# Project 1: Trading ETFs

## Formalities, structure and expectations for the first mandatory project

The assignment consists of two parts. The first part focuses on descriptive analysis of the data. The second part is primarily about confidence intervals and hypothesis tests.

The assignment is formulated in such a way that it can be solved in small "easy" steps. In practice, the assignment must be solved using the statistical software R. Some R code is provided in order to make it easy to get started with the project. However, the code is not complete, and you are encouraged to explore new features in R while working on the project. For example, you could add suitable titles to the plots, or use R's built-in functions for computing confidence intervals and testing hypotheses.

The results of the analysis must be documented in the report using tables, figures, mathematical notation, and explanatory text. Relevant figures and tables must be included within the text, not in the appendix. Present the results of your analysis as you would when explaining them to one of your peers.

Divide the report into subsections, one for each of the questions to be answered.

The report must be handed in as a pdf file. R code should not be included in the report itself but must be handed in as an appendix (a .R-file). The report and appendix must be handed in under Assignments on Learn at: `Projekt 1: Handel med ETF`

The report text should not exceed 6 pages (excluding figures, tables, and the appendix). A normal page contains 2400 characters.

Figures and tables cannot stand alone - it is important that you describe and explain

the R output in words.

Figures and tables are not included in the assessment of the length of the report. However, it is not in itself an advantage to include many figures, if they are not relevant!

You may work together in groups, but the report must be written individually. Questions about the project can be addressed to the teaching assistants, see the guidelines on the `Projects` page of the course website.

## Introduction

This project focuses on the weekly returns from a selection of ETFs. An ETF (Exchange Traded Fund) can be described as a structured, publicly traded pool of shares. ETFs are bought and sold in the same way as ordinary shares on a stock exchange. An ETF is a pooled investment fund, similar to a unit trust or mutual fund. For investors, ETFs combine the benefits of pooled funds and shares. If you buy, for example, a simple ETF which covers the SP100 Index in the United States, it is equivalent to owning a part of all 100 stocks in the index. Thus, you avoid having to buy 100 individual securities, instead, you can just purchase one.

There are many different ETFs – actually, the market for ETFs is under explosive development. There are also various strategies under which the ETFs are administered. An ETF with a passive strategy seeks to track the return of the underlying index as closely as possible. That is, it aims to provide the investors with the same return as the underlying market. Such an ETF is called an index fund. The EURO STOXX Index of leading Eurozone company shares is an example of such a fund. For example, if the EURO STOXX50 Index increases by 10% in one year, an ETF tracking this index aims to provide investors with the same return, minus fees which, in the case of an ETF, are referred to as the Total Expense Ratio (TER). In order to be able to deliver the same return as a market index, the ETF holds all the index constituents, or a representative subset of these.

The advantages of ETFs, compared to e.g. investment funds, are flexibility, cost effectiveness, and high liquidity. That is, the ETFs are cheap compared to other investment products. However, as with all types of investments, there are risks involved with the buying and selling of ETFs. The value of the investment can go down, and the money invested can be lost. [1]

[1] The above sections were written based on information from: http://www.ishares.com and https://falconinvest.dk/hvad-er-en-etf/.

## Reading the data into R

Make a folder for the project on your computer. Download the project material from Learn and unzip it to the folder that you just made.

Then, open the data file `finans1_data.csv` (e.g., in RStudio, File → Open File) in order to see the contents of the file. Note that the first row (referred to as a *header*) contains variable names, and that the subsequent rows contain the actual observations. Variable names and observations of the individual variables are separated by a ';' (therefore `.csv`: "comma separated values", though here it is a semi-colon).

The dataset contains weekly returns (ratio between the final and initial price for the week in question minus 1) for 95 ETFs. There are 96 columns in the dataset. The first column is a date column, while each of the remaining 95 columns contains the weekly returns from one ETF. The column name specifies the name of the ETF in question.

