Statistisk Dataanalyse

Indholdsfortegnelse

Lektion 1

Opgave 1.1

The table below shows the height of students in classroom A (total of 15 students) and classroom B (total of 16 students), measured in centimeters. For each of the classroom, calculate the following:

a. Median

b. Mean

c. Mode

d. Midrange

Display the results on a table.

Løsning

A. For at finde medianen i R, anvendes *median()* funktionen på datasættet.

```
1   classA <-
    c(156,175,189,165,160,154,158,170,171,169,180,175,172,169,162)
2   classB <-
    c(185,175,169,182,179,163,191,182,180,174,161,180,176,174,182,173)
3   median(classA)
4   median(classB)
```

```
> median(classA)
[1] 169
> median(classB)
[1] 177.5
```

B. For at finde middelværdien i R, anvendes *mean()* funktionen på datasættet.

```
5   classA <-
    c(156,175,189,165,160,154,158,170,171,169,180,175,172,169,162)
6   classB <-
    c(185,175,169,182,179,163,191,182,180,174,161,180,176,174,182,173)
7   mean(classA)
8   mean(classB)
```

```
> mean(classA)
[1] 168.3333
> mean(classB)
[1] 176.625
```

C. For at finde typetallet i R, anvendes *sort()* funktionen til at sortere datasættet og derefter tælles værdierne.

```
9   sort(classA)
10  sort(classB)
```

```
> sort(classA)
 [1] 154 156 158 160 162 165 169 169 170 171 172 175 175 180 189
> sort(classB)
 [1] 161 163 169 173 174 174 175 176 179 180 180 182 182 182 185 191
```

D. For at finde Midrange i R, anvendes *sort()* funktionen til at sortere datasættet og derefter findes subtraktionen mellem største- og mindsteværdien.

```
9   sort(classA)
10  sort(classB)
```

```
> sort(classA)
 [1] 154 156 158 160 162 165 169 169 170 171 172 175 175 180 189
> sort(classB)
 [1] 161 163 169 173 174 174 175 176 179 180 180 182 182 182 185 191
> (189+154)/2
[1] 171.5
> (191-161)/2
[1] 15
```

Opgave 1.2

Find the mean of the following data:

20, 26, 40, 36, 23, 42, 35, 24, 30

Løsning

- Vi skal bruge *mean()* funktionen til at kunne bestemme middelværdien.

```
1   vektor1 <- c(20,26,40,36,23,42,35,24,30)
2   mean(vektor1)
```

```
> mean(vektor1)
[1] 30.66667
```

Opgave 1.3

Find the median of the following measurements:

713, 300, 618, 595, 311, 401, and 292

Løsning

- For, at kunne finde medianen anvendes derfor *median()* funktionen.

```
1   vektor2 <- c(713,300,618,595,311,401,292)
2   median(vektor2)
```

```
> median(vektor2)
[1] 401
```

Opgave 1.4

Find the median of the following measurements:

684, 764, 656, 702, 856, 1133, 1132, 1303

Løsning

- For, at kunne finde medianen anvendes derfor *median()* funktionen.

```
1   vektor3 <- c(684,764,656,702,856,1133,1132,1303)
2   median(vektor3)
```

```
> median(vektor3)
[1] 810
```

Opgave 1.5

Find the mode of the following measurements:

8, 9, 9, 14, 8, 8, 10, 7, 6, 9, 7, 8, 10, 14, 11, 8, 14, 11

Løsning

- For at kunne finde typetallet, anvendes *sort()* funktionen og derefter tælles de specifikke værdier.

```
1  vektor4 <- c(8,9,9,14,8,8,10,7,6,9,7,8,10,14,11,8,14,11)
2  sort(vektor4)
```

```
> sort(vektor4)
 [1]  6  7  7  8  8  8  8  8  9  9  9 10 10 11 11 14 14 14
```

Opgave 1.6

Find the mode of the following measurements:

110, 731, 1031, 84, 20, 118, 1162, 1977, 103, 752

Løsning

- For at kunne finde typetallet, anvendes *sort()* funktionen og derefter tælles de specifikke værdier.

```
1  vektor5 <- c(110,731,1031,84,20,118,1162,1977,103,752)
2  sort(vektor5)
```

```
> sort(vektor5)
 [1]   20   84  103  110  118  731  752 1031 1162 1977
```

