Project: Trading with ETF

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 Python. Some Python 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 while working on your code for the project. For example, you could add suitable titles to the plots, or 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. Code should not be included in the report itself but must be handed in as an appendix (a .py or .ipynb file). The report and appendix must be handed in under Assignments on Learn under Opgaver/Assignments.

The report text should not exceed 9 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 Python 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.

The problem

In this project, the weekly returns for a selection of ETFs are analyzed and modelled. 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 blend the benefits of pooled funds and shares.

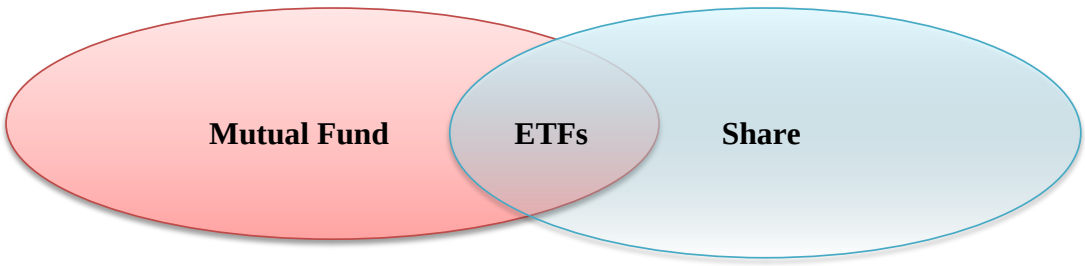

Figure 1.1: Illustration of the ETF concept.

If you buy for example a simple ETF covering the SP100 Index in the United States it is equivalent to owning a part of all 100 stocks in the index. Thus you avoid buying 100 individual securities and can instead just buy a single.

There are many different ETFs - actually the ETF market is under explosive development. Various strategies are available - for example passive and active approach under which EFTs are administered.

An ETF with passive strategy seek to track the underlying index return as close as possible. Such an ETF is called an index fund. An example would be the EURO STOXX Index of leading Eurozone company shares. This means that the aim of an ETF is to provide investors with the same return as the underlying market. For example, if the EURO STOXX50 Index goes up by 10% during a year, an ETF tracking this index aims to provide investors with the same return, minus fees, which in the case of an ETF is called the Total Expense Ratio (TER). To deliver the same return as the market index ETFs hold all the index constituents, or a representative subset of the index constituents.

The advantage of ETFs compared to e.g. investment funds are flexibility, cost-effectiveness and high liquidity. The EFTs are cheap compared to other investment products.

Remember: as with all other types of investments there are risks involved with buying and selling ETFs. Your investment can go up and down and the amount invested can be lost, see [1].

The available data is found in the file `finans1_data.csv` and consists of 96 columns. The first column is a date column and the subsequent 95 columns indicates the weekly returns (i.e. the ratio between the final and initial price for that week minus 1) of 95 ETFs. The column name specifies the name of the ETF.

Descriptive analysis

The first part of the project is to perform a descriptive analysis to examine the data and its quality, in order to get some knowledge about methods and models we can use to analyze the data.

Make a folder on your PC for the project, download and unzip the material into this folder.

First open the data file `finans1_data.csv` for example in Visual Studio Code and see what is in the file. It is seen that the first line contains variable names

[1] The above sections are written based on the following references: http://www.ishares.com/dk/private/da/literature/brochure/brochure-introducing-ishares-and-etfs-en.pdf and https://falconinvest.dk/hvad-er-en-etf/

and that subsequent lines are the actual observations. Variable names and observation values are separated by a "," (comma).

You have to make a Python-script to be attached to the assignment as an appendix to document the performed analysis. Open the file `finans3_english.ipynb`, there is a template for the Python-script (Jupyter Notebook) you have to produce.

First set the "working directory" to where the script and data file is located on your computer. See "Read Data" in the Jupyter notebook. `D` is a pandas DataFrame containing the data. (see the Python intro in Chapter 1 of the eNotes)

To get an overview of the data first run the command in "Simple summary of data" in the Jupyter notebook.

a) Make a short description of the data - how many observations does the data-set contain, which period is covered, when is the first observation and when is the last observation recorded, how is the data quality e.g. in the form of missing observations? etc.

In the following, we select 4 different ETFs that we will continue to analyze. A description of the selected ETFs listed in the table in Appendix 1.

