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:

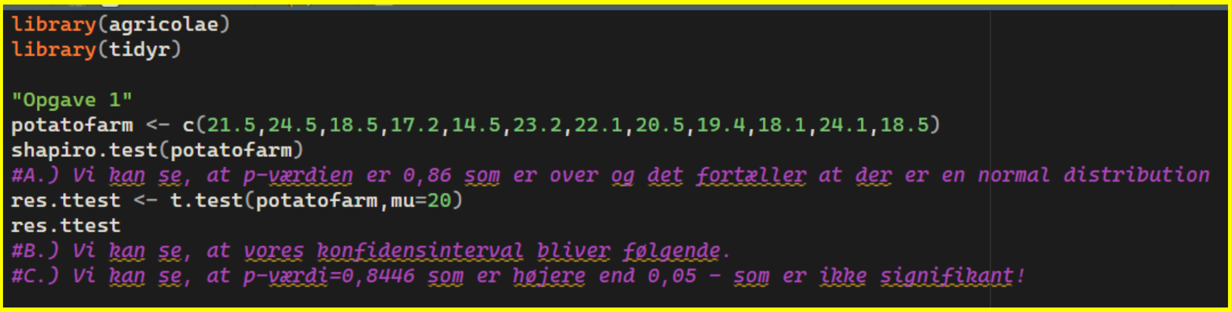
1) We have measured the potato yield from 12 different farms. We know that the standard potato yield for the given variety is μ=20.

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 to the following questions:

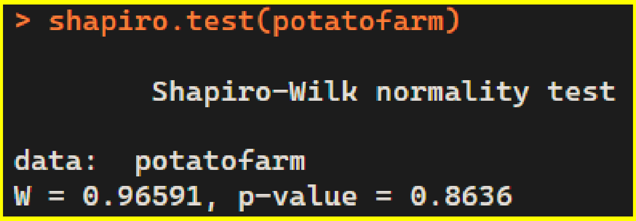
a. Does the population follow a normal distribution?
b. What is the sample mean?
c. What is the population mean μ (consider a 95% confidence level)?
d. Is there evidence that the potato yield from these farms is significantly different than the standard yield?

j. **Såden har vi løst opgaven!**
   - **Husk, at shapiro.test er normal distributionstesten.**
   - **Hvorimod t.test er kendt som one sample t-test**



```
library(agricolae)
library(tidyr)

"Opgave 1"
potatofarm <- 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(potatofarm)
#A.) Vi kan se, at p-værdien er 0,86 som er over og det fortæller at der er en normal distribution
res.ttest <- t.test(potatofarm,mu=20)
res.ttest
#B.) Vi kan se, at vores konfidensinterval bliver følgende.
#C.) Vi kan se, at p-værdi=0,8446 som er højere end 0,05 - som er ikke signifikant!
```

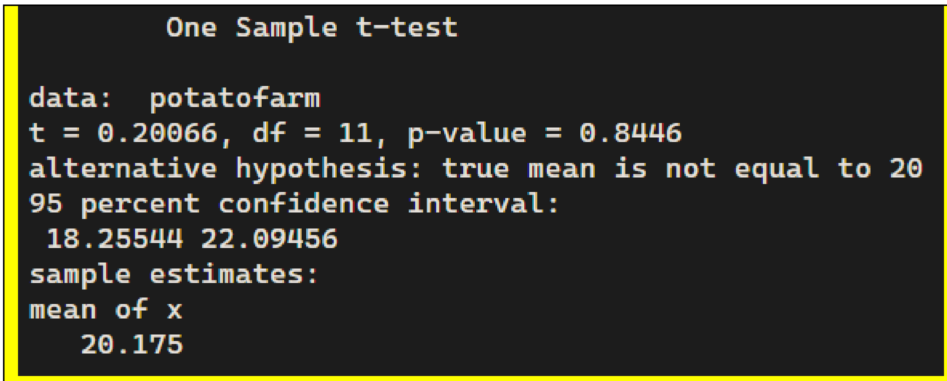
a. **Følgende resultat fås:**



```
> shapiro.test(potatofarm)

        Shapiro-Wilk normality test

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

- **Vi kan se, at resultatet også viser at der er snak om en normal distribution fordi p-værdien er over grænsen som er 0,05=95%**

b. **Følgende resultat fås:**



