Statistisk Dataanalyse 2023

Instruktør: Vivek Misra

Exercise Class NR8

Solutions to the Tasks

Task 1 - Description

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

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?
b. What is the sample mean?
c. What is the population mean μ (consider a 95% confidence level)?
d. Is there evidence that the potato yield from these farms is significantly different than the standard yield?

Task 1 – Solution

- We will write the x-value as potato_farm in this case.
  - Then we will assign the values to the vector-name.
  - We will then use the shapiro.test to find the normal distribution.
  - Then we will use res.ttest to print the result from the confidence level.

```
potato_farm <- c(21.5,24.5,18.5,17.2,14.5,23.2,22.1,20.5,19.4,18.1,24.1,18.5)
shapiro.test(potato_farm)
res.ttest <- t.test(potato_farm, mu=20)
res.ttest
```

```
> shapiro.test(potato_farm)

        Shapiro-Wilk normality test

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

```
> res.ttest

        One Sample t-test

data:  potato_farm
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
```

Task 1 – Solution

A. Yes, there is a normal distribution, because the p-value lies over the confidence level which is 95%=0,05. 0,86>0,05 = not significant.
B. Here we have the sample to 20,175.
C. Confidence Interval is: -18,25<20,175<+18,25
D. No, there is a evidence that there is not a significant difference between the potato yield farm and the others. This is due, that the p-value is stronger than the conf.level.

```
> shapiro.test(potato_farm)

        Shapiro-Wilk normality test

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

```
> res.ttest

        One Sample t-test

data:  potato_farm
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
```

Task 2 - Description

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

Task 2 - Solution

- We have started by creating a data.frame, where the values are inside.
- Shapiro.test has been included, but paired has been added due to a two sample dependent t-test.

```
cholostrol <- data.frame(week0=c(215,239,208,190,172,244),week6=c(184,160,201,188,169,219))
summary(cholostrol)
shapiro.test(cholostrol$week0)
shapiro.test(cholostrol$week6)
res.ttest <- t.test(cholostrol$week0,cholostrol$week6,paired=TRUE)
res.ttest
```

Task 2 - Solution

- CONCLUSION: We can see, that there is not a significant difference between the cholostrols measured over the six-weeks period.
  - This means, that the cholostrol level has not changed.
  - And that there is a normal distribution in the datase, while the p-value is not significant.

```
> summary(cholostrol)
     week0           week6
 Min.   :172.0   Min.   :160.0
 1st Qu.:194.5   1st Qu.:172.8
 Median :211.5   Median :186.0
 Mean   :211.3   Mean   :186.8
 3rd Qu.:233.0   3rd Qu.:197.8
 Max.   :244.0   Max.   :219.0
```

```
> shapiro.test(cholostrol$week0)

        Shapiro-Wilk normality test

data:  cholostrol$week0
W = 0.95365, p-value = 0.7697
```

```
> shapiro.test(cholostrol$week6)

        Shapiro-Wilk normality test

data:  cholostrol$week6
W = 0.97885, p-value = 0.9457
```

```
> res.ttest

        Paired t-test

data:  cholostrol$week0 and cholostrol$week6
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
```

Task 3- Description

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 **90%** 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

Task 3- Solution

- As we saw in the shapiro.test, we just need to write conf.level in the outprint of the shapiro.test which is the res.ttest.

```
res.ttest <- t.test(cholostrol$week0,cholostrol$week6,paired=TRUE,conf.level = 0.90)
res.ttest
```

- As you can see, that the confidence level is 0,41<24,5<48,59.
- There is not a significant difference in the measurement.
- And there is also a normal distribution.

```
> res.ttest

        Paired t-test

data:  cholostrol$week0 and cholostrol$week6
t = 2.0494, df = 5, p-value = 0.09572
alternative hypothesis: true mean difference is not equal to 0
90 percent confidence interval:
  0.4105483 48.5894517
sample estimates:
mean difference
           24.5
```

Task 4 - Description

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

Task 4 - Solution

A. We will use a two-sample independent t-test.

B. We have use the following commands.

```
summary(yoghurt)
shapiro.test(yoghurt$brandA)
shapiro.test(yoghurt$brandB)
res.ttest <- t.test(yoghurt$brandA,yoghurt$brandB)
res.ttest
```

C. There is a significant different between the two yoghurt-brands.

   A. There is not a normal distribution of the dataset.

```
> summary(yoghurt)
     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
```

```
> shapiro.test(yoghurt$brandA)

        Shapiro-Wilk normality test

data:  yoghurt$brandA
W = 0.93556, p-value = 0.1975

> shapiro.test(yoghurt$brandB)

        Shapiro-Wilk normality test

data:  yoghurt$brandB
W = 0.91529, p-value = 0.08044
```

```
> res.ttest

        Welch Two Sample t-test

data:  yoghurt$brandA and yoghurt$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
```

Task 5 - Description

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".
    a. Construct a boxplot for each of the players, to better visualize the data.
    b. When using the appropriate statistical test, is there a significant difference among the scores obtained by each of the players?
    c. In case there is a difference, which player/players obtained a higher or lower score than the other/others?