The Excel file `ETF_dokumentation.xls` contains a description of all the ETFs. The table in Appendix 1 is taken from this file.

b) Examine the empirical distribution/density of each of the 4 EFTs, as a minimum, plot the empirical density and box plots (see eg. Ex1-26 and 1.28).
Further, fill out the following table:

EFT	(Number of obs.)	(Sample mean)	(Sample variance)	(Std. dev.)	(Lower quartile)	(Median)	(Upper quartile)
AGG							
VAW							
IWN							
SPY							

For filling out the table you may run the commands in "Descriptive analysis of selected variables" in the notebook.

c) Based on the above analysis, briefly describe the distribution of the weekly returns for each of the 4 ETFs. Are they symmetrical or skewed, and if they are skewed, what kind of skewness. Describe interesting details, including account for extremes / outliers. What values can the weekly returns assume (state perhaps minimum and maximum) and is it as you expect?

Note, that a density is skewed if the probability mass is unevenly distributed (i.e. not symmetrical. If it is a *left-skewed* distribution, the longer tail is located to the left of the median (in general, the mean is located to the left of the median) and similarly for a *right-skewed* distribution the longest tail to the right of the median (in general, the mean is located to the right of the median).

Statistical analysis I

We will now start at the statistical analysis and make statistical inference, hence we will analyze portfolios, formulate models, test hypotheses and determine confidence intervals.

Problem 1 - ETF Portfolio

In relation with the construction of a portfolio of ETFs diversification of risk is a key concept. It's about "not putting all your eggs in one basket". Risk can be measured in several ways - e.g. the standard deviation of the weekly returns. Another used measure of risk is the concept of volatility, which is the standard deviation of the logarithm of the weekly returns (also the same as log (return)).

When you construct a portfolio of ETFs, the covariance between the various ETFs is an essential tool to determine how to allocate your investment between the ETFs (how much you want to invest in the various ETFs).

d) Determine the covariance and correlation between the following ETFs: AGG, VAW and IWN. Use the Python code in "Covariance between ETFs" in the notebook.

e) Make pairwise scatterplot of all combinations of the four ETFs, and comment on the results and relate to the correlations calculated in the previous question.

f) Make a portfolio of two ETFs so that the variance of the portfolio is minimized doing it in following steps:

1. Let P_1 be a random variable that describes the portfolio consisting of EFT'erne: AGG og VAW: $P_1 = \alpha \cdot X_{AGG} + (1 - \alpha) \cdot X_{VAW}$, where X_{AGG} respectively X_{VAW} are random variable that indicates the weekly returns for AGG respectively VAW. α specifies the proportion of the portfolio invested in AGG.

   Define corresponding random variables, describing portfolios with all combinations of ETFs from the four ETF considered: AGG, VAW, IWN, and SPY.
2. Determine an expression for $var(P_1)$ (see Remark 2.59 and Theorem 2.60).
3. Determine $var(P_1)$ as a function of α (dvs. $V(\alpha)$), where the values of $var(X_{AGG})$, $var(X_{VAW})$ and $cov(X_{AGG}, X_{VAW})$ (determined in question d)) are inserted.
4. Make a graph of $V(\alpha)$.
5. Determine α_m, that provides minimum variance of P_1 - Consider the monotony conditions for the function. What is the implication of $\alpha_m > 1$ or $\alpha_m < 0$?
6. Make a table of all α_m -values for the investigated portfolio combinations.
7. Make a table of minimum variance for the investigated portfolio combinations.
8. Make a table of the expected weekly returns for the portfolios with minimum variance ($E(P_i)$).
9. Choose the optimal portfolio - arguments for your choice of portfolio.

Problem 2 - Best investment

In this part we will investigate what the best investment is: saving your money under the pillow or investing in one of the 4 selected ETFs.

g) Set up models in which we can assess the weekly returns for each of the 4 ETFs. State the assumptions of the model. Estimate the model parameters and perform model control.
Carry out a model validation, i.e. examine if the assumptions of the test are fulfilled (see Section 3.1.8).
The Python code in "Model validation" in the Jupyter notebook can be used for validation of model.
Assess whether assumptions are fulfilled. Remember to also include the Central Limit Theorem in the assessment (Theorem 3.14).
If the normal distribution assumption is not met (important if the distribution is highly skewed and sample size is small) a transformation of data should be considered - typically by the logarithm function. Since this data is financial, it could be considered to investigate the geometric average rather than the arithmetic average. In this assignment you can omit these considerations.

h) Determine a 95% confidence interval for the average weekly return for each of the 4 ETFs.State formulas and insert numbers. Describe the confidence intervals, e.g.: Do they have the same width? If they do not have the same width, what is the reason?

Compare your results of the determination of confidence intervals with the results from the Python code in "Confidence intervals" in the Jupyter notebook.