```
        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
```

- **Vi kan se, at vores sample mean ender med at være 20,175**

c. Følgende resultat fås:

- **Populationens mean er blevet sat til 0.**

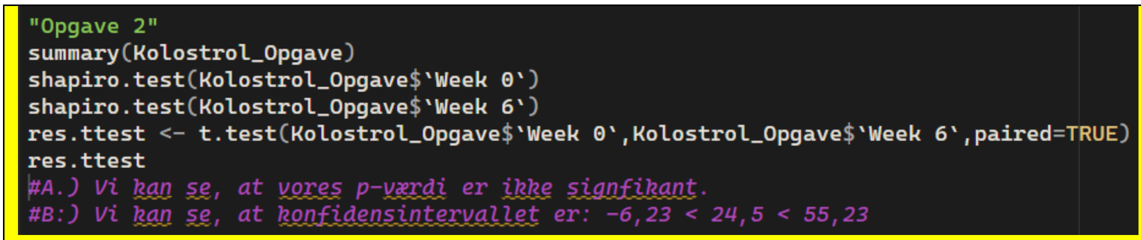
d. Følgende resultat fås:

- **Set på billedet på oven ved opgave b, kan det ses at selve p-værdien ligger over 0,05=95% og det er markeret med blåt i one sample t-test.**

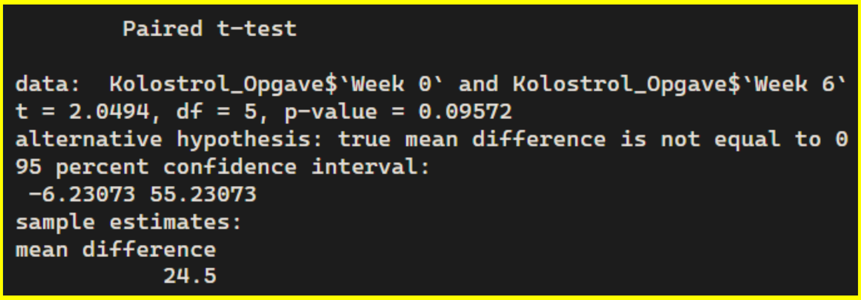
2) A researcher wants to see if a vitamin included in the diet changes the cholesterol. Six subjects were pretested at Week 0, and then they took the vitamin during 6-weeks. After the 6-weeks period, their cholesterol level was measured again. Using R, can we conclude (with 95% confidence level) that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

- **Vi løser Two-Sample T-Test på følgende måde.**
- **Bemærk, at paired TRUE er den som karakteriserer denne dependent metode.**



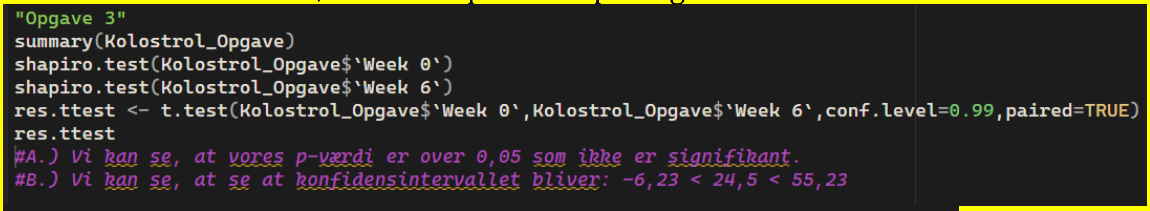
```
"Opgave 2"
summary(Kolostrol_Opgave)
shapiro.test(Kolostrol_Opgave$`Week 0`)
shapiro.test(Kolostrol_Opgave$`Week 6`)
res.ttest <- t.test(Kolostrol_Opgave$`Week 0`,Kolostrol_Opgave$`Week 6`,paired=TRUE)
res.ttest
#A.) Vi kan se, at vores p-værdi er ikke signfikant.
#B:) Vi kan se, at konfidensintervallet er: -6,23 < 24,5 < 55,23
```



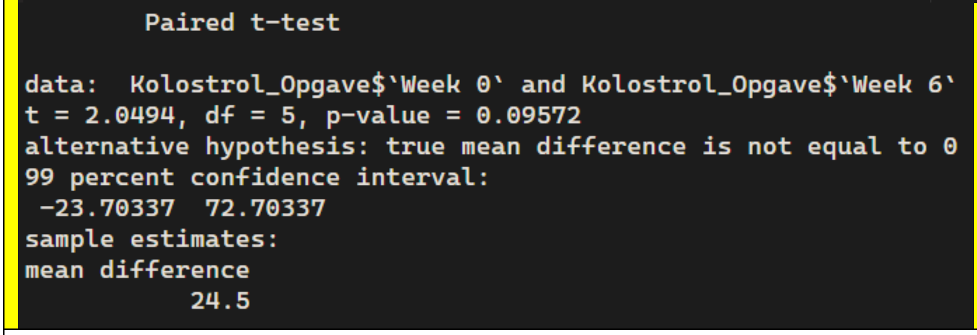
```
        Paired t-test