Opgave 1.7

Find the midrange of these data:

2, 3, 6, 8, 4, 1

Løsning

- For, at kunne finde midrange anvendes derfor *sort()* funktionen hvor derefter lægges størsteværdien og mindsteværdien sammen og divideres med 2.

```
1  vektor6 <- c(2,3,6,8,4,1)
2  sort(vektor6)
3  (8+1)/2
```

```
> vektor6 <- c(2,3,6,8,4,1)
> sort(vektor6)
[1] 1 2 3 4 6 8
> (8+1)/2
[1] 4.5
```

Opgave 1.8

A researcher wants to collect data on 100 inhabitants living in one specific town. Classify the following collected variables according to their type (Nominal, ordinal, discrete, or continuous)

i. Occupation (“blue collar”, “white collar”, “unemployed”)

ii. Highest attained education (“low”, “medium”, high”)

iii. Monthly salary

iv. Civil status (“single”, “married”, “widow”)

v. Number of children

Løsning

I. Det er Norminal Variable.
II. Det er Ordinal Variable.
III. Det er Continuous Variable.
IV. Det er Norminal Variable.
V. Det er Discrete Variable.

Opgave 1.9

Evaluate the following statements as true or false:

a. In statistics, a population always refers to humans.

b. A sample is a subset of the study population.

c. Inferential statistics are statistical techniques used to draw conclusions about one specific sample.

d. A survey will be given to 100 students randomly selected from the freshmen class at Odense High School. The sample is all the freshmen at Odense High School.

Løsning

A. Det referer ikke kun til humans, men kan også refereres til andre ting.
B. Det er rigtigt, fordi et sample er lille håndfuld gruppe fra populationen.
C. Det er omvendt, men at en Inferential Statistik går ud på at man bruger sample til at lave konklusioner om Populationen.
D. Det er omvendt, men at de 100 elever fra Freshmen er sample størrelsen og alle freshmen fra Odense Gymnasium er Populationen.

Opgave 1.10

Find the range, variance and standard deviation for the data set for the samples of Brand A and Brand B paint.

Løsning

A. For, at kunne finde Variationsbredde anvendes *range()* funktionen.

```
1  brandA <- c(10,60,50,30,40,20)
2  brandB <- c(35,45,30,35,40,25)
3  range(brandA)
4  range(brandB)
```

```
> range(brandA)
[1] 10 60
> range(brandB)
[1] 25 45
```

B. For at kunne finde Variansen anvendes *var()* funktionen.

```
1  brandA <- c(10,60,50,30,40,20)
2  brandB <- c(35,45,30,35,40,25)
3  var(brandA)
4  var(brandB)
```

```
> var(brandA)
[1] 350
> var(brandB)
[1] 50
```

C. For at kunne finde Standardafvigelsen anvendes *sd()* funktionen.

```
1  brandA <- c(10,60,50,30,40,20)
2  brandB <- c(35,45,30,35,40,25)
3  sd(brandA)
4  sd(brandB)
```

```
> sd(brandA)
[1] 18.70829
> sd(brandB)
[1] 7.071068
```

Opgave 1.11

If the variance of a distribution is 9, the standard deviation is:

a. 3

b. 6

c. 9

d. 81

e. impossible to determine without knowing n.

Løsning

- Standardafvigelsen findes ved at tage kvadratroden af Variansen. Derfor anvendes *sqrt()* funktionen.

```
1   sqrt(9)
```

```
> sqrt(9)
[1] 3
```

Opgave 1.12

The standard deviation of a dataset is 10. If 5 were subtracted from each measurement, the standard deviation of the new

dataset would be:

a) 2

b) 10/25

c) 5

d) none of these.

Løsning

- 5 trækkes fra 10 og ned ad indtil 0. Derfor ender vi med et resultat på 0.
- Derfor er det ingen af delene ved Svar D.

```
1   > 10-5
2   [1] 5
3   > 9-5
4   [1] 4
5   > 8-5
6   [1] 3
7   > 7-5
8   [1] 2
9   > 6-5
10  [1] 1
11  > 5-5
12  [1] 0
13  > 4-5
14  [1] -1
15  > 3-5
16  [1] -2
17  > 2-5
18  [1] -3
19  > 1-5
20  [1] -4
21  > 0-5
22  [1] -5
23  5+4+3+2+1+0+(-1)+(-2)+(-3)+(-4)+(-5)
```

```
> 5+4+3+2+1+0+(-1)+(-2)+(-3)+(-4)+(-5)
[1] 0
```

