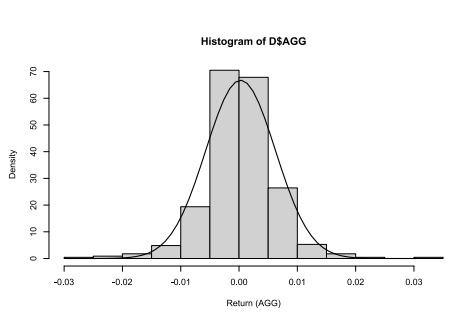
Trading With ETF

Fedir Vasyliev

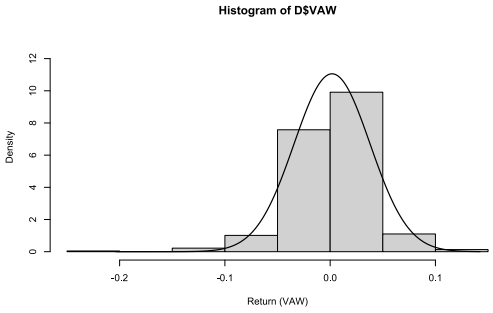
15/03/2005

1 Data Description

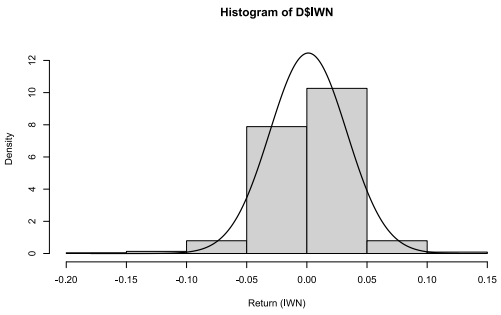
The data consists of the weekly returns for 95 ETFs. The first row contains the names of different ETFs. The columns are the actual return values of ETF with respect to the dates. Here the ETFs returns are a quantitative variable because it is numerical and represent a measurable quantity. Dates are the date variable but can be interpreted as a categorical variable if we group the data specifically by year, month, day, etc. There are 454 observations for each ETF. The first observation happened to be on $2006-5-5$ and the last one was on $2015-5-8$. Using $sum(is.na(D))$ can be seen that the data set has no missing values.



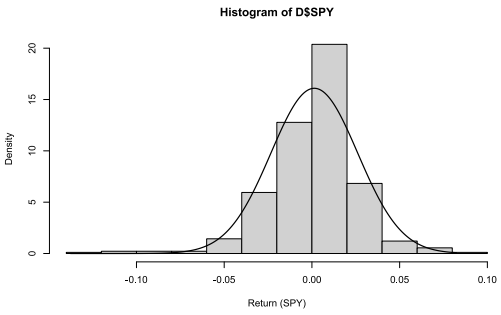
(a) Histogram of AGG



(b) Histogram of VAW



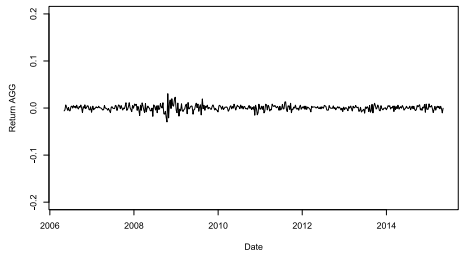
(c) Histogram of IWN



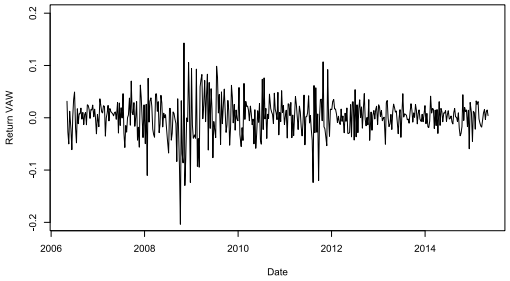
(d) Histogram of SPY

Figure 1: Histograms of different ETFs with simulated normal distribution curve with the mean and the standard deviation of the corresponding data set

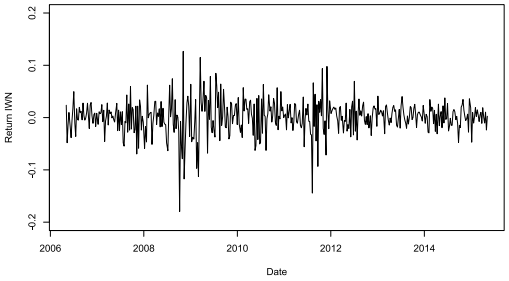
Each of the empirical density histograms looks symmetric, VAW histogram happens to have a longer left tail, with both positive and negative return rates. The bell curve of the density distribution of AGG is thin and tall which indicates lower variance. The density distribution of VAW is wider and one also can notice that return rates are larger than the ones of AGG and SPY, the variance is higher compared to the AGG's. The IWN's density distribution looks very similar to VAW. The histogram of SPY looks wider than the histogram of AGG, however, density distribution spread out more evenly from the mean.



