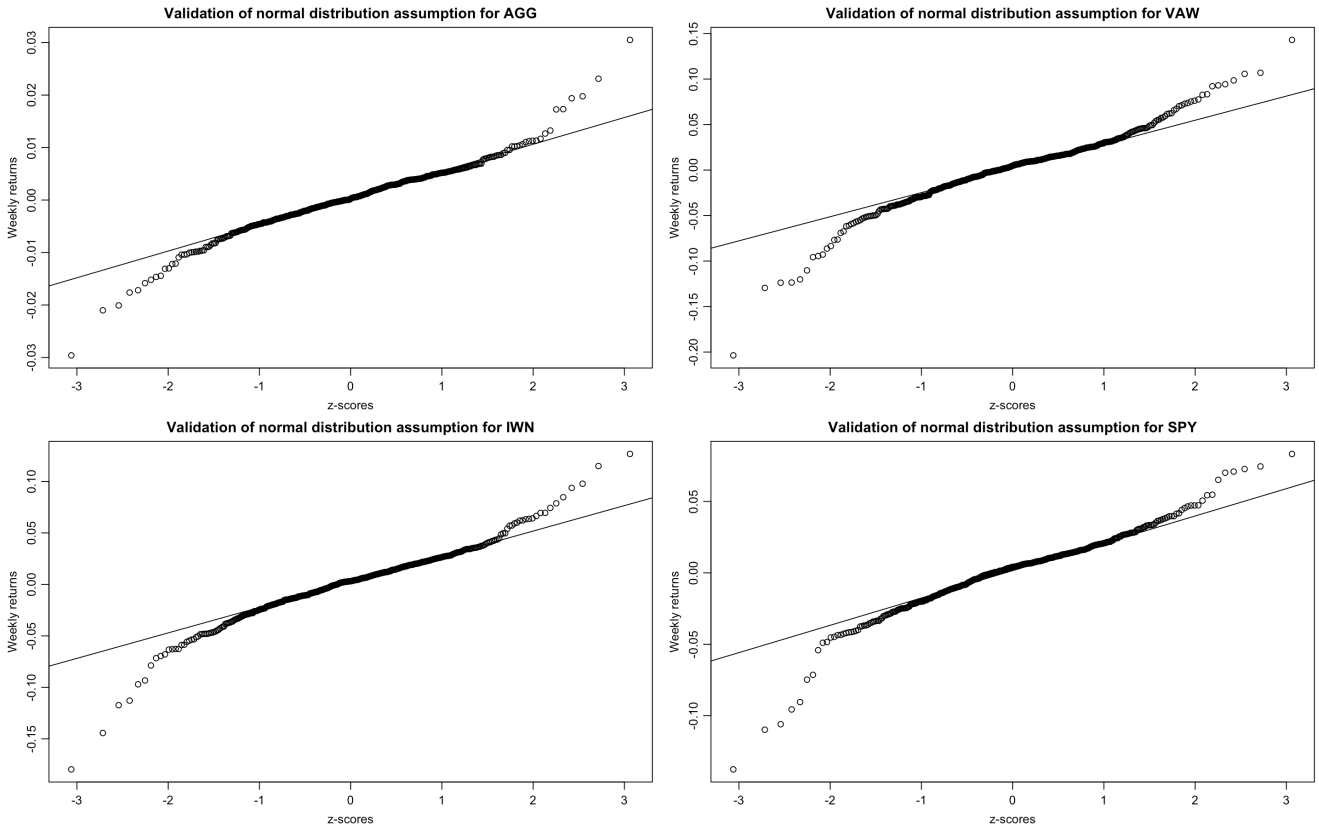


Figure 4: Validation of normal distribution assumption for AGG

When judging whether or not a dataset is normally distributed from a qqplot, we judge how well the datapoints fit the qqline. Overall when looking at the four qqplots they look neither right nor leftskewed, but towards the ends of the dataset they seem deviated from the qqline. This is a sign that the data is too centered around the mean, but since most of the data is on the qqline, we can still validate that the data is approximatly normally distributed.

g) To compute a confindence interval for the average weekly return we can use the following formula:

$$\bar{x} \pm t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}} \tag{18}$$

where $\alpha$ is our signiifance level, which is set at 0.05
We can now insert the values for each of the four ETFs to get the confidence intervals for each.

