LESSON 8. STATICS IN R – PART 2

Class content:

1) Basic operations in R
2) Types of variables in R
3) Inferential statistics in R: Hypothesis testing + ANOVA

Exercises:

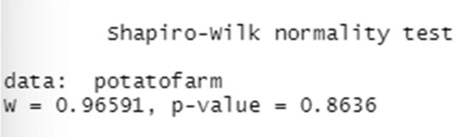
1) We have measured the potato yield from 12 different farms. We know that the standard potato yield for the given variety is μ=20.

x = [21.5, 24.5, 18.5, 17.2, 14.5, 23.2, 22.1, 20.5, 19.4, 18.1, 24.1, 18.5]

Use R to reply the following questions:

a. **Does the population follow a normal distribution?**

- In this case we will start by using the Shapiro.test(potatofarm) because in that way we will be able to conclude the normal distribution.



```
        Shapiro-Wilk normality test

data:  potatofarm
W = 0.96591, p-value = 0.8636
```

-

b. **What is the sample mean?**

- Because we have defined our arraylist, then we can use the mean code which we learned from last time and use it to conclude the mean of the sample.
- In this case, we can see that the sample mean is equal 0,8636 and because it is over the confidence interval of 95%=0,05 then there is a normal distribution.

c. **What is the population mean μ (consider a 95% confidence level)?**

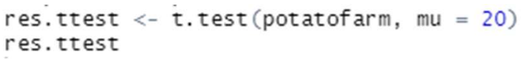
- In this case we can see that the population has the following interval.

$$18{,}255 < 20{,}175 < 22{,}094$$

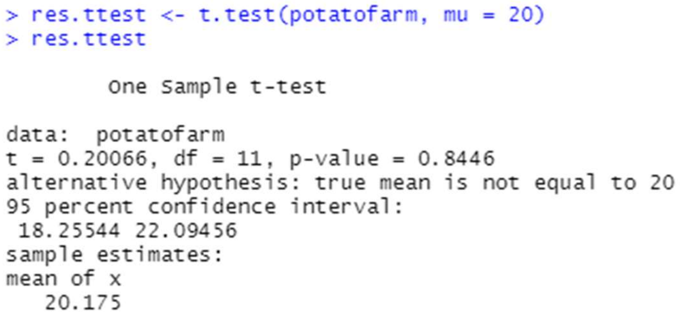
- And because we have got a different p-value, then we can see that it is different from 1,116e^-10 and therefore we can see that it is significantly different.

d. **Is there evidence that the potato yield from these farms is significantly different than the standard yield?**

- To solve this Question, we need to write res.ttest <- t.test(potatofarm, mu = 20). Thereafter we need to write res.ttest, which results a printning of the different values such as the p-value. We can see that the p-value is equal to 0,8446 and the confidence interval is equal to 95% or 0,05.
- In this case we can see, that the potatfarms yield are NOT significantly different than the standard yield.



```
res.ttest <- t.test(potatofarm, mu = 20)
res.ttest
```



```
> res.ttest <- t.test(potatofarm, mu = 20)
> res.ttest

        One Sample t-test

data:  potatofarm
t = 0.20066, df = 11, p-value = 0.8446
alternative hypothesis: true mean is not equal to 20
95 percent confidence interval:
 18.25544 22.09456
sample estimates:
mean of x
   20.175
```

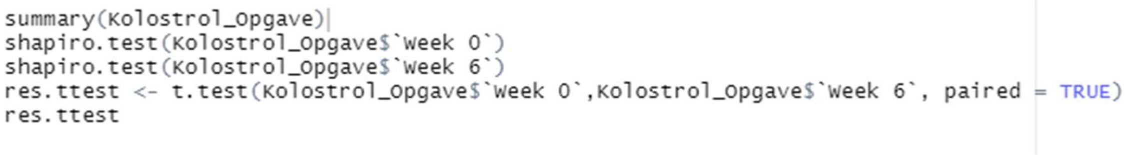
2) A researcher wants to see if a vitamin included in the diet changes the cholesterol. Six subjects were pretested at Week 0, and then they took the vitamin during 6-weeks. After the 6-weeks period, their cholesterol level was measured again. Using **R**, can we conclude (with 95% confidence level) that the cholesterol level has been changed? Assume the variable is approximately normally distributed.

Subject	1	2	3	4	5	6
Week 0	215	239	208	190	172	244
Week 6	184	160	201	188	169	219

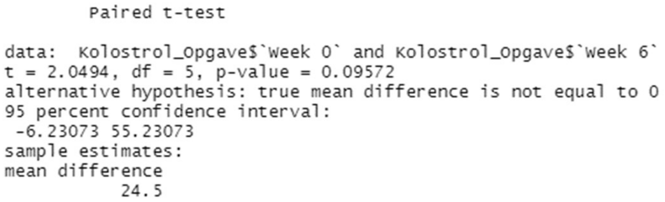
- In this case we will start by importing the table into R so we don't get problems in summarizing the data of the table.

3) Repeat the previous exercise using now a 90% confidence level.