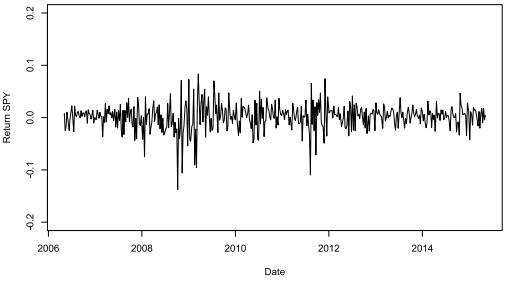
(a) Weekly return of AGG



(b) Weekly return of VAW



(c) Weekly return of IWN



(d) Weekly return of SPY

Figure 2: Plots that illustrate the weekly return over time for each of the four ETFs

To prove the previous point, one can look at the Figure above to see how weekly returns developed over time. It is easy to see volatility regions in return rates around 2009 and 2012. The return rates don't seem to change a lot over time. What is obvious is that the return rate of AGG is the lowest one, with VAW and IWN having the higher one, and SPY in the middle, however, higher return rates mean higher risks.

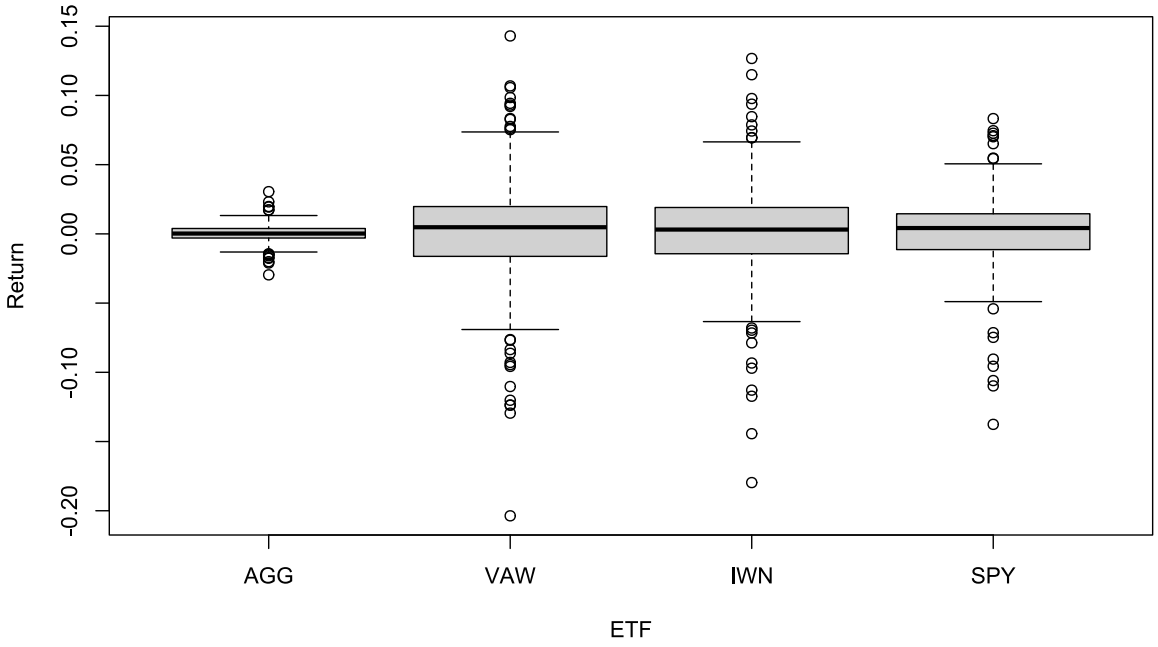


Figure 3: Box plot of weekly returns by ETF

AGG distribution appears to be relatively symmetrical with a small interquartile range, indicating low volatility. There is a small number of outliers, suggesting that most weekly returns are close to the median. VAW distribution seems slightly skewed to the right, with a larger interquartile range compared to AGG. There are several outliers on both ends. IWN is similar to VAW, the distribution for IWN also appears to be

slightly right-skewed and has multiple outliers on both ends. The interquartile range is similar to that of VAW, indicating a similar level of volatility. The distribution for SPY appears to be more symmetrical compared to VAW and IWN, with an interquartile range similar to them. It has multiple outliers on both ends, indicating occasional weeks with significantly higher or lower returns.