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
```

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

- Vi kan se, at vi har opskrevet på følgende måde.



```
"Opgave 3"
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`,conf.level=0.99,paired=TRUE)
res.ttest
#A.) Vi kan se, at vores p-værdi er over 0,05 som ikke er signifikant.
#B.) Vi kan se, at se at konfidensintervallet bliver: -6,23 < 24,5 < 55,23
```

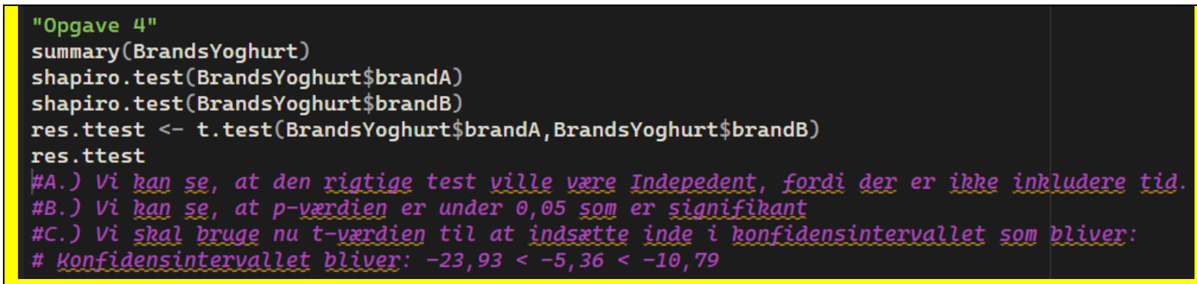