- So, first we will start by looking at the Question and here we can see that it is a two-sample dependent t-test because we are measuring same things, with a gap of 6 weeks.
- In this case we will start by writing the Shapiro-test because we want to know the distribution value. In our program we have written Shapiro.test(Kolostrol_Opgave$Week0) and Shapiro.test(Kolostrol_Opgave$Week6) and the reason is because Shapiro.test cannot receive to values at the same time but needs to receive it individually.
- Then we need to afterwards use the res.ttest where we write the t.test and the name of the file in paranthesis with a dollar sign in between the file name and the week 0.



```
summary(Kolostrol_Opgave)
shapiro.test(Kolostrol_Opgave$`Week 0`)
shapiro.test(Kolostrol_Opgave$`Week 6`)
res.ttest <- t.test(Kolostrol_Opgave$`Week 0`,Kolostrol_Opgave$`Week 6`, paired = TRUE)
res.ttest
```



```
        Paired t-test

data:  Kolostrol_Opgave$`Week 0` and Kolostrol_Opgave$`Week 6`
t = 2.0494, df = 5, p-value = 0.09572
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -6.23073 55.23073
sample estimates:
mean difference
           24.5
```

- We can see that the following data shows the following interval:

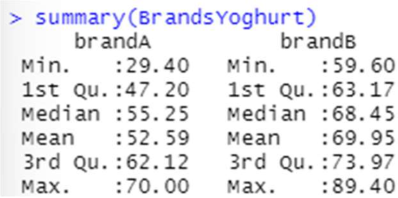
$$-6,230 < 24,5 < 55,230$$

- We can see that in this case we have got the following p-value and that is 0,09572 and our confidence level which is 90% is equal to 0,1. Here we can see that our p-value is lower than 0,1, which concludes that our data is significantly different.

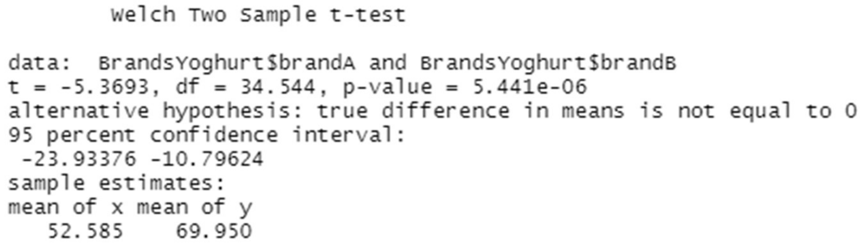
4) **We would like to know if the concentration of a compound in two brands of yogurt is different. We select 20 bottles of Brand A and 20 bottles of Brand B. The results are shown in the excel file "yogurt.xslx".**

   a) **What is the appropriate test to use to respond our research question?**

- So, like the question number 2. We can use the two-sample t-test, but instead use the independent test form.
- First of all, we will start by using the independent t.test formula we will start by writing the summary (file name which is BrandsYoghurt). Afterwards we will write the res.ttest <- t.test(BrandsYoghurt$brandA, BrandsYoghurt$brandB) and then we get the following result:



```
> summary(BrandsYoghurt)
     brandA          brandB     
 Min.   :29.40   Min.   :59.60  
 1st Qu.:47.20   1st Qu.:63.17  
 Median :55.25   Median :68.45  
 Mean   :52.59   Mean   :69.95  
 3rd Qu.:62.12   3rd Qu.:73.97  
 Max.   :70.00   Max.   :89.40  
```



```
        Welch Two Sample t-test

data:  BrandsYoghurt$brandA and BrandsYoghurt$brandB
t = -5.3693, df = 34.544, p-value = 5.441e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -23.93376 -10.79624
sample estimates:
mean of x mean of y 
   52.585    69.950 
```

- In this case we can see that the following answer which tells the following answer in task 4b.

   b) **What is the main assumption to be tested before performing the test? After importing the data to R, test the assumption using the appropriate method.**

- So, the main assumption is that the data is normally distributed, as the confidence level is 95=0,05. In this case the dataset for both brand A and B are normally distributed because the t-value shown from the question A is equal -5,3693. In this case we have the following interval.

$$-23,93 < -5,36 < -10,79$$

– In this case we can see that the interval shows us that the two brands for Yoghurt are normally distributed.

   c) **Can you conclude whether the compound's concentration in the two brands of yogurts is significantly different?**

- But if you look closer towards the p-value, then you can see that the p-value is lower the 0,000005 which tells us that our two brands of yoghurts are significantly different.

5) **We want to compare the scores obtained by three professional basketball players. The data with all the scores obtained by the players in the pre-season games is available in the csv file "basketball.csv".**

- We have downloaded the file.

   a. **Construct a boxplot for each of the players, to better visualize the data.**

- So, because we want to construct a boxplot, we need to use the keyword "boxplot" to inherit the data from the basketball file into the boxplot function.

   b. **When using the appropriate statistical test, is there a significant difference among the scores obtained by each of the players?**

Solution

1. Import the dataset.
2. Start by using the hypothesis testing 3.

   c. **In case there is a difference, which player/players obtained a higher or lower score than the other/others?**

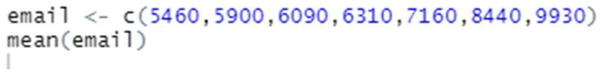
- Skip

6) **According to the Harvard Business Review (in the article: "How to Spend Way Less Time on Email Every Day"), the average professional checks his/her emails 15 times per day.**
The data represent a sample of the number of times/years, that 7 employees in a company check their emails:
5460 5900 6090 6310 7160 8440 9930
Use R to find out: which one of the following statements is correct?