ETF	Number of obs. (n)	Sample mean ($\bar{x}$)	Sample variance (s^2)	Std. dev. (s)	Lower quartile ($Q1$)	Median ($Q2$)	Upper quartile ($Q3$)
AGG	454	0.0002658	3.571e-05	0.005976	-0.002973254	0.0002374461	0.003893193
VAW	454	0.001794	0.001302	0.03608	-0.01609575	0.004797925	0.01968522
IWN	454	0.001188	0.001025	0.03202	-0.01430519	0.003119637	0.01905639
SPY	454	0.001360	0.0006143	0.02479	-0.01132501	0.004215788	0.01449757

Table 1: Quantified data for each ETF

The table shows the number of observations for each ETF, it also states standard deviation and variance and provides exact values for quartiles and means.

2 Statistical Analysis

As one can see from Figure 1 by looking at the histogram and the normal distribution curve simulated based on the corresponding data set(with the same mean and variance). The random data samples for ETFs returns follow normal distribution. One can also

be interpret it as follows:

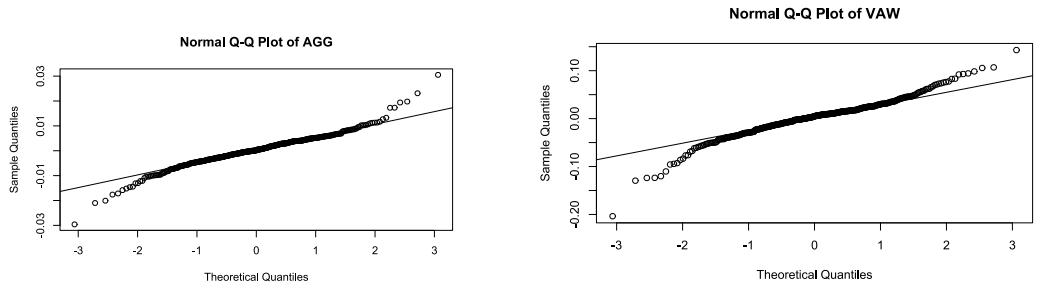
$$A_i \sim N(0.0002658, 3.571 \times 10^{-5}), \quad \text{where } i = 1, 2, \ldots, 454$$

$$V_i \sim N(0.001794, 0.001302), \quad \text{where } i = 1, 2, \ldots, 454$$

$$I_i \sim N(0.001188, 0.001025), \quad \text{where } i = 1, 2, \ldots, 454$$

$$S_i \sim N(0.001360, 0.0006143), \quad \text{where } i = 1, 2, \ldots, 454,$$

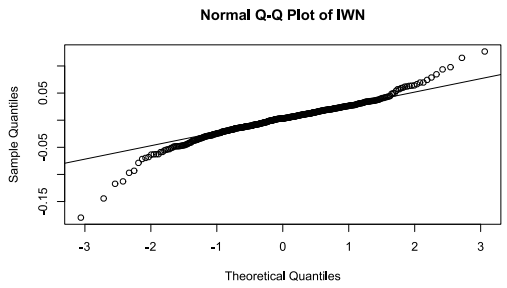
A_i, V_i, I_i, S_i are random variables of AGG, VAW, IWN, and SPY respectively. Hence the n random variables are independent and randomly distributed. So that is the assumption that we want to test. The statement about independence can only be assessed by knowing the sampling procedure, on the other hand, the assumption about normality can be assessed by performing data analysis on the given data sample. We have already seen simulated bell curves of the normal distribution in Figure1, which is the first evidence of normality, however, we won't stop on that note and go further by doing q-q plot for each sample. If the observations are normally distributed then the observed are close to the expected and this plot is close to a straight line. One can



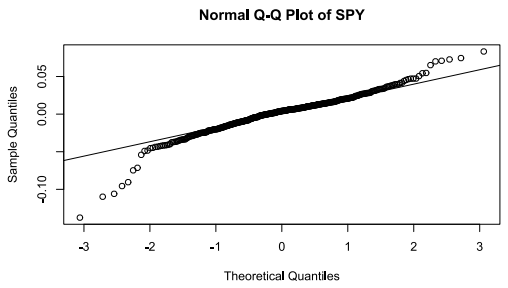
(a) qqPlot of AGG

(b) qqPlot of VAW

Figure 4: Normal Q-Q Plot for AGG and VAW



(c) qqPlot of IWN



(d) qqPlot of SPY

Figure 4: Normal Q-Q Plot for IWN and SPY

see from the plots above that points are close to the straight line, hence we can assume that normal distribution holds.