```
        Paired t-test

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
99 percent confidence interval:
 -23.70337  72.70337
sample estimates:
mean difference
           24.5
```

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?

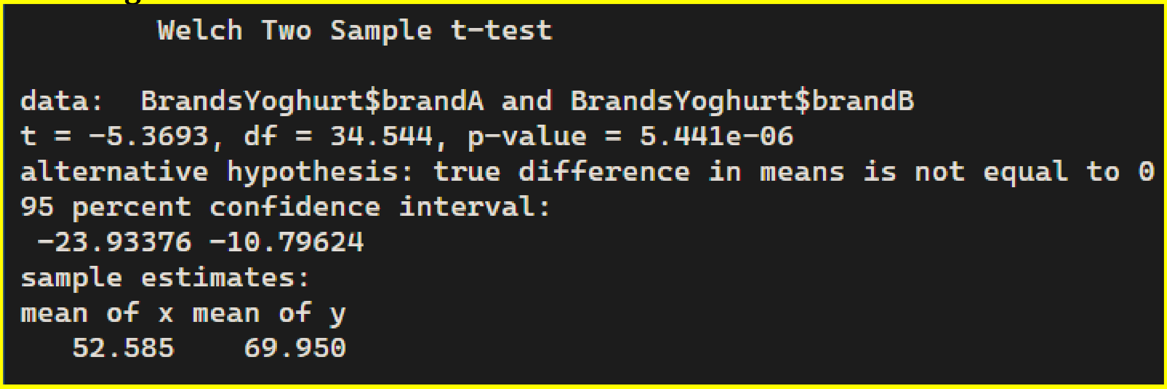
A. Det er Two-Sample T-test Independent, fordi vi sammenligner to ting med hinanden 😉

B. Følgende kode er opskrevet i R.



```
"Opgave 4"
summary(BrandsYoghurt)
shapiro.test(BrandsYoghurt$brandA)
shapiro.test(BrandsYoghurt$brandB)
res.ttest <- t.test(BrandsYoghurt$brandA,BrandsYoghurt$brandB)
res.ttest
#A.) Vi kan se, at den rigtige test ville være Indepedent, fordi der er ikke inkludere tid.
#B.) Vi kan se, at p-værdien er under 0,05 som er signifikant
#C.) Vi skal bruge nu t-værdien til at indsætte inde i konfidensintervallet som bliver:
# Konfidensintervallet bliver: -23,93 < -5,36 < -10,79
```

C. **Følgende resultat er fået i R og derfra kan vi konkludere om tingene er signifikante.**

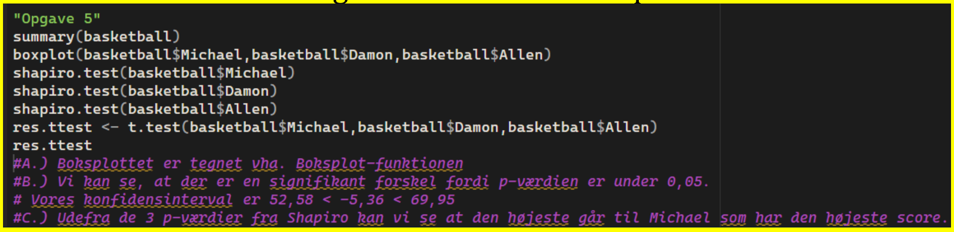


```
        Welch Two Sample t-test

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 
```

- **Vi kan se, at fordi vores p-værdi ligger under grænsen af 95% og derfor kan vi sige at de 2 brands koncentrationer er signifikante.**

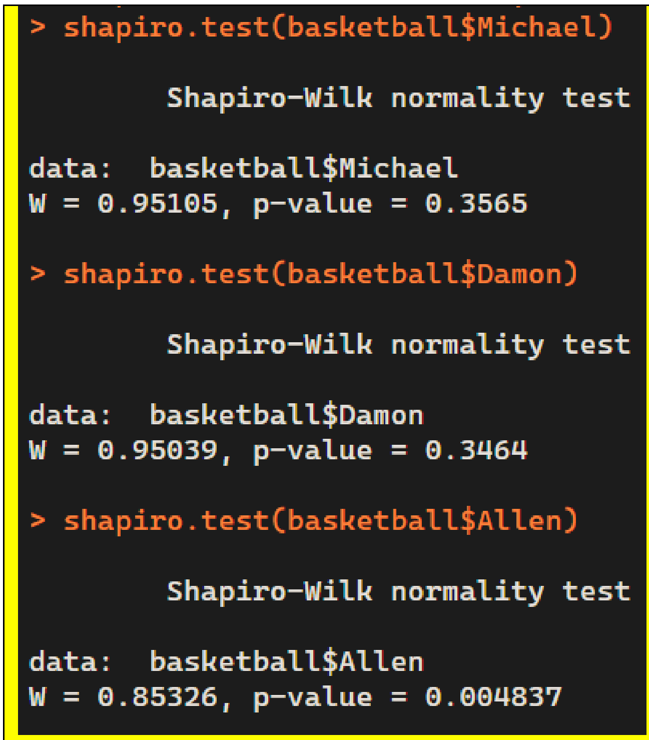
5) We want to compare the scores obtained by three professional basketball players. The data with all the scores obtained by the players in the pre-season games is available in the csv file "basketball.csv".
   a. Construct a boxplot for each of the players, to better visualize the data.
   b. When using 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?

a. Vi har skrevet følgende R-kommando op.



