**Confidence Intervals and Hypothesis Testing**

This breakout session has two parts on the statistical concepts of confidence intervals and hypothesis testing. These are important concepts when estimating an average value for a whole population based on results from a sample. For example, if we would like to know the average spending of our supermarket customers and we cannot ask all the customers, only few of them, then we would like to understand how accurately the mean of this sample of customers estimates the mean of all our customers. The 95% confidence interval gives us a range around the sample mean such that we are 95% confident that the mean spending of all customers is within this range.

If we would like to find out whether a statement about the mean spending of customers is likely to be true or not, e.g. “the mean spending of all customers is £50 per visit”. Then by applying the concept of hypothesis testing, we can either reject or not reject this hypothesis, with a specific significance level, e.g. 5% (which corresponds to 95% confidence).

Format:

Part I consists of short exercises and Part II is a short workshop to analyse financial returns using confidence intervals and hypothesis testing. Please work in your group and discuss your findings. If you have any problems with the following exercises, please ask the lecturer or the workshop tutor for help.

**Part I: Short Exercises**

**Exercise 1**

The financial return of a commodity is estimated to be normally distributed with a mean of 1% and standard deviation of 1.5%. Find the probabilities that the return will be

a) at least 2.5%

b) between –0.5% and 4%

c) less than -1%

d) between 0% and 2%

**Exercise 2**

True or false:

a) If we know the mean of the normal distribution we know its shape

b) In the normal distribution, the mean, median, mode, and standard deviation are all at the same position on the horizontal axis since the distribution is symmetric.

c) Normal distribution that is wide has a high standard deviation

**Exercise 3**

A survey involves selecting a random sample of 256 middle managers for study. One item of interest is annual income. The sample mean is computed to be £35,420, and the sample standard deviation is £2,050.

a) What is the estimated mean income of all middle managers (the population)?

b) Give a 95 percent confidence interval (rounded to the nearest £10) for your estimate of the mean income. Do you have to make any assumptions?

c) Interpret the meaning of the confidence interval.

**Exercise 4**

Please discuss or write down the answers to the following questions:

a) If you collect 4 times more data, how much narrower will your confidence interval (CI) be? Same question for collecting 100 times more data.

b) Assume you work for a manager who says one day "I got the budget to collect twice as much data; that's great because our estimates will be twice as precise." Is anything wrong with his statement?

c) Your manager says "Let's just calculate our CIs with 90% coverage probability instead of 95%; this will make the CIs narrower." Is she right or wrong? Your manager adds: "We get better precision this way." What is the manager's misconception?

**Exercise 5:** The processing times for a particular product follow a normal distribution with an assumed mean of 40 minutes, and a standard deviation of 5 minutes. Suppose we want to test that the process is under control using a 5% significance test of an observed sample mean.

A sample of 100 units yields an average processing time of 41.2 minutes.

i) Is the machine producing at its expected speed? (test at 5% significance level)

ii) Is the assumption that the processing times follow a normal distribution necessary?