   A. We can be 99% confident that the mean number of times that the employees of this company check their email each year is between 4785 and 9298.

Solution

1. Start by writing the dataset.
- So, to solve this Question we need to write of the dataset in the R-program.
2. Write the Mean-Function.
- Then, we will afterwards write the mean-keyword where we will the value out.

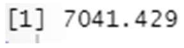


```
email <- c(5460,5900,6090,6310,7160,8440,9930)
mean(email)
```

- REMEMBER TO INSERT THE LETTER C BEFORE THE PARANTHESIS

3. Conclude the Answer in the End.

-

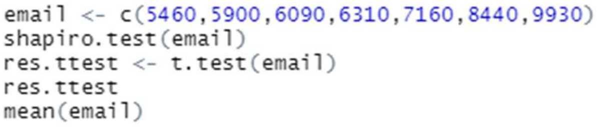


```
[1] 7041.429
```

B. We can be 99% confident that the mean number of times that the employees of this company check their email is not significantly different from that of the "average professional".

Solution

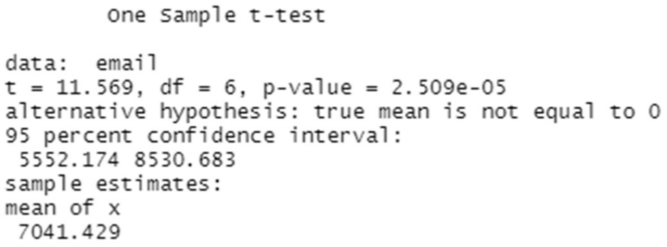
1. Write the summary, Shapiro test and res.ttest.



```
email <- c(5460,5900,6090,6310,7160,8440,9930)
shapiro.test(email)
res.ttest <- t.test(email)
res.ttest
mean(email)
```

2. Look at the p-value and compare with the confidence level.

-



```
        One Sample t-test

data:  email
t = 11.569, df = 6, p-value = 2.509e-05
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 5552.174 8530.683
sample estimates:
mean of x
 7041.429
```

- The p-value is lower than the confidence level 0,05, which the p-value is equal to 0,00002509.

3. Conclude

- In this case we can see that because we have a confidence level of 99%, which is equal to 0,05. In this case we can see that our p-value is x the confidence level, which in conclusion tells us that that the employees checking their email is xxx significantly different from that of the average professional.

C. None of the previous responses is correct.
D. A and B are correct.

7) The number of children born in 7 towns in a region is:

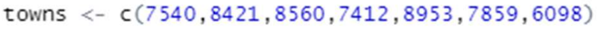
7540 8421 8560 7412 8953 7859 6098

Find the 99% confidence interval for the mean number of children born annually per town.

Solution

1. Start by writing the dataset.

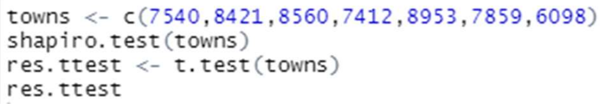
-



```
towns <- c(7540,8421,8560,7412,8953,7859,6098)
```

2. Thereafter write the Shapiro.test.

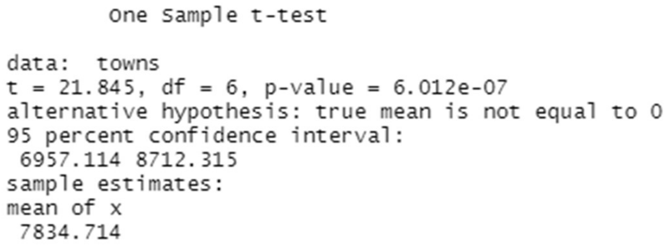
-



```
towns <- c(7540,8421,8560,7412,8953,7859,6098)
shapiro.test(towns)
res.ttest <- t.test(towns)
res.ttest
```

3. Out print the values from the function.

-



```
        One Sample t-test

data:  towns
t = 21.845, df = 6, p-value = 6.012e-07
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 6957.114 8712.315
sample estimates:
mean of x
 7834.714
```

4. Conclude in the end.

- In this case the following Interval is equal to this:

$$6957,114 < 7834,71 < 8712,315$$

8) We want to evaluate three different methods to lower the blood pressure of individuals that have been diagnosed with high blood pressure. Eighteen subjects are randomly assigned to three groups (6 per group): the first group takes medication, the second group exercises, and the third one follows a specific diet. After four weeks, the reduction in each person's blood pressure is recorded. Is there a significant difference among the reduction obtained from each of the three methods? If yes, which method was more effective?

Medication	Exercise	Diet
12	14	6
8	9	10
11	2	5
17	5	9
16	7	8
15	4	6