```
"Opgave 5"
summary(basketball)
boxplot(basketball$Michael,basketball$Damon,basketball$Allen)
shapiro.test(basketball$Michael)
shapiro.test(basketball$Damon)
shapiro.test(basketball$Allen)
res.ttest <- t.test(basketball$Michael,basketball$Damon,basketball$Allen)
res.ttest
#A.) Boksplottet er tegnet vha. Boksplot-funktionen
#B.) Vi kan se, at der er en signifikant forskel fordi p-værdien er under 0,05.
# Vores konfidensinterval er 52,58 < -5,36 < 69,95
#C.) Udefra de 3 p-værdier fra Shapiro kan vi se at den højeste går til Michael som har den højeste score.
```

b. Her har vi kigget på de individuelle p-værdier fra enhver spiller og sammenlignet med konfidensniveauet.



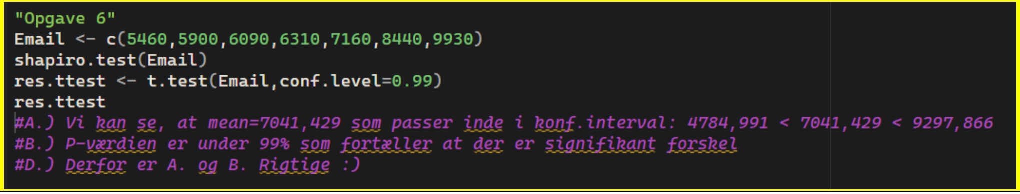
```
> shapiro.test(basketball$Michael)

        Shapiro-Wilk normality test

data:  basketball$Michael
W = 0.95105, p-value = 0.3565

> shapiro.test(basketball$Damon)

        Shapiro-Wilk normality test

data:  basketball$Damon
W = 0.95039, p-value = 0.3464

> shapiro.test(basketball$Allen)

        Shapiro-Wilk normality test

data:  basketball$Allen
W = 0.85326, p-value = 0.004837
```

- Vi kan se, at Michael og Damon er ikke signifikante, men vi kan se at Allen er.

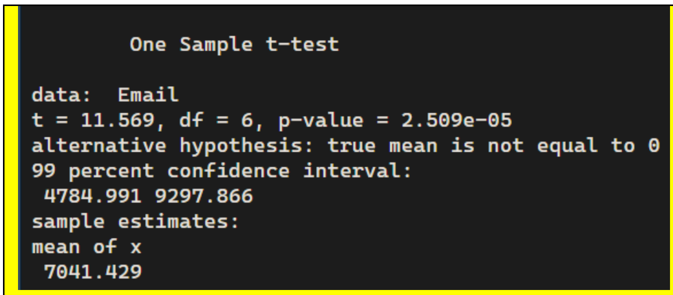
c. I tilfældet, kan du tage gennemsnittet af hver af spillerne og sammenligne disse. Dog er dette ikke helt den

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.
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.

a. Her tager vi middelværdien og putter ind mellem de to angivet intervaller gennem one-sample t-test ved t.test.

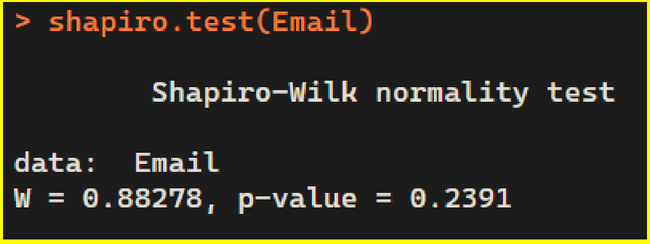