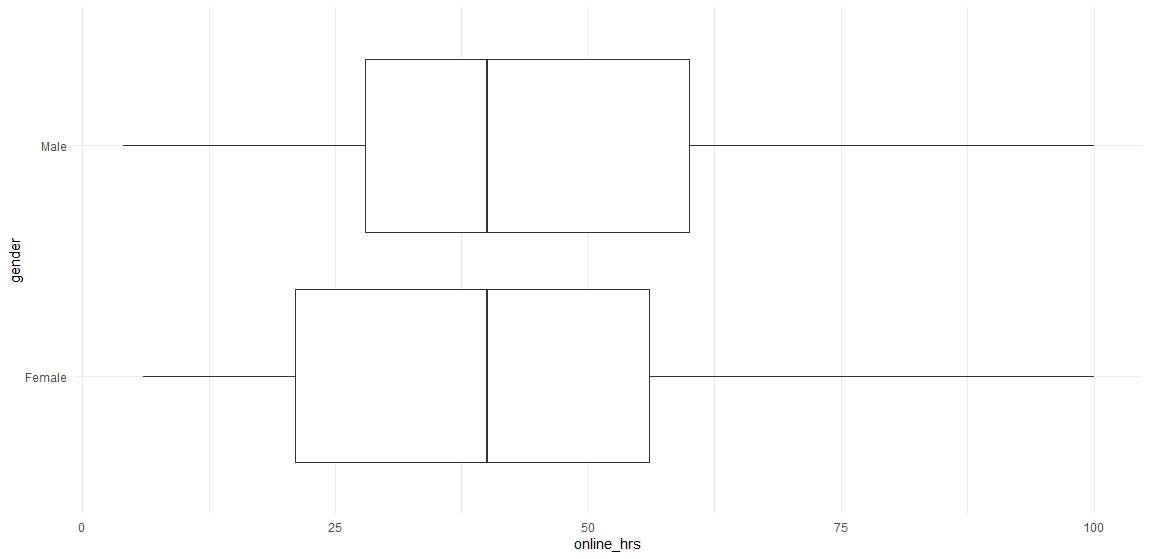
**Exercise 6:** Supporters claim that a new wind turbine can generate an average of at least 800 kilowatt-hours (kWh) of power per day. A random sample of 9 days is taken to test this claim and the sample mean is 776 kWh and the sample standard deviation is 48 kWh.

* Can you test whether the sample mean you measured (776 kWh) is statistically different from the claimed mean (800 kWh)? Assume 5% significance.
* What sample mean value would have you 95% convinced that the wind turbine is \*NOT\* producing 800 kWh?

**Exercise 7**: The following table contains summary statistics from the LBS early careers (EC) students. We want to check whether there is a significant difference between EC men and women in how much time they spend online, how much time they exercise, and how many facebook friends they have