Lektion 2

Opgave 2.1

The table below shows the height of students in classroom A (total of 15 students) and classroom B (total of 16 students), measured in centimeters.

a) Develop an ungrouped frequency table for all 31 students (1 table in total)

b) Construct a grouped frequency table for all 31 students (1 table in total)

c) Plot the frequencies of each class for all 31 students (1 histogram in total)

Løsning

A. Så når vi snakker om en Ungrouped Frequency Table, så snakker vi oftest om at lave en tabel hvor vi indskriver værdierne fra datasættet og tæller hvor mange de optræder i den anden kolonne. Vi starter allerførst med, at sortere rækkefølge og derefter vises eksemplet nedenfor:

Klasse A	156,175,189,165,160,154,158,170,171,169,180,175,172,169,162
Klasse A (Sorteret)	154,156,158,160,162,165,169,169,170,171,172,175,175,180,189

Klasse B	185,175,169,182,179,163,191,182,180,174,161,180,176,174,182,173
Klasse B (Sorteret)	161,163,169,173,174,174,175,176,179,180,180,182,182,182,185,191

- Nu laver vi en ungrouped frekvens tabel:

Ikke-grupperede Frekvens Tabel for Klasse A												
154	156	158	160	162	165	169	170	171	172	175	180	189
1	1	1	1	1	1	2	1	1	1	2	1	1

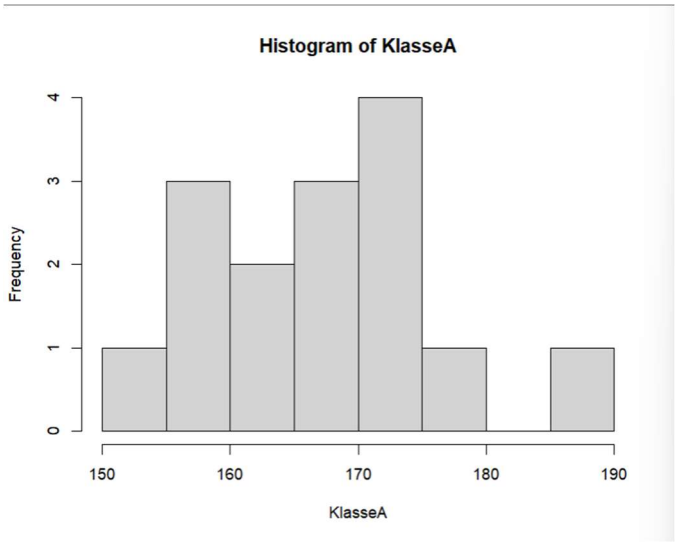
Ikke-grupperede Frekvens Tabel for Klasse B											
161	163	169	173	174	175	176	179	180	182	185	191
1	1	1	1	2	1	1	1	2	3	1	1

B. Nu skal vi lave en grouped frekvens tabel og dette er ved at danne intervaller fra A til B, og fra C til D osv. Her inde i de forskellige intervaller skal vi indsætte værdier på, hvor mange gange den samme værdi fra datasættet optræder flere gange.

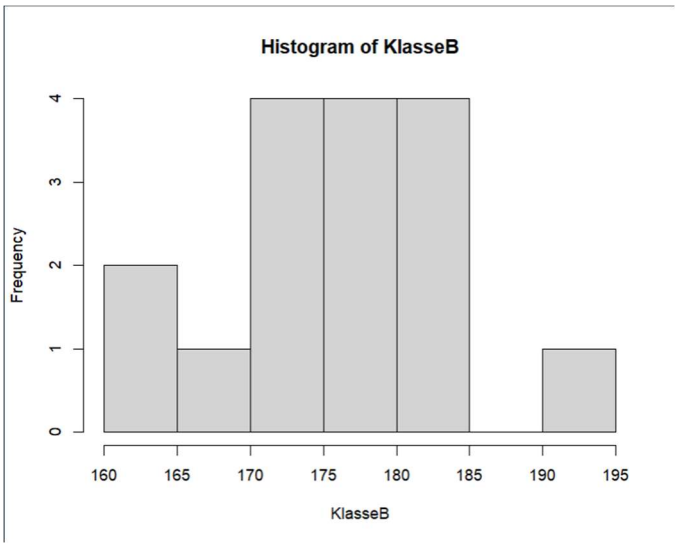
Grupperet Frekvens Tabel for Klasse A				
(154-160(	(160-165(	(165-170(	(170-175(	(175-189)
3	2	3	3	4