```
"Opgave 6"
Email <- c(5460,5900,6090,6310,7160,8440,9930)
shapiro.test(Email)
res.ttest <- t.test(Email,conf.level=0.99)
res.ttest
#A.) Vi kan se, at mean=7041,429 som passer inde i konf.interval: 4784,991 < 7041,429 < 9297,866
#B.) P-værdien er under 99% som fortæller at der er signifikant forskel
#D.) Derfor er A. og B. Rigtige :)
```



```
        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
99 percent confidence interval:
 4784.991 9297.866
sample estimates:
mean of x
 7041.429
```

b. Vi kan i tilfældet se på shapiro.testen og derved se omkring p-værdien.



```
> shapiro.test(Email)

        Shapiro-Wilk normality test

data:  Email
W = 0.88278, p-value = 0.2391
```

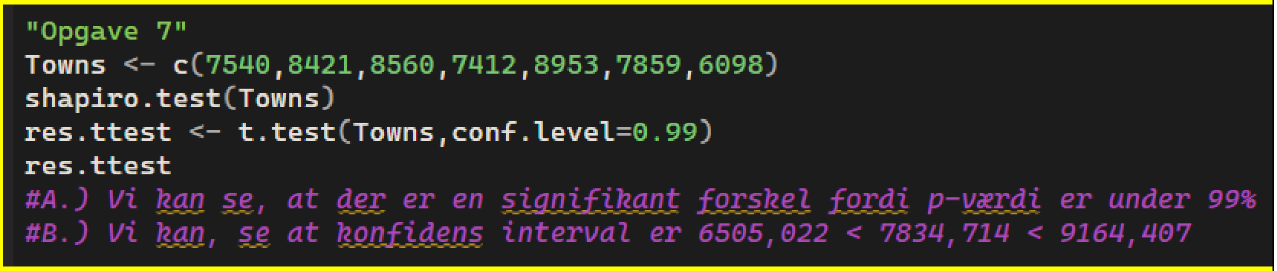
- Vi kan i tilfældet, se at vores p-værdi er større end 0,01=99%. Derfor er det ikke signifikant.

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

7540 8421 8560 7412 8953 7859 6098

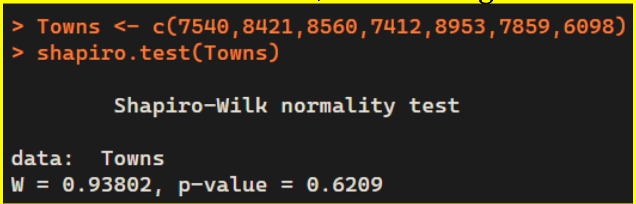
Find the 99% confidence interval for the mean number of children born annually per town.

I. Følgende kode er blevet skrevet op:

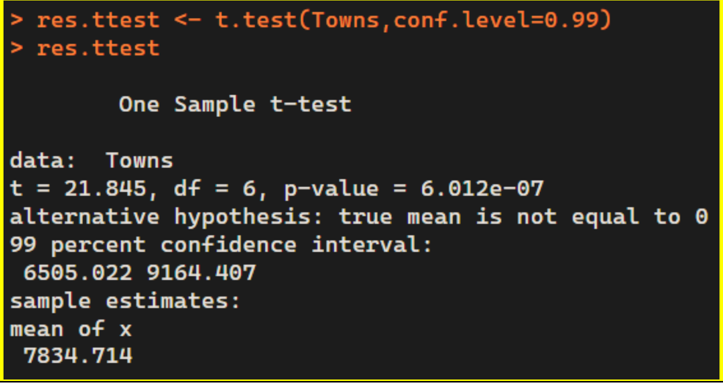