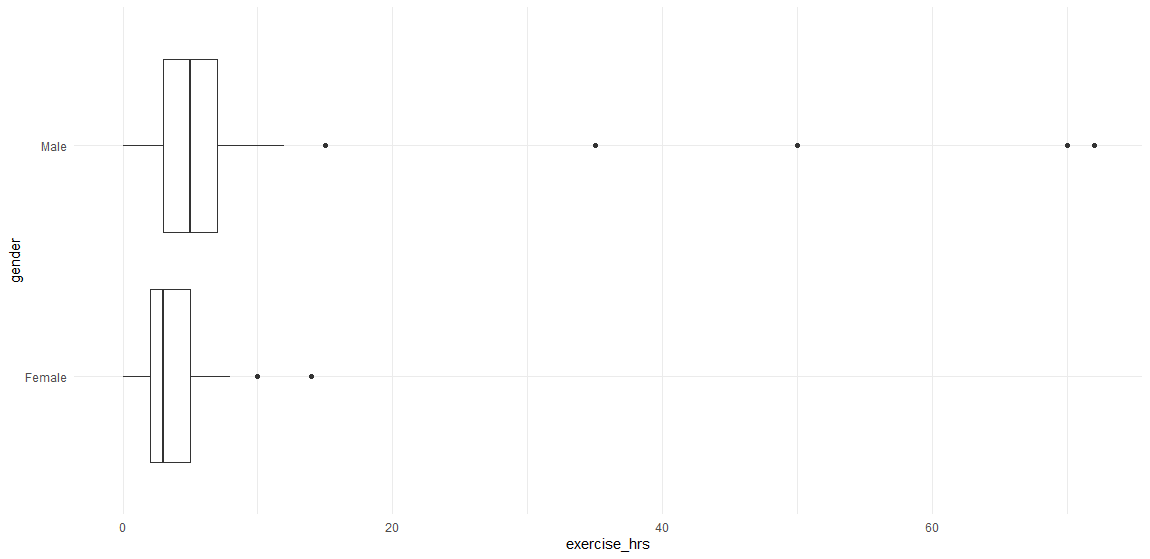
**Online Hours**





**Exercise Hours**

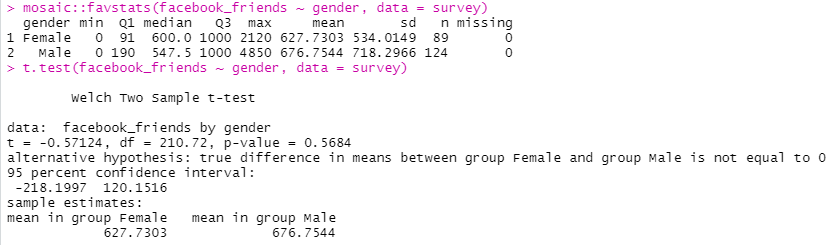


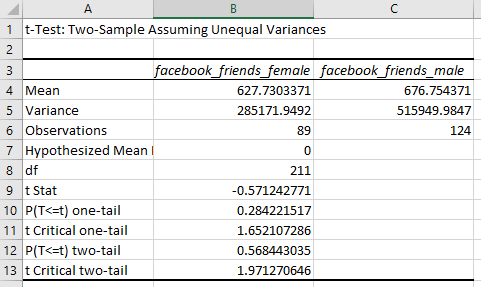


1. Before you go on, do you think there is any difference between how much time EC men and women spend online? Or time spend exercising?
2. Now, write down the null and alternative hypothesis, first for online and then for exercise hours
3. Use the formula to calculate the SE for the difference between two means
4. Calculate the t-stat, namely how far away is the actual difference you measured in your sample, from the difference you assumed

For facebook friends, we have run a t test within R and Excel and show you both outputs. What do you conclude? Men have on average 49 (=676.7 – 627.7) more facebook friends