Open the file `finans1_english.R`, which contains some R code that can be used for the analysis. First, the "working directory" must be set to the directory on the computer, which contains the files for the project:

```
## In RStudio the working directory is easily set via the menu
## "Session -> Set Working Directory -> To Source File Location"
## Note: In R only "/" is used for separating in paths
## (i.e. no backslash).
setwd("Replace with path to directory containing project files.")
```

Now the data may be read into R using the following code:

```
## Read data from finans1_data.csv
D <- read.table("finans1_data.csv", header=TRUE, sep=";", as.is=TRUE)
## Keep only the dates and the ETFs AGG, VAW, IWN, and SPY
D <- D[ ,c("t","AGG","VAW","IWN","SPY")]
```

`D` becomes a "data.frame" (a kind of table), which contains the data that was read into R (see the introduction to R in Section 1.5 of the book). Four different ETFs, *AGG*, *VAW*, *IWN*, and *SPY*, are selected. These are the only ETFs that will be used for the further analysis in this project. See Appendix 1 on p. 10 for a description of the selected ETFs. The file `ETF_dokumentation.xls` contains descriptions of all 95 ETFs.

## Descriptive analysis

The purpose of the first part of the project is to carry out a descriptive analysis of the data. In a report it is important to present the data and describe it to the reader. For example, this can be done using summary statistics and suitable figures.

Start by running the following commands to get a simple overview of the data:

```
## Dimensions of D (number of rows and columns)
dim(D)
##  Column/variable names
names(D)
## The first rows/observations
head(D)
## The last rows/observations
tail(D)
## Selected summary statistics
summary(D)
## Another type of summary of the dataset
str(D)
```

a) Write a short description of the data. Which variables are included in the dataset? Are the variables *quantitative* and/or *categorized* (or date variables)? (Categorized variables are introduced in Chapter 8, but they are simply variables which divide the observations into categories/groups - e.g. three categories: low, medium, and high). How many observations are there? Which time period is covered by the observations (date of first and last observations)? Are there any missing values?

The following code may be used to generate a "density histogram" describing the empirical density of the weekly returns from the ETF AGG (see Section 1.6.1):

```
## Histogram describing the empirical density of the weekly returns from
## AGG (histogram of weekly returns normalized to have an area of 1)
hist(D$AGG, xlab="Return (AGG)", prob=TRUE)
```

b) Make a density histogram of the weekly returns from the ETF AGG. Use this histogram to describe the empirical distribution of the returns. Is the empirical density symmetrical or skewed? Can the returns be both positive and negative? Is there much variation to be seen in the observations?

Note: In a *skewed* distribution, the probability mass is not symmetrically distributed around the median. In a left-skewed distribution, the left tail is longer than the right tail (and, typically, the mean will lie to the left of the median). Similarly, in a right-skewed distribution, the right tail is the longer of the two (usually, with the mean to the right of the median).

When observations are recorded regularly over time, the data is often referred to as a *time series*. Thus, the data from each ETF constitutes a time series. For time series, it is often relevant to make figures illustrating the data over time. Here, it is first necessary to tell R that the variable `t` should be treated as a date variable. This can be done using the following code:

```
## Converts the variable 't' to a date variable in R
D$t <- as.Date(x=D$t, format="%Y-%m-%d")
## Checks the result
summary(D$t)
```

Plots illustrating the weekly returns over time for each ETF can be made using the following R code:

```
## Plots of weekly return over time for each of the four ETFs
ylim <- c(-0.2,0.2)
## Plot of weekly return over time for AGG
plot(D$t, D$AGG, type="l", ylim=ylim, xlab="Date", ylab="Return AGG")
## Similar plots for the three other ETFs
plot(D$t, D$VAW, type="l", ylim=ylim, xlab="Date", ylab="Return VAW")
plot(D$t, D$IWN, type="l", ylim=ylim, xlab="Date", ylab="Return IWN")
plot(D$t, D$SPY, type="l", ylim=ylim, xlab="Date", ylab="Return SPY")
```

c) Make plots illustrating the weekly return over time for each of the four ETFs. Use these plots to describe the development of the weekly returns in words. Does the level of return seem to change over time? Are there specific periods of time during which the weekly returns are notably different? Are there overall differences in the returns from the different ETFs?