Grupperet Frekvens Tabel for Klasse B				
(160-165(	(165-170(	(170-175(	(175-180(	(180-191)
2	1	3	3	7

C. Følgende Histogrammer er dannet gennem *hist(x=dataframe)* funktionen.



- Her har vi lavet det for Klasse B.



Opgave 2.2

The distribution of entrance test scores of freshmen in a particular university has the following percentile scores. How may the distribution be described?

a. Symmetrical bell-shaped

b. Skewed left (negatively skewed)

c. Skewed right (positively skewed)

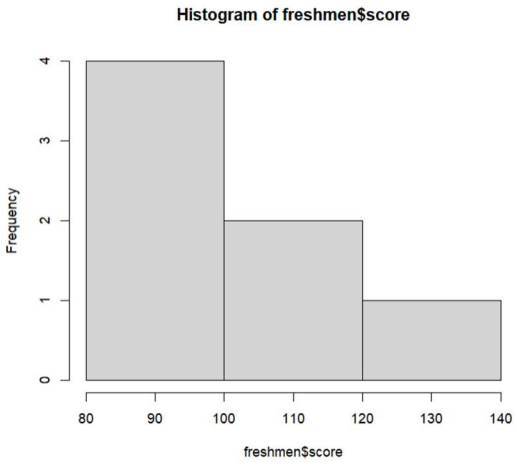
d. Impossible to tell from the above

Løsning

- For, at kunne find ud af skævheden. Skal vi danne Histogram fra *hist(x=dataframe)* funktionen.

```
1  freshmen <-
   data.frame(percentile=c(95,80,65,50,35,20,5),score=c(140,120,101,94,91
   ,87,80))
2  hist(x=freshmen$score)
```

- Vi kan se, at følgende diagram bliver dannet som er Positiv Højre Skævhed.



Opgave 2.3

The cumulative frequency graph below shows the salaries of 100 employees who work for Welsh Bank (black) and 100 employees who work for the Bank of Finland (blue).

Based on the given graph, evaluate the following sentences as TRUE or FALSE:

i. The interquartile range of the data for the Bank of Finland is 60000

ii. The median for the Welsh Bank is £62000

iii. 78000 is an outlier for the Bank Welsh

iv. The range for both banks is the same

Løsning

I. Vi starter med at kigge på 75% og 25% ved andenaksen og derved kigger vi førsteaksen og ser inputtet.

- Vi kan se at Q3=69 og Q1=50, hvor Kvartilbredden bliver 69-50=19.
- Derfor kan vi sige at det er falsk.

II. Vi kan se, at medianen er 51 og ikke 62.

III. Vi finder Outlier ved at først finde den rigtige kvartilbredde for Welsh Bank.

- $61 + 1{,}5 \cdot (16) = 85$
- $45 - 1{,}5 \cdot (16) = 21$
- Vi kan se, at 78 befinder sig imellem 21 < x < < 85.

IV. Vi kan se, at Variationsbredden er den samme for begge Banker. (80-20=60).

Opgave 2.4

If the mean, median and mode of a distribution are 8, 7, 6 respectively, then the distribution is:

a. negatively skewed

b. not skewed

c. positively skewed

d. symmetrical

e. bimodal.

Løsning

- Vi kan se, at vores middelværdi er større end medianen.
- Vi kan se at middelværdien er 8 og medianen er 7.
- Derfor er der ift. Præsentationen snak om en Positiv Højre Skævhed.

Opgave 2.5

The following Type 1 boxplot was drawn using a list of numbers.

What is the incorrect statement regarding this boxplot?

a) The number 1 must be in the list of numbers from which the plot was drawn.

b) The dataset has no median.

c) The boxplot could be derived from the following dataset: 1, 1, 5, 5, 7

d) More than half of the data falls between 1 and 5.

e) The range is 6.

Løsning

- Vi kan se, at vores Boksplot model viser os at vores størsteværdi er svarende til 7.
- Vi kan se, at vores mindsteværdi er 1 og 3. Kvartil er 5.
- I Datasættet kan det ses at den midterste tal er 5, som befinder sig i ulige talrække.