```
"Opgave 7"
Towns <- c(7540,8421,8560,7412,8953,7859,6098)
shapiro.test(Towns)
res.ttest <- t.test(Towns,conf.level=0.99)
res.ttest
#A.) Vi kan se, at der er en signifikant forskel fordi p-værdi er under 99%
#B.) Vi kan, se at konfidens interval er 6505,022 < 7834,714 < 9164,407
```

- Vi kan se, at vi får følgende resultater fra testen.



```
> Towns <- c(7540,8421,8560,7412,8953,7859,6098)
> shapiro.test(Towns)

        Shapiro-Wilk normality test

data:  Towns
W = 0.93802, p-value = 0.6209
```



```
> res.ttest <- t.test(Towns,conf.level=0.99)
> res.ttest

        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
```

- I tilfældet, kan det ses at vores middelværdi som er markeret med blåt, ligger mellem de to intervaller: 6505,022 < 7834,714 < 9164,407

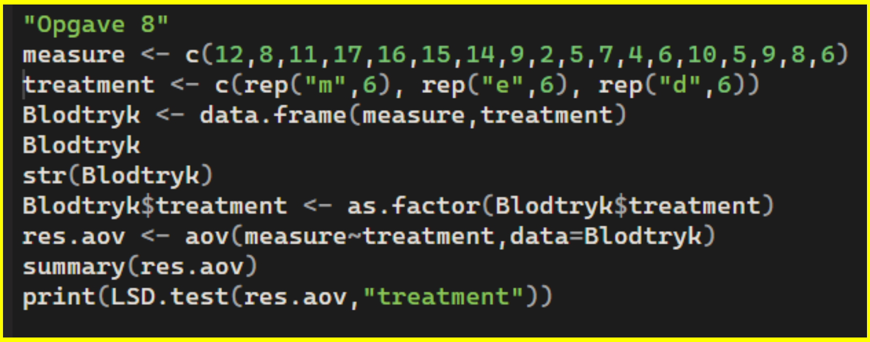
8) We want to evaluate three different methods to lower the blood pressure of individuals that have been diagnosed with high blood pressure. Eighteen subjects are randomly assigned to three groups (6 per group): the first group takes medication, the second group exercises, and the third one follows a 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

I. Vi har startet med at skrive alle scores op. Derefter har vi inddelt dem inde i 3 kategorier.

