## LESSON 7. STATISTICS IN R – PART 1

**Class content:**

1) What is R and what is R Studio?
2) Installing R and R studio
3) Support materials
4) R components and layout
5) Opening the data in R
6) Descriptive statistics in R: summary functions and basic plots

**Exercises:**

1) **Install R and check the version of the software you have installed. You can do that by typing R.Version() in the console.**

- We have successfully downloaded the application.

2) **Create the following vector in R:**
   {8, 9, 9, 14, 8, 8, 10, 7, 6, 9, 7, 8, 10, 14, 11, 8, 14, 11}

   **a) For the data assigned to this vector, calculate the following:**

   **I. Mean**

```
students <- c(8, 9, 9, 14, 8, 8, 10, 7, 6, 9, 7, 8, 10, 14, 11, 8, 14, 11)
elever <- c(8, 9, 9, 14, 8, 8, 10, 7, 6, 9, 7, 8, 10, 14, 11, 8, 14, 11)
pupils <- c(8, 9, 9, 14, 8, 8, 10, 7, 6, 9, 7, 8, 10, 14, 11, 8, 14, 11)
```

- The right picture shows how we have written our students dataset. We can see, that in this case the (<-) sign means equal to the variable instead of (==).
- If we want to calculate the mean, then we can see that we have written mean and then the parentheses. Mean(students). Look, at the picture to the right again!

```
mean(students)
max(students)
min(students)
```

- The a

   **II. Median**

- Now we will go for median, and here we have used the same method. As we can see that the answer has become 9, and just to note that the run the process of loading we need to press Ctrl and Enter.

   **III. Standard deviation**

- The same method is also used for standard deviation, the only difference is that the standard deviation has the initialization sd which is written before the parentheses. We can see that the result would be: 2,45.

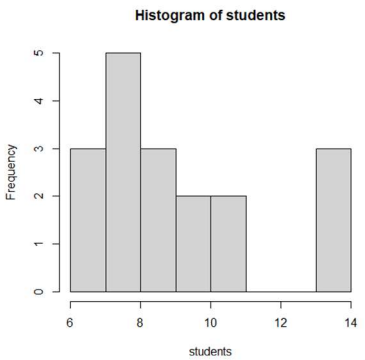
   **b) Construct a histogram for the data**

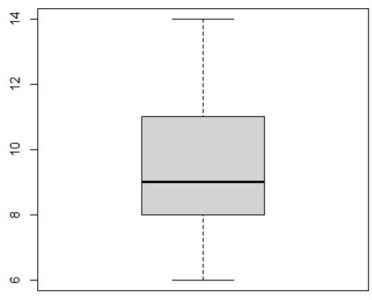
- To construct a histogram, we need to write hist(x=students), and then we need to press Ctrl and Enter and then look at the plot.

   **c) Construct a boxplot for the data**

- We must use the same method, and in this case, we have created a boxplot.

3) This exercise is divided in three steps. These steps are the following:



**Step 1: Create a data frame in R called *data.comput* with data on 5 laptop computers regarding their memory, storage and display size. You know the following:**

- Computer 1 has 8 GB RAM of memory, 500 GB storage drive and 13 inches display.
- Computer 2 has 16 GB RAM of memory, 500 GB storage drive and 15 inches display.
- Computer 3 has 16 GB RAM of memory, 1000 GB storage drive and 13 inches display.
- Computer 4 has 8 GB RAM of memory, 240 GB storage drive and 15 inches display.
- Computer 5 has 16 GB RAM of memory, 500 GB storage drive and 17 inches display.

- So, the way, we have decided to solve this question is by setting three rows. One for the RAM, the other for STORAGE DRIVE, the third for INCHES DISPLAY. We have created an array list or observation dataset for these three variables.
- After, that we have written the following code with my data.

```
ram<-c(8,16,16,8,16)
storage<-c(500,500,1000,240,500)
display<-c(13,15,13,15,17)

mydata <- data.frame(ram,storage,display)
```

- From this code, we have created a frame or a nice table, where we have written all the arrays from ram, storage, and display. The following has resulted in a table like this to the right picture.

| | ram | storage | display |
|---|---|---|---|
| 1 | 8 | 500 | 13 |
| 2 | 16 | 500 | 15 |
| 3 | 16 | 1000 | 13 |
| 4 | 8 | 240 | 15 |
| 5 | 16 | 500 | 17 |

**Step 2: Calculate the mean, median and standard deviation for the variable's "memory", "storage", and "display".**

- Now, because we have created the three different arrays, we can use the same method as we did with the student's array list and we can remove the name student and replace it with the name of the variables.
- Mean for ram: 12,8.
- Mean for storage: 548
- Mean for display: 14,6

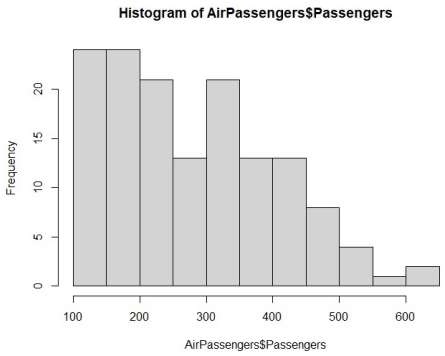
**Step 3: Save the workspace (environment) containing the data frame *data.comput* in a work directory that is convenient to you. To practice how to open it again, close the R session and open the workspace again and see if you can easily recover the objects (i.e. data, values) of the previous session.**

- We have done it, and it works 😊

4) **Import the dataset "Air_passengers.xlsx", which contains data on the number of passengers that have flew in a specific airplane per month. Now do the following:**
    a. **Summarize the data. What is the minimum and maximum number of passengers who flew in this airplane?**

**Solution**

1. Start by importing the data into the R-program.
2. Thereafter use the histogram function to create a graph for the Passengers.




```
hist(x=AirPassengers$Passengers)
```

- Looking at the picture above, we can clearly see that we have created a Histogram.
- Remember, that the graph needs to be created with single number and not with intervals in this case.
    b. **Make a histogram using the default *hist()* function. How would you describe the data distribution?**

**Solution**

1. Use the skews 😊

- We can in this case see, that the graph is positively skewed because we can see that the graph is going from left to right and they are matching the amount of people at each block.
    c. **Define the number of breaks and choose 5 breaks. The HELP tab can help you here.**

**Solution**

1. X
    d. **Change the number of breaks now to 20. Compare this histogram with the one obtained in item b.**

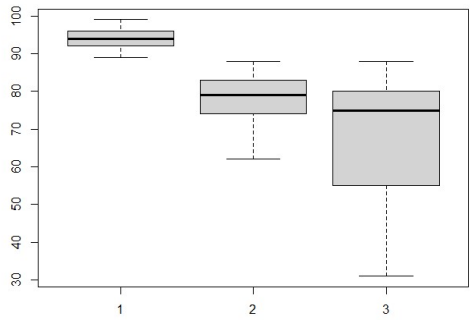
**Solution**

1. X

5) **Import the dataset "basketball.csv", which contains the scores obtained by three professional basketball players in the pre-season games. Make a boxplot for each of the players. When looking at the boxplots, who seems to be the best player? Can we be sure on this result?**

**Solution**

1. Start by importing the data from the basketball Excel data.
2. Thereafter use the boxplot graph.




```
boxplot(Basketball$Michael,Basketball$Damon,Basketball$Allen)
```

You can also use the time in the class to check your answers on the point-giving activity (in ItsLearning) and ask the instructors in case there is any remaining question. :)