Opgave 2.6

A teacher gives a 20-point test to 10 students. Find the percentile rank of a score of 12.

18, 15, 12, 6, 8, 2, 3, 5, 20, 10

Løsning

- For at kunne løse denne opgave, er det nødvendigt at sortere rækkefølgen.

```
1  teacher <- c(18,15,12,6,8,2,3,5,20,10)
2  sort(teacher)
```

```
> sort(teacher)
 [1]  2  3  5  6  8 10 12 15 18 20
```

- Vi kan se, at for at kunne løse denne opgave skal vi tælle antal pladser op til tallet før 12 som er 10.

- Derfor kan vi anvende den selviske koncept gennem udregningen nedenfor og derved sige at 60% af eleverne har fået score under 60%.

$$\frac{6}{10} \cdot 100 = 60\%$$

Opgave 2.7

A teacher gives a 20-point test to 10 students. Find the value corresponding to the 25th percentile.

18, 15, 12, 6, 8, 2, 3, 5, 20, 10

Løsning

- Vi skal bruge den sorterede rækkefølge fra den sidste opgave.

```
1   teacher <- c(18,15,12,6,8,2,3,5,20,10)
2   sort(teacher)
```

```
> sort(teacher)
 [1]  2  3  5  6  8 10 12 15 18 20
```

- Nu skal vi anvende Procentil formlen.

$$L_k = \frac{k}{100} \cdot (n+1)$$

$$L_k = \frac{25}{100} \cdot (10+1) = \frac{11}{4} = 2{,}75$$

- Vi kan se, at vores 25th procentil befinder sig imellem 2 og 3.

Opgave 2.8

Find Q1, Q2, and Q3 for the data set.

15, 13, 6, 5, 12, 50, 22, 18

Løsning

- Vi starter med, at anvende *sort()* funktionen til at kunne sortere rækkefølgen.

```
1   dataset <- c(15,13,6,5,12,50,22,18)
2   sort(teacher)
```

```
> sort(dataset)
 [1]  5,6,12,13,15,18,22,50
```

- Nu finder vi medianen af de to tal i midten, da vi har lige talrække.

$$\frac{5+12}{2} = \frac{17}{2} = 8{,}5$$

- Nu skal vi finde 1. Kvartil.

$$\frac{13+6}{2} = \frac{19}{2} = 9{,}5$$

- Nu skal vi finde 3. Kvartil.

$$\frac{50+22}{2} = 36$$

Opgave 2.9

The mean of the population of ten scores, 78, 91, 91, 94, 74, 23, 63, 22, 78, 89 is 70.3, and the modes are 78 and 91. The skewness of the population is:

a. negative

b. zero

c. positive

d. not determined

e. positive or negative depending on the score.

Løsning

- Vi kan se, at der skal anvendes *median()* funktionen til at kunne løse opgaven.

```
1  set <- c(78,91,91,94,74,23,63,22,78,89)
2  median(set)
```

```
> median(set)
 [1]  78
```

- Derefter sammenlignes middelværdien med medianen og her kan det ses at medianen er større end medianen.
- Vi kan se, at: 70.3<78 som er Negativ Venstre Skævhed.

Opgave 2.10

A percentile score of 40 indicates that a person:

a. answered 40% of the questions correctly on the test.

b. knows 40% of the material covered by the examination.

c. has earned a score equal to or better than 40 persons in his class.

d. has earned a score equal to or better than 40% of the persons in his class.

Løsning

- Vi skal huske at anvende den selviske princip hvor vi er bedre end andre i procent.
- Derfor er svaret D.

Opgave 2.11

Construct two boxplots ("Type 1" and "Type 2") for the data.