Going further in analyzing the whole population the analysis of the mean can be performed, specifically finding the confidence interval for the mean.

As follows from the Central Limit Theorem: for large enough n, which in our case is 454, it holds approximately that $\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim N(0,1^2)$, which means that margin of the mean error can now be calculated as $z_{1-\alpha/2}\frac{s}{\sqrt{n}}$. For AGG the confidence interval for the mean can be computed by

$$CI = \overline{\mu} \pm z_{1-\alpha/2}\frac{s}{\sqrt{n}} = 0.0002658 \pm qnorm(0.975)\frac{3.571 \times 10^{-5}}{sqrt(454)} =$$

$$= [-0.0002839348\ 0.0008154487]$$

$$CI = \overline{\mu} \pm t_{1-\alpha/2}\frac{s}{\sqrt{n}} = 0.0002658 \pm qt(0.975, df = 453)\frac{3.571 \times 10^{-5}}{sqrt(454)} =$$

$$= [-0.0002854073\ 0.0008169213]$$

The $(t.test(D\$AGG, \text{conf.level} = 0.95)$ can be used to check the values, and it yields to be $[-0.0002854073\ 0.0008169213]$. The CLT assumption works, but in order to prevent

any misunderstandings in communication the t distribution will be used in calculations of confidence intervals. The following table represents the lower and the upper bounds of CI for each ETF calculated using $CI = \overline{\mu} \pm t_{1-\alpha/2}\frac{s}{\sqrt{n}}$ formula.

	Lower bound of CI	**Upper bound of CI**
AGG	-0.0002854073	0.0008169213
VAW	-0.001534208	0.005121788
IWN	-0.001765174	0.004140533
SPY	-0.000925960	0.003646171

Table 2: Lower and upper bounds of CI for each ETF

$(t.test(D\$AGG, \text{conf.level} = 0.95)$ function obviously gives the same results as we've already got using a regular formula.

Now let's investigate whether the mean weekly return from AGG and the other ETFs deviates significantly from the return obtained by saving money under the pillow, that is gaining nothing.

To do it one can test the null hypotheses, to begin with, AGG:

$$H_0 : \mu_{AGG} = 0,$$

$$H_1 : \mu_{AGG} \neq 0$$

The following code finds the p-value for AGG with a significance level of $\alpha = 0.05$(we want our mean to fall within 95% interval across the mean of t distribution).

```
tobs_AGG <- AGG_mean/AGG_sd*sqrt(454)
p_val_AGG <- 2*(1-pt(abs(tobs_AGG), df=453))
```

```
p_VAL_AGG = 0.3438511
```

p-value for AGG $>$ significance level $\alpha = 0.05$, which means that the effect is not significant, in our case the effect is the difference between the sample mean and the zero mean, so we accept the null hypothesis.

```
t.test(D$AGG, mu=0)
```

gives a p-value of 0.3439 and what is also worth mentioning is that the p-value doesn't depend on the significance level. It is also worth mentioning that we could have already concluded whether the null hypotheses can be rejected or not of the **Confidence Interval**, it is obvious from the CI that the null hypothesis can be accepted as $\mu = 0$ is within the range of CI.