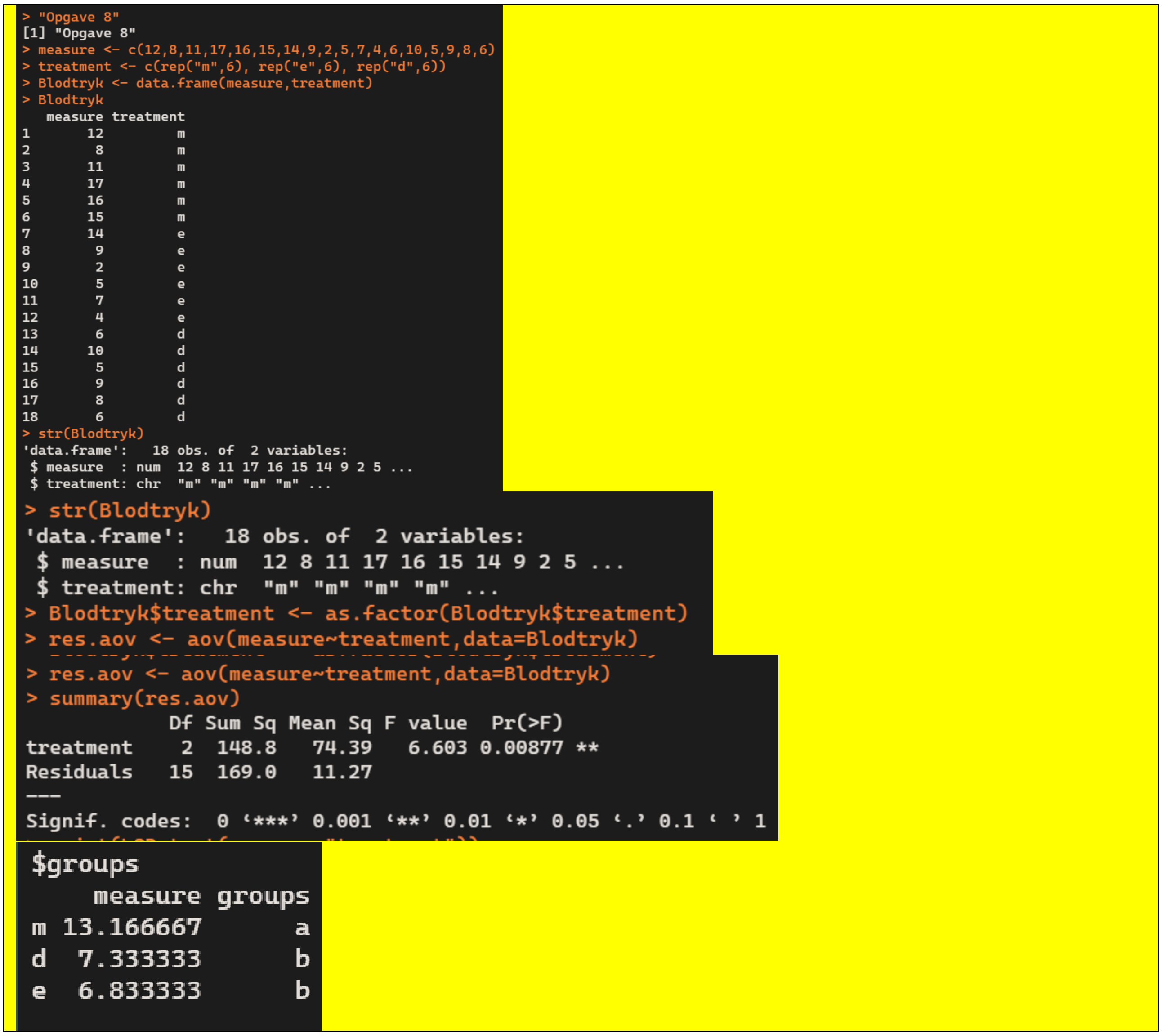
II. Derefter har vi sørget for at ændre de 3 variabler, medicine, exercise og diet er blevet til faktorer gennem as.factor().

III. Hvorimod har vi efterfølgende anvendt anova-metode, hvor der efterfølgende er anvendt LSD-metoden for at kunne se om værdierne sammenfletter sammen!



```
"Opgave 8"
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))
Blodtryk <- data.frame(measure,treatment)
Blodtryk
str(Blodtryk)
Blodtryk$treatment <- as.factor(Blodtryk$treatment)
res.aov <- aov(measure~treatment,data=Blodtryk)
summary(res.aov)
print(LSD.test(res.aov,"treatment"))
```

- Nu får vi følgende resultater!



```
> "Opgave 8"
[1] "Opgave 8"
> 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))
> Blodtryk <- data.frame(measure,treatment)
> Blodtryk
   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(Blodtryk)
'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" ...
```

```
> str(Blodtryk)
'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" ...
> Blodtryk$treatment <- as.factor(Blodtryk$treatment)
> res.aov <- aov(measure~treatment,data=Blodtryk)
```

```
> res.aov <- aov(measure~treatment,data=Blodtryk)
> 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
```