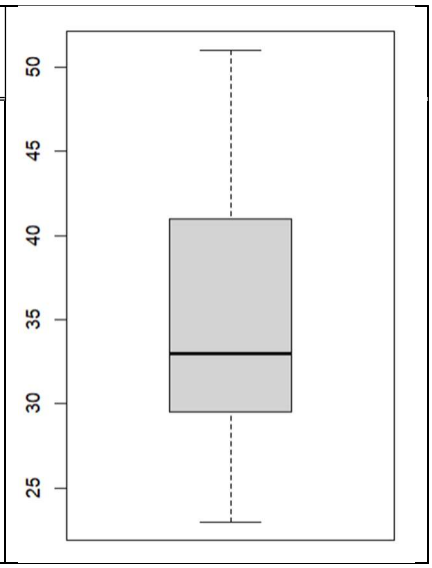
33, 38, 43, 30, 29, 40, 51, 27, 42, 23, 31

Løsning

- For, at besvare opgaven opskrives *boxplot()* funktionen, hvor derved indsættes data.

```
1  typer <-
   c(33,38,43,30,29,40,51,27,42,23,31))
2  boxplot(typer)
```

```
> boxplot(typer)
```



Opgave 2.12

In this data set:

15, 13, 6, 5, 12, 50, 22, 18

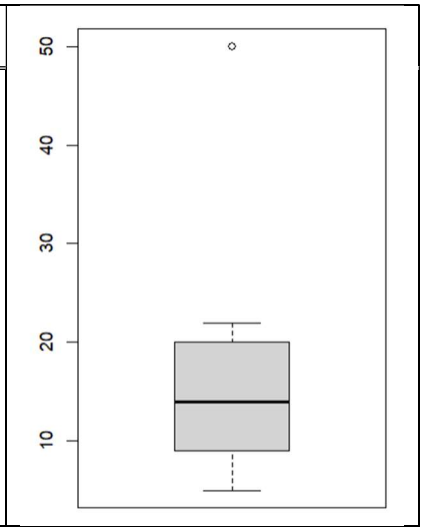
Is there any outlier?

Løsning

- For, at kunne besvare opgaven anvendes *boxplot()* funktionen, hvor derefter kigges på om der findes en cirkular prik udenfor bokxplot grafen.

```
13 outlier <- c(15,13,6,5,12,50,22,18)
14 boxplot(outlier)
```

```
> boxplot(outlier)
```



Opgave 2.13

Twenty-five people were given a blood test to determine their blood type.

Raw Data: A,B,B,AB,O O,O,B,AB,B B,B,O,A,O A,O,O,O,AB AB,A,O,B,A

a) Can you construct a histogram? Can you construct a bar graph?

b) Considering your reply in item a, construct the correct graph

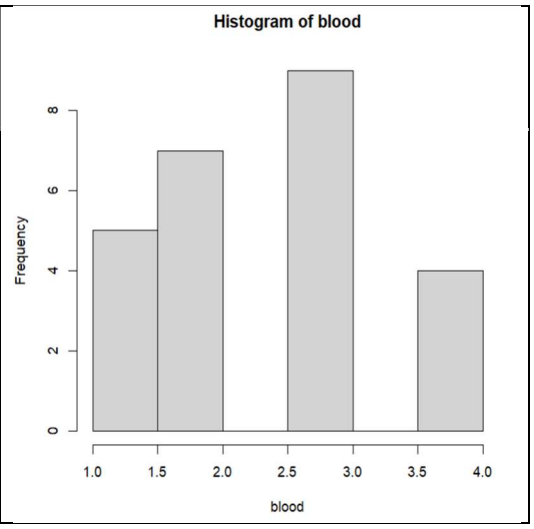
Løsning

A. Vi kan bare definere de adskillige bogstaver med adskillige tal.
- Vi kan eksempelvis definere A=1, B=2, O=3, AB=4

B. Nu kan vi udefra de definerede værdier, lave en vektor som indeholder datarækken og danne en histogram.

```
1  blood <-
   c(1,2,2,4,3,3,3,2,4,2,2,2,3,1,3,1
   ,3,3,3,4,4,1,3,2,1)
2  hist(x=blood)
```

```
> hist(x=blood)
 [1]  78
```



Lektion 3

Opgave 3.1

Find the sample space for the gender of the children if a family has three children. Use B for boy and G for girl.

Løsning

- Vi starter allerførst med at definere B for Dreng og G for Pige.

Udfald	1 B og 2 G	2 B og 1 G	3 B	1 G og 2 B	2 G og 1 B	3 G
Muligheder	BGG	BBG	BBB	GBB	GGB	GGG

Opgave 3.2

If a family has three children, find the probability that all the children are girls.

Løsning

- Nu skal vi danne kombinationerne for dannelsen af fødte børn i forhold til Køn.