The following R code makes a box plot of the weekly returns by ETF:

```
## Box plot of weekly returns by ETF
boxplot(D$AGG, D$VAW, D$IWN, D$SPY, names=c("AGG", "VAW", "IWN", "SPY"),
        xlab="ETF", ylab="Return")
```

d) Make a box plot of the weekly returns by ETF. Use this plot to describe the empirical distribution of the weekly returns from each of the four ETFs. Are the distributions symmetrical or skewed? Does there seem to be a difference between the distributions (if so, describe the difference)? Are there extreme observations/outliers?

The empirical distribution of the weekly returns from each of the four ETFs may also be quantified using summary statistics as in the following table:

| **ETF** | **Number of obs.** | **Sample mean** | **Sample variance** | **Std. dev.** | **Lower quartile** | **Median** | **Upper quartile** |
|---|---|---|---|---|---|---|---|
| | $n$ | $(\bar{x})$ | $(s^2)$ | $(s)$ | $(Q_1)$ | $(Q_2)$ | $(Q_3)$ |
| AGG | | | | | | | |
| VAW | | | | | | | |
| IWN | | | | | | | |
| SPY | | | | | | | |

R code like the following may be used to fill in the empty cells in the table (see also the remark on p. 13):

```
## Total number of observations
## (doesn't include missing values if there are any)
sum(!is.na(D$AGG))
## Sample mean for weekly returns from AGG
mean(D$AGG, na.rm=TRUE)
## Sample variance for weekly returns from AGG
var(D$AGG, na.rm=TRUE)
## etc.
##
## The argument 'na.rm=TRUE' ensures that the statistic is
## computed even in cases where there are missing values.
```

e) Fill in the empty cells in the table above by computing the relevant summary statistics for each of the four ETFs. Which additional information may be gained from the table, compared to the box plot?

# Statistical analysis

The purpose of the second part of the project is to perform a simple statistical analysis of the weekly returns. This includes specifying statistical models for the weekly returns, estimating the parameters of these models, performing hypothesis tests, and computing confidence intervals.

## Confidence intervals and hypothesis tests

The following R code may be used to make a qq-plot. This plot can be used to investigate whether AGG's weekly returns may be assumed to be normal distributed:

```
## qq-plot of AGG's weekly returns
qqnorm(D$AGG)
qqline(D$AGG)
```

f) Specify separate statistical models describing the weekly return for each of the four ETFs (see Remark 3.2). Estimate the parameters of the models (mean and standard deviation). Perform model validation (see Chapter 3 and Section 3.1.8). Since, in this case, confidence intervals and hypothesis tests involve the distribution of an average, it might also be useful to include the central limit theorem (Theorem 3.14) in the discussion.

In practice, situations will arise where it is *not* appropriate to assume that the assumptions of a model are satisfied. In these cases, one often considers whether a transformation of the data might improve the situation. (See Section 3.1.9). Note that after a transformation, the interpretation of the results on the original scale changes. In this specific project, however, the intention is *not* for you to transform the data.

g) State the formula for a 95% confidence interval (CI) for the mean weekly return from AGG (see Section 3.1.2). Insert values and calculate the interval. Compute corresponding intervals for the three remaining ETFs and fill in the table below.

| | **Lower bound of CI** | **Upper bound of CI** |
|---|---|---|
| AGG | | |
| VAW | | |
| IWN | | |
| SPY | | |

Compare the CI for AGG computed above with the result of the following R code:

```
## CI for the mean weekly return from AGG
t.test(D$AGG, conf.level=0.95)$conf.int
```

h) Perform a hypothesis test in order to investigate whether the mean weekly return from AGG deviates significantly from the return obtained by saving money under the pillow (that is, nothing). This can be done by testing the following hypothesis:

$$H_0 : \mu_{\text{AGG}} = 0,$$
$$H_1 : \mu_{\text{AGG}} \neq 0.$$

Specify the significance level $\alpha$, the formula for the test statistic, as well as the distribution of the test statistic (remember to include the degrees of freedom). Insert relevant values and compute the test statistic and $p$-value. Write a conclusion in words. In particular, comment on whether it was necessary to perform the hypothesis test, or whether the same conclusion could have been reached using the confidence interval for AGG from before.