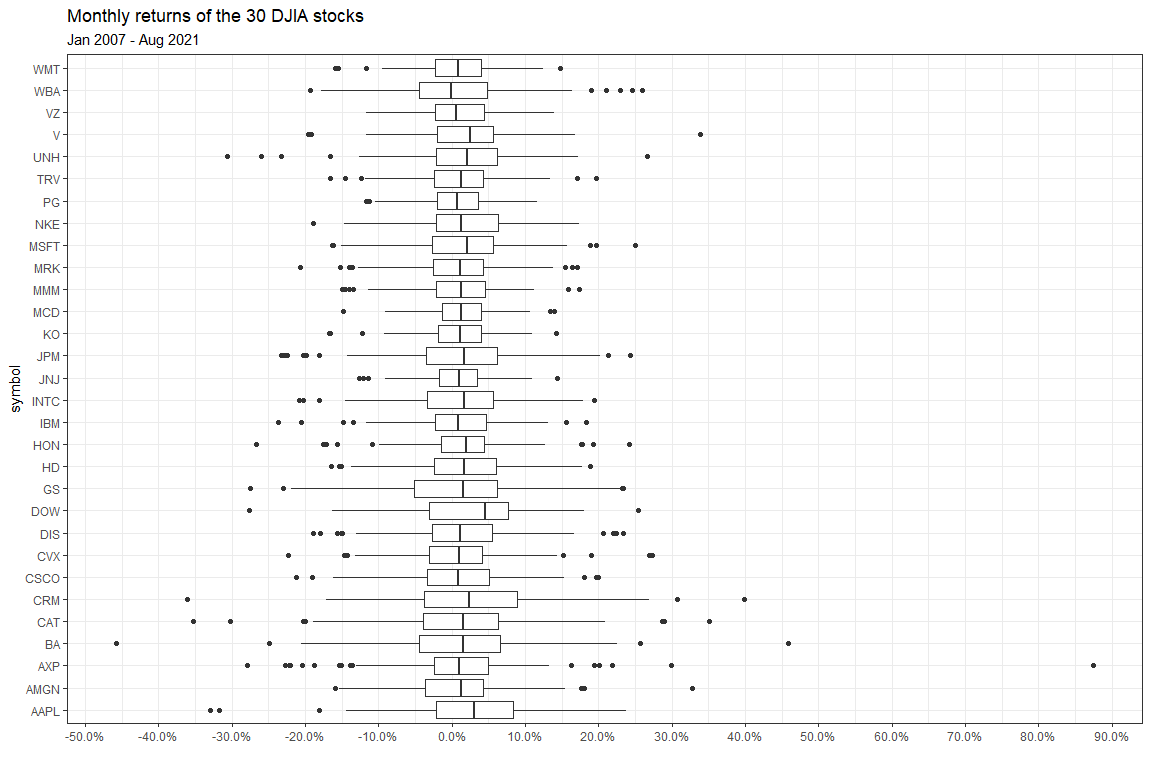
**Facebook friends**

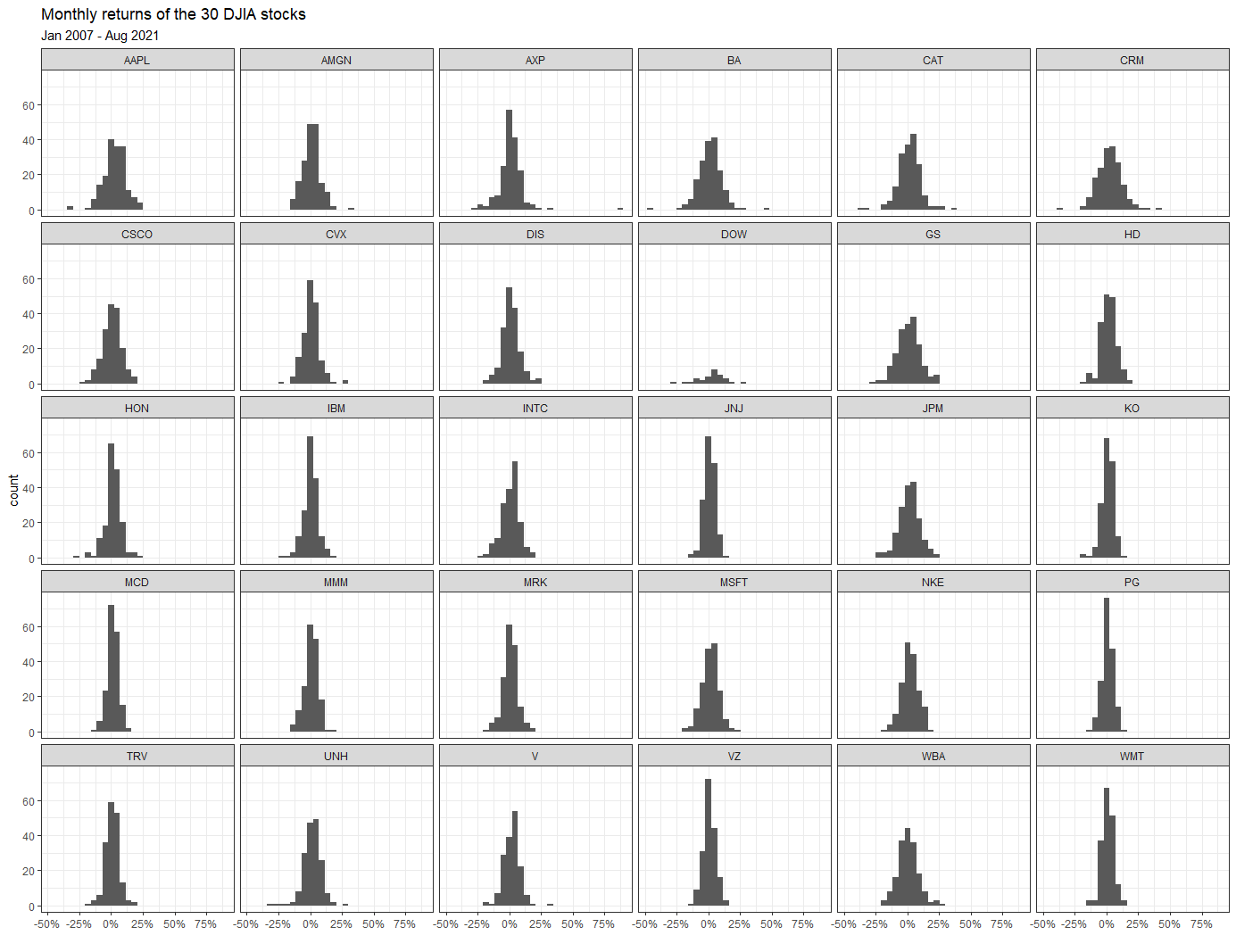




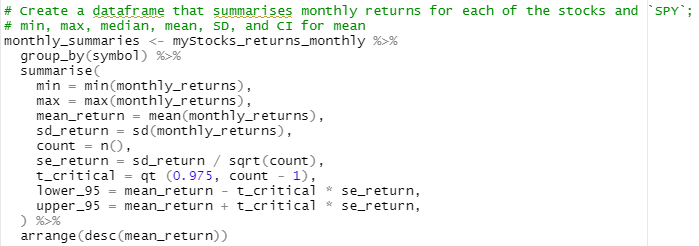
**Part II: Financial Return Analysis**

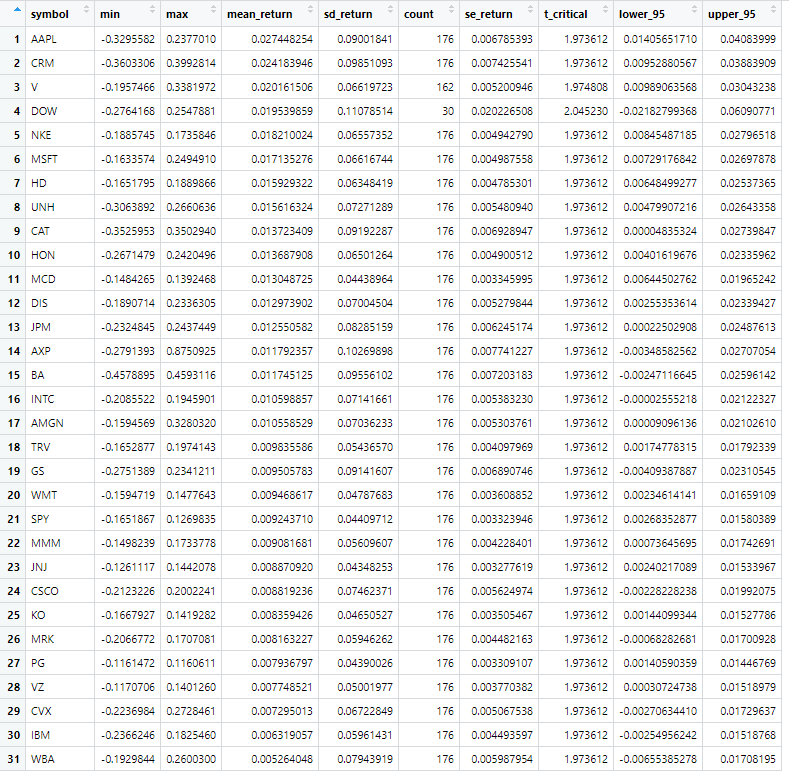
In this exercise, you will analyse monthly financial  returns for the thirty stocks that make up the DJIA and a stock market index, namely SPY, an ETF that tracks the SP500. The R code that downloads the data off the internet and generates all graphs is given on github and you can replicate them whenever you want.





**Calculate summary statistics for the returns**





We have constructed 95% confidence intervals for the mean of each stock; make sure you understand how (just read the code given earlier). How many stocks do **NOT** contain zero in the CI?

For each stock, can you reject the null hypothesis that the mean return is zero (use a 5% significance level)?