$$\begin{aligned}
AGG: &\quad \bar{X}_{AGG} \pm t_{0.975} \cdot \frac{s_{AGG}}{\sqrt{n_{AGG}}} = 0.0002658 \pm 1.9652 \cdot \frac{0.005976}{\sqrt{454}} = [-0.0002854, 0.0008169] \\
VAW: &\quad \bar{X}_{VAW} \pm t_{0.975} \cdot \frac{s_{VAW}}{\sqrt{n_{VAW}}} = 0.001794 \pm 1.9652 \cdot \frac{0.03608}{\sqrt{454}} = [-0.001534, 0.005122] \\
IWN: &\quad \bar{X}_{IWN} \pm t_{0.975} \cdot \frac{s_{IWN}}{\sqrt{n_{IWN}}} = 0.001188 \pm 1.9652 \cdot \frac{0.03202}{\sqrt{454}} = [-0.001793, 0.004169] \\
SPY: &\quad \bar{X}_{SPY} \pm t_{0.975} \cdot \frac{s_{SPY}}{\sqrt{n_{SPY}}} = 0.001360 \pm 1.9652 \cdot \frac{0.02479}{\sqrt{454}} = [-0.0009264, 0.003646]
\end{aligned}$$

If we instead want to compute the confidence interval of the variance parameter for each of the ETFs, we can use the formula:

$$\sigma^2 = [\frac{(n-1) \cdot s^2}{\chi^2_{1-\alpha/2}}, \frac{n-1) \cdot s^2}{\chi^2_{\alpha/2}}] \tag{19}$$

We now insert the values for each of the four ETFs

$$\begin{aligned}
AGG: &\quad \left[\frac{(454-1) \cdot 0.00003571}{513.8655}, \frac{(454-1) \cdot 0.00003571}{395.9219}\right] = [0.00003148, 0.00004086] \\
VAW: &\quad \left[\frac{(454-1) \cdot 0.001302}{513.8655}, \frac{(454-1) \cdot 0.001302}{395.9219}\right] = [0.001148, 0.001490] \\
IWN: &\quad \left[\frac{(454-1) \cdot 0.001025}{513.8655}, \frac{(454-1) \cdot 0.001025}{395.9219}\right] = [0.0009036, 0.001173] \\
SPY: &\quad \left[\frac{(454-1) \cdot 0.0006143}{513.8655}, \frac{(454-1) \cdot 0.0006143}{395.9219}\right] = [0.0005416, 0.0007029]
\end{aligned}$$

To compare the confidence intervals of the average weekly return and variance of the four ETFs we can compare the width of the intervals:

| | AGG | VAW | IWN | SPY |
|---|---|---|---|---|
| Average weekly return width | 0.0011023 | 0.006656 | 0.005962 | 0.0045724 |
| Variance width | 0.00000938 | 0.0003420 | 0.0002694 | 0.0001613 |

If we look at the widths of the intervals its clear to see the AGG has a significantly lower average weekly return width, as well as a lower variance width. This also makes sense, as AGG had a much lower variance than the others, which in turn makes its confidence interval width lower. We see that the other three ETFs confidence interval width are in roughly the same range, which also makes sense, as their variances are closer to each other.

h) We can find the non paramtric bootstrap confidence intervals for the mean and variance by sampling from our data 10.000 times:

| | non parametric mean confidence interval | non parametric variance confidence interval |
|---|---|---|
| AGG | [-0.0002853;0.0008161] | [0.00002839; 0.00004396] |
| VAW | [-0.001511; 0.005136] | [0.001033; 0.001608] |
| IWN | [-0.001859 0.004096] | [0.0008095; 0.001275] |
| SPY | [-0.0009658; 0.003650] | [0.0004793; 0.0007643] |

These new bootstrap confidence intervals can be compared with the previous confidence intervals. The general tendency is that for the mean confidence intervals they cover almost the same interval with both methods, but with the variance confidence intervals the bootstrap intervals are wider than the normal confidence interval. The normal variance intervals are entirely covered by the bootstrap intervals.

i) We can test if there is a significant difference between getting the average weekly returns (AWR) from the four ETFs compared to not investing at all, by looking at whether or not 0 is within the confidence intervals of the AWRs of the four ETFs. If 0 is in the interval for a given ETF, we must accept the null hypothesis that there is no significant difference between getting the AWR and not investing in the ETF. However if 0 is not in the intervals, we must reject the null hypthosis and can therefore conclude that the AWR of the ETF difer significantly from not investing at all. The four 95% confidence intervals of the AWR for AGG, VAW, IWN and SPY are as follows: [-0.0002854,0.0008169], [-0.001534,0.005122], [-0.001793,0.004169] and [-0.0009264,0.003646]. Since all of the four intervals include the number 0, we have to accept our null hypothesis for all four, which means that the average weekly return for these four ETFs do not difer statistically significantly from not investing at all.