Compare the results of the hypothesis test with the results of the following R code:

```
## Testing the hypothesis mu=0 for weekly AGG returns
t.test(D$AGG, mu=0)
```

Now, we would like to investigate whether the weekly returns from VAW and AGG differ.

i) Perform a hypothesis test in order to investigate whether the mean weekly return differs between the two ETFs VAW and AGG. Specify the hypothesis as well as the significance level $\alpha$, the formula for the test statistic, and the distribution of the test statistic (remember the degrees of freedom). Insert relevant values and compute the test statistic and $p$-value. Write a conclusion in words. In particular: Is there a significant difference between the daily returns from VAW and AGG? If so, which of the two ETFs has the highest level of return?

Compare the results of the hypothesis test with the results of the following R code:

```
## Comparing the mean weekly returns from VAW and AGG
t.test(D$VAW, D$AGG)
```

j) Comment on whether it was necessary to carry out the statistical test in the previous question, or if the same conclusion could have been drawn using the confidence intervals from before? (See Remark 3.59).

## Correlation

In connection with the construction of a portfolio of ETFs, diversification of risk is a key concept. Popularly speaking, it concerns "not putting all your eggs in one basket". Risk can be measured in several ways - for example, it may be quantified using the standard deviation of the weekly returns.

When constructing a portfolio of ETFs, pairwise correlations between the ETFs are an essential tool in determining how much to invest in the various ETFs. Generally, the lower the correlation between ETFs which are combined, the higher the diversification of risk. Thus, a combination of ETFs with negative correlation can contribute to decreasing the volatility (i.e. the risk) of the total portfolio.

k) State the formula for computing the correlation between the two ETFs VAW and IWN. Insert values and determine the correlation (note, insert only numbers in the correlation formula, i.e. three numbers). Make a scatter plot illustrating the weekly returns for these two ETFs. Assess whether the relation between the plot and the correlation is as you would expect.

Compare the correlation computed above to the result of the following R code:

```
## Computing the correlation between selected ETFs
cor(D[ ,c("AGG","VAW","IWN","SPY")], use="pairwise.complete.obs")
```

# Appendix 1 Description of the ETFs

The table below shows an overview and description of the 4 selected ETFs.

The Excel file `ETF_dokumentation.xls` contains a description of all the ETFs. The following table was taken from this file.