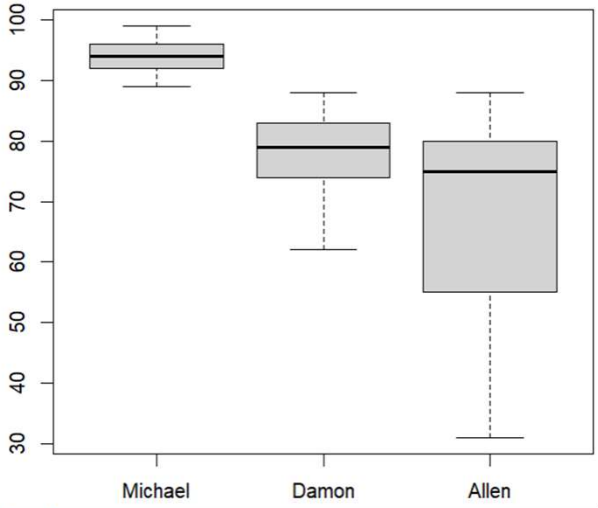
Task 5 - Solution

- We have created the following commands:

```
summary(basketball)
boxplot(basketball)
shapiro.test(basketball$Michael)
shapiro.test(basketball$Damon)
shapiro.test(basketball$Allen)
res.ttest <- t.test(basketball)
res.ttest
```

Task 5 - Solution

A. We have the boxplot here ->



B. The p-value is below 0,05 and therefore we can say that there is a significant difference.

```
> res.ttest

        One Sample t-test

data:  basketball
t = 38.52, df = 62, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 75.30673 83.55042
sample estimates:
mean of x
 79.42857
```

C. Just by looking at the boxplot, we can see that Michael has the highest score. Then we can see, that Damon is the next highest score, along with Allen.

Task 6 - Description

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/year, 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.
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".
C. None of the previous responses is correct.
D. A and B are correct.

Task 6 - Solution

- We will use the one sample t-test to solve this question.
- NOTE: We have used conf.level=0.99 to define the confidence level.

```
employee <-c(5460,5900,6090,6310,7160,8440,9930)
shapiro.test(employee)
res.ttest <- t.test(employee,conf.level=0.99)
res.ttest
```

- We have got the following result.

```
        One Sample t-test

data:  employee
t = 11.569, df = 6, p-value = 2.509e-05
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 4784.991 9297.866
sample estimates:
mean of x
 7041.429
```

- **We can see, that the null-hypothesis is accepted.**
  - The sample mean is located in between the confidence interval.
  - ANSWER: D is 100% correct, which means that A and B are both correct.

Task 7 - Description

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.

Task 7 - Solution

- We have used commands for one sample t-test.

```
towns <-c(7540,8421,8560,7412,8953,7859,6098)
shapiro.test(towns)
res.ttest <- t.test(towns,conf.level=0.99)
res.ttest
```

- We have got the following results:

```
        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
99 percent confidence interval:
 6505.022 9164.407
sample estimates:
mean of x
 7834.714
```

- Null-Hypothesis is accepted due to the sample mean lying between the confidence level interval of 99%.

Task 8 - Description

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

Task 8 - Solution

```r
measure <- c(12,8,11,17,16,15,14,9,2,5,7,4,6,10,5,9,8,6)
treatment <- c(rep("m",6), rep("e",6), rep("d",6))
bloodpressure <- data.frame(measure,treatment)
bloodpressure
str(bloodpressure)
bloodpressure$treatment <- as.factor(bloodpressure$treatment)
res.aov <- aov(measure~treatment,data=bloodpressure)
summary(res.aov)
print(LSD.test(res.aov,"treatment"))
```

- We will use One Way ANOVA to solve this Question.
  1. We have written all the data.
  2. We have divided the data into 3 parts.
  3. We have created a dataframe, so that the 3 parts are sat on a column.
  4. We have looked for each datatype of the value.
  5. The treatment value is with the headers are treated as string-values.
  6. One-Way Anova is Conducted from the Bloodpressure database.
  7. A Summary has been created to find the significant difference with the p-value.
  8. LSD-Test shows how different the treatment-methods are from eachother.

Task 8 - Solution

```
> measure <- c(12,8,11,17,16,15,14,9,2,5,7,4,6,10,5,9,8,6)
> treatment <- c(rep("m",6), rep("e",6), rep("d",6))
> bloodpressure <- data.frame(measure,treatment)
> bloodpressure
   measure treatment
1       12         m
2        8         m
3       11         m
4       17         m
5       16         m
6       15         m
7       14         e
8        9         e
9        2         e
10       5         e
11       7         e
12       4         e
13       6         d
14      10         d
15       5         d
16       9         d
17       8         d
18       6         d
```

```
> str(bloodpressure)
'data.frame':   18 obs. of  2 variables:
 $ measure  : num  12 8 11 17 16 15 14 9 2 5 ...
 $ treatment: chr  "m" "m" "m" "m" ...
```

```
> bloodpressure$treatment <- as.factor(bloodpressure$treatment)
> res.aov <- aov(measure~treatment,data=bloodpressure)
> summary(res.aov)
            Df Sum Sq Mean Sq F value  Pr(>F)
treatment    2  148.8   74.39   6.603 0.00877 **
Residuals   15  169.0   11.27
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
$groups
      measure groups
m 13.166667      a
d  7.333333      b
e  6.833333      b
```

Task 8 - Solution

- **We can see, that there is a significant difference.**
  - *P-VALUE=0,008 and it is below 0,05=95%.*
  - ***Null-Hypothesis is Rejected and Alternative Hypothesis is Accepted.***
- **LSD-Test**
  - *All treatments are different.*
  - ***Medications has the highest effectivity among the other treatments.***
  - *Exercise has the lowest effectivity amon the other treatments.*
  - *Diet ranks on the second spot, just behind Medication.*

Tak for i dag!

Instruktør: Vivek Misra