Is the mean weekly income of the AGG and VAW different? To answer this question one can perform The (Welch) two-sample t-test statistic. The null hypotheses will look like this:

$$\delta = \mu_{VAW} - \mu_{AGG}$$

$$H_0 : \delta = \delta_0 = 0$$

$$t_{obs} = \frac{(\mu_{AGG} - \mu_{VAW}) - \delta_0}{sqrt(s^2_{AGG}/n + s^2_{VAW}/n)}$$

$$\text{with } v = \frac{(s^2_{AGG}/n + s^2_{VAW}/n)^2}{(s^2_{AGG}/n)^2/(n-1) + (s^2_{VAW}/n)^2/(n-1)} \text{ degrees of freedom,}$$

$$p - value = 2 * P(T > |t_{obs}|)$$

```
tobs_mu = (AGG_mean - VAW_mean)/sqrt(AGG_sd^2/454 + VAW_sd^2/454)

v = (AGG_sd^2/454 + VAW_sd^2/454)^2/((AGG_sd^2/454)^2/453 + (VAW_sd^2/454)^2/453)

pval_mu = 2*(1-pt(abs(tobs_mu),df=v))
```

```
pval_mu

[1] 0.3738104
```

Since the p-value is $> \alpha = 0.05$ we accept the null hypothesis as we don't have enough evidence against it, which means that there is no significant difference between the weekly income from these two ETFs. And as we accepted the null hypothesis it is now hard to conclude which of the two ETFs has the highest level of return, even though computing the means of both indicates that VAW index has a higher mean.

```
> AGG_mean

[1] 0.000265757

> VAW_mean

[1] 0.00179379

t.test(D$VAW, D$AGG)
```

gives the same p-value as the one that has been computed above. The same result could be drawn from accepting the null hypothesis of both AGG and VAW which would automatically mean that $\delta_0 = 0$.

2.1 Correlation

Correlation between assets can be seen as the measure of the volatility of a portfolio, the lower the correlation the lower the risk. The correlation between the two samples can be found as follows

$$r = \frac{1}{n-1} \sum_{i=1}^{n} (\frac{x_i - \overline{x}}{s_x})(\frac{y_i - \overline{y}}{s_y})$$

Here is the R code that calculates the correlation between VAW and IWN indices

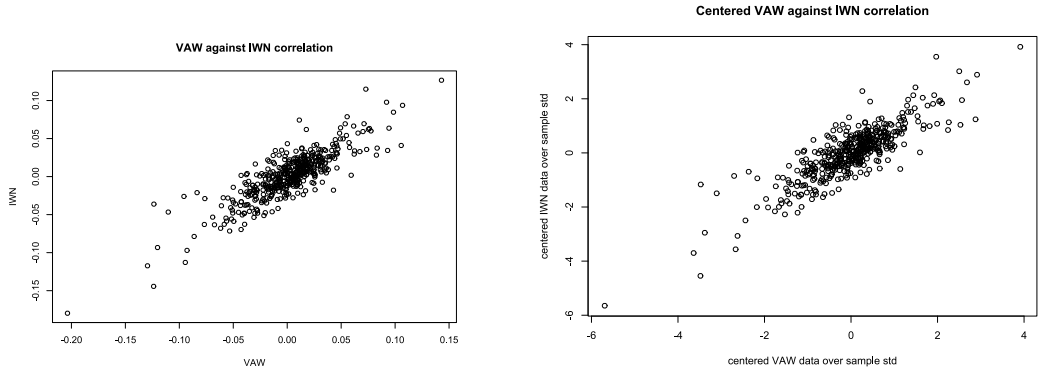
```
>VAW_centered_overSD <- (D$VAW - VAW_mean)/VAW_sd
> IWN_centered_overSD <- (D$IWN - IWN_mean)/IWN_sd
> IWN_VAW_corr <- sum(IWN_centered_overSD * VAW_centered_overSD)/453
> IWN_VAW_corr
[1] 0.8516407
```

To check the result the in-build function can be used

```
> cor(D[ ,c("VAW","IWN")], use="pairwise.complete.obs")
          VAW       IWN
VAW 1.0000000 0.8516407
IWN 0.8516407 1.0000000
```

Both correlations are the same and equal to 0.8516407. The value is lying between 0 and 1, which means the relation is rather linear but the points are not on the one line. The result can be visualized by the following plots

```
> plot(D$VAW, D$IWN, xlab="VAW", ylab="IWN", main="VAW against IWN correlation")
> plot(VAW_centered_overSD, IWN_centered_overSD, xlab="centered VAW data over sample s
```



(a) VAW against IWN correlation

(b) Centered VAW against IWN correlation

Figure 5: Correlation between VAW and IWN data samples

This result was expected because if we look at the weekly returns plots(Figure 2), we can see that the values of returns react in the same way to the market: i.e. when one index grows the second one also grows and vice versa. This behavior is the reason why it is less risky to have indices with negative correlation, negative correlation would mean the opposite behavior of indices: if one index falls, the other one will rise.