## 2.3 Problem 3

j) In order to find out if there are significant differences betweeen the average weekly returns between the four ETFs it makes sense to compare the two ETFs which have the biggest difference in AWR, as if they do not difer significantly, we can conclude than none of them will. The AGG-ETF has the lowest AWR at 0.000256 per week and that the VAW-ETF has the highest AWR at 0.001302 per week. Therefore this comparison should be sufficient in order to find out if there are significant differences between the ETFs. For the comparison the significance level, $\alpha$, will be set at 0.05. We can use the welch two sample t-test statistic:

$$t_{obs} = ((\bar{x}_1 - \bar{x}_2) - \delta_0)/sqrt(s_1^2/n_1 + s_2^2/n_2) \tag{20}$$

where $\delta_0$ is the difference between $\mu_1$ and $\mu_2$

We can insert our numbers for AGG and VAW:

$$t_{obs} = ((0.0002658 - 0.001794) - 0.001528)/sqrt(0.00003571/454 + 0.001302/454) = -1.7805 \tag{21}$$

We also need to calculate the degrees of freemdom. We can do this with the following formula:

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}} \tag{22}$$

$$v = \frac{\left(\frac{0.00003571}{454} + \frac{0.001302}{454}\right)^2}{\frac{(0.00003571/454)^2}{454-1} + \frac{(0.001302/454)^2}{454-1}} = 477.8302 \tag{23}$$

Now that $t_{obs}$ and v has been calculated we can use the following formula to investigate the p-value.

$$p - value = 2 \cdot P\left(T > |t_{\text{obs}}|\right) = 0.07564 \tag{24}$$

Since the p-value of of the t-test is above the critical value of $\alpha = 0.05$, we must keep our null hypthosis, which was that there is no signicant difference between the average weekly return of the AGG and VAW-ETF. Since these

were the two most extreme returns it can therefore be concluded that there are no significant difference between any of the four ETFs.

# 3 Statistical Analysis II

## 3.1 Problem 4

k) In order to analyze the correlation between the different variabes in finance2 it makes sense to create scatterplots.