| ETF | Description |
|---|---|
| AGG | iShares Core Total US Bond Market ETF, formerly iShares Lehman Aggregate Bond Fund (the Fund) seeks investment results that correspond generally to the price and yield performance of the total United States investment-grade bond market as defined by the Lehman Brothers U.S. Aggregate Index (the Index). The Index measures the performance of the United States investment-grade bond market, which includes investment-grade United States Treasury bonds, government-related bonds, investment-grade corporate bonds, mortgage pass-through securities, commercial mortgage-backed securities and asset-backed securities that are publicly offered for sale in the United States. The securities in the Index must have at least one year remaining to maturity. In addition, the securities must be denominated in United States dollars, and must be fixed rate, non-convertible and taxable. The Index is market capitalization weighted. The Fund uses a representative sampling strategy to track the Index. The Fund's investment advisor is Barclays Global Fund Advisors (BGFA). |
| VAW | Vanguard Materials ETF (the Fund), formerly known as Vanguard Materials VIPERs, is an exchange-traded share class of Vanguard Materials Index Fund, which employs a passive management or indexing investment approach designed to track the performance of the Morgan Stanley Capital International (MSCI) US Investable Market Materials Index (the Index). The Index is an index of stocks of large, medium and small United States companies in the materials sector, as classified under the Global Industry Classification Standard (GICS). This GICS sector is made up of companies in a range of commodity-related manufacturing industries. Included within this sector are companies that manufacture chemicals, construction materials, glass, paper, forest products and related packaging products, as well as metals, minerals and mining companies, including producers of steel. The Fund attempts to replicate the Index by investing all, or substantially all, of its assets in the stocks that make up the Index, holding each stock in approximately the same proportion as its weighting in the Index. The Fund also may sample its target index by holding stocks that, in the aggregate, are intended to approximate the Index in terms of key characteristics, such as price/earnings ratio, earnings growth and dividend yield. |
| IWN | iShares Russell 2000 Value Index Fund (the Fund) seeks investment results that correspond generally to the price and yield performance of the Russell 2000 Value Index (the Index). The Index measures the performance of the small-capitalization value sector of the United States equity market. It is a subset of the Russell 2000 Index. The Index is a capitalization-weighted index and consists of those companies or portion of a company, with lower price-to-book ratios and lower forecasted growth within the Russell 2000 Index. The Index represents approximately 50% of the total market capitalization of the Russell 2000 Index. The Fund invests in a representative sample of securities included in the Index that collectively has an investment profile similar to the Index. iShares Russell 2000 Value Index Fund's investment advisor is Barclays Global Fund Advisors. |
| SPY | SPDR Trust, Series 1 (the Trust) is a unit investment trust. The Trust is an exchange-traded fund created to provide investors with the opportunity to purchase a security representing a proportionate undivided interest in a portfolio of securities consisting of substantially all of the common stocks, in substantially the same weighting, which comprise the Standard and Poor's 500 Composite Price Index (the SP Index). Each unit of fractional undivided interest in the Trust is referred to as a Standard and Poor's Depositary Receipt (SPDR). The Trust utilizes a full replication approach. With this approach, all 500 securities of the Index are owned by the Trust in their approximate market capitalization weight. |

**Remark 2.1 Extra R tips**

This is an optional extra remark about different ways to take subsets in R (useful but not necessary for solving the project):

```
## Optional extra remark about taking subsets in R
##
## A logical vector with a TRUE or FALSE for row value in D.
## E.g.: The weeks with losses (negative returns) from AGG
D$AGG < 0
## Can be used to extract all AGG losses
D$AGG[D$AGG < 0]
## Alternatively, use the 'subset' function
subset(D, AGG < 0)
## More complex logical expressions can be made, e.g.:
## Find all observations from 2009
subset(D, "2009-01-01" < t & t < "2010-01-01")
```

### Remark 2.2 Extra R tips

Optional remark with some extra R tips. The table can also be generated more effectively using a 'for'-loop:

```
## Use a 'for'-loop to calculate the summary statistics
## and assign the result to a new data.frame
num <- 2:5
Tbl <- data.frame()
for(i in num){
  Tbl[i-1,"mean"] <- mean(D[ ,i])
  Tbl[i-1,"var"] <- var(D[ ,i])
}
row.names(Tbl) <- names(D)[num]
## View the contents of Tbl
Tbl

## In R there are even more condensed ways to perform such
## calculations, e.g.:
apply(D[, num], 2, mean, na.rm=TRUE)
## or several calculations in one expression
apply(D[, num], 2, function(x){
  c(mean=mean(x, na.rm=TRUE),
    var=var(x, na.rm=TRUE))
})
## See more useful functions with: ?apply, ?aggregate and ?lapply
## For extremely efficient data handling see, e.g., the packages:
## dplyr, tidyr, reshape2 and ggplot2

## LaTeX tips:
##
## The R package "xtable" can generate LaTeX tables written to a file
## and thereby they can automatically be included in a .tex document.
##
## The R package "knitr" can be used very elegantly to generate .tex
## documents with R code written directly in the document. This
## document and the book were generated using knitr.
```