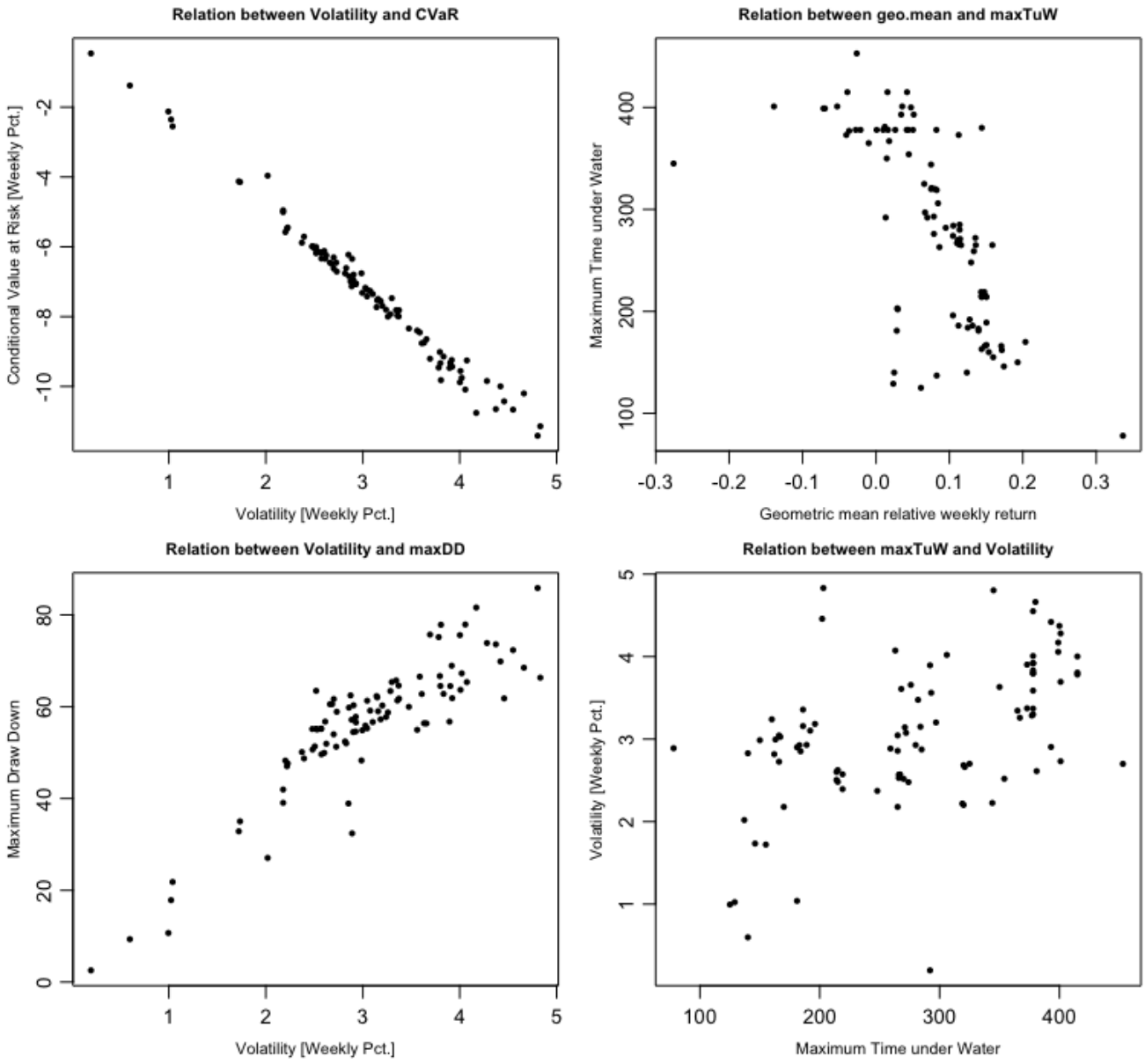


Figure 5: The four scatterplots

When looking at the four scatterplots it looks lige the Volatility and Conditional Value at Risk are strongly negativily correlatated and that Volatility and Maximum Draw Down are positivily correlated, whereas Geometric mean and max Time Under Water, as well as max Time Under Water and Volatility seem less correlated. If we want the exact values of correlation between the different variables we can use the following formula:

$$Corr(x, y) = \frac{Cov(x, y)}{\sigma_x \cdot \sigma_y} \tag{25}$$

We can insert our values for each of the four plots:

$$Corr(Volatility, CVaR) = \frac{-1.8406}{0.8790 \cdot 2.1104} = -0.9922 \tag{26}$$

$$Corr(Geo.Mean, maxTuW) = \frac{-4.9635}{0.08086 \cdot 91.8358} = -0.6684 \tag{27}$$

$$Corr(Volatility, maxDD) = \frac{11.3254}{0.8790 \cdot 14.6440} = 0.8798 \tag{28}$$

$$Corr(maxTuW, Volatility) = \frac{11.3253}{91.8358 \cdot 0.8790} = 0.5108 \tag{29}$$

As expected we can see that there is a strongly negative correlation between Volatility and CVar and strongly positive between Volatility and maxDD. We also see that there is a weaker negative correlation between volatility and MaxTuW and a weaker positive correlation between maxTuW and Volatility.

l) In order to create a linear regression model with the geometric mean as the dependent variable ($Y_i$), we first neeed to choose an explanitory variable. I have chosen to compare the geomatric mean to the Weekly Value-At-Risk, as this variable had the highest positive correlation with the geometric mean out of our given variables. They have a correlation of 0.4111. It is assumed that the residuals are normally distributed, as well as the residuals are independendt and identically distrubuted (i.i.d). Our linear regression model will be:

$$E(Y_i) = \beta_0 + \beta_1 * x_i \tag{30}$$

m) In order to calculate the least square estimates we can use the following formula:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{s_{xx}} \tag{31}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{32}$$

where $s_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$

We can insert our values for the geometric mean and Weekly Value-At-Risk:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - 0.07690)(x_i - (-4.6545))}{187.6348} = 0.02353 \tag{33}$$

$$\hat{\beta}_0 = 0.07690 - 0.02353 \cdot -4.6545 = 0.1864193 \tag{34}$$

Because our value for $\hat{\beta}_1$ estimate is positive, we can conclude that the slope is positive between the geometric mean and the weekly value-at-risk in our model. From our value of $\hat{\beta}_0$ we can conclude that even when there is no Weekly Value-At-Risk the geometric mean is still positive.

In order to calculate the model variance we use the following formula:

$$\hat{\sigma}^2 = \frac{\text{RSS}\left(\hat{\beta}_0, \hat{\beta}_1\right)}{n-2} \tag{35}$$

Where RSS($\hat{\beta}_0, \hat{\beta}_1$) is the sum of squared residuals.

With the values inserted it gives us the following:

$$\hat{\sigma}^2 = \frac{0.5108}{95-2} = 0.0005492 \tag{36}$$

The value of the model variance says how well our linear model fits.

These values have been verified in R. When we make the summary in R we also get other values such as our t and p values. with t-values of roughly 7 and 4 and very low p values, we can interpret there is a very low probability that there is no correlation between the geometric mean and the weekly value of risk.

n) Because of our high t values and low p values from the summary above we can be very sure that there is a correlation between the geometric mean and the weekly Value-at-Risk, as we had to reject that there is no correlation. This can also be confirmed in R, which gave the correlation of 0.4111.