Udfald	1 B og 2 G	2 B og 1 G	3 B	1 G og 2 B	2 G og 1 B	3 G
Muligheder	BGG	BBG	BBB	GBB	GGB	GGG

- Vi kan se, at vi har sorteret de adskillige muligheder som opstår flere end en gang.

Udfald	1 B og 2 G	2 B og 1 G	3 B	3 G
Muligheder	BGG	BBG	BBB	GGG

- Vi kan se udefra Kombinationerne, at sandsynligheden for at få 3 Piger er ¼ da vi har 4 kombinationer, hvor 3 piger er født i en familie.

Opgave 3.3

If the probability that a person lives in an industrialized country of the world is 1/5, find the probability that a person does not live in an industrialized country.

Løsning

- Vi kan se, at fordi 1/5 er sandsynligheden for at leve i en industrialiseret verden.
- Men de muligheder som mangles tilbage i tælleren mod, er den sandsynlighed, som findes ved ikke at leve i en industrialiseret verden.
- Derfor kan vi sige, at vores sandsynlighed er 4/5.

Opgave 3.4

The table below contains information on the number of daily emergency service calls received by the volunteer ambulance service of Happytown for the last 50 days: 22 days of which 2 emergency calls were received, 9 days of which 3 emergency calls were received, 8 days of which no emergency calls were received, etc.

Number of Service Calls per Day (X)	Number of Days (fj)
0	8
1	10
2	22
3	9
4	1
Total	50

What is the probability that 2 or more emergency calls are received on a day?

Løsning

- Vi kan se, at fordi vi skal finde sandsynligheden for at der er opstået flere opkald end 2 (eller) flere opkald.
- Derfor kan vi sige at hvis dagene uden for antal opkald, lægges sammen kan vi derfor finde sandsynligheden for 2 eller flere opkald.
- 22+9+1=32 og derved 32/50.

Opgave 3.5

Determine which of the following distributions is a cumulative probability distribution.

i)

X	1	2	3	4	5
P(X≤x)	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{9}{20}$	$\frac{6}{8}$	$\frac{20}{20}$

ii)

X	22	33	44	55	66	77
P(X≤x)	-0.4	0.2	0.4	0.7	0.8	1

iii)

X	0	3	5	6
P(X≤x)	0.25	0.25	0.25	0.25

iv)

X	0	2	4	6
P(X≤x)	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{6}$	$\frac{4}{16}$

Løsning

- Vi kan se, at billedet ii er vores svar da vi kan se at law of large numbers gælder herhenne.
- Dette betyder, at vi kan i vores tilfælde se at jo mere vi kaster med en terning, desto mere end vi med en præcist udfald som forventet.
- Her kan vi se, at vi har kastet 0,4 men vi er end med en 0,4 igen på billedet.

Opgave 3.6

The grades of a group of 1000 students in an exam are normally distributed with a mean of 70 and a standard deviation of 10. Approximately, how many students have grades greater than 80?

a) 680 students

b) 840 students

c) 160 students

d) 50 students

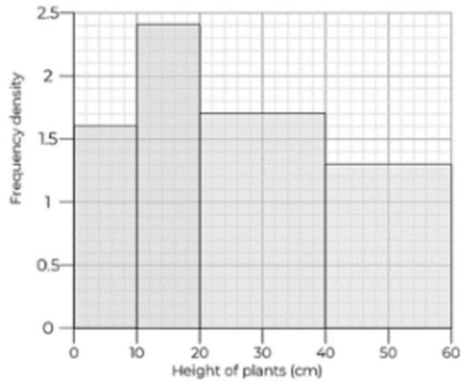
e) 320 students

Løsning

- Vi skal forestille os at vi har en bølgegraf hvor middelværdien er placeret midterst som er x = 70.
- Vi kan se, at hvis man danner to intervaller så kan man lægge standardafvigelsen sammen med middelværdien (mod højre) og på den anden side kan vi se at (mod venstre) kan vi trække standardafvigelsen fra middelværdien.
- 70+10=80
- 70-10=60
- Hvis vi forsøger at udvide vores interval mod højre, så kan det ses at vi har 80+10=90.
- Vi kan se, at i den første interval har vi 34,134 og derefter har vi 13,591 som er den næste sammen med 2,14.
- Hvis selve 13,591 lægges sammen med 2,14 => 13,591+2,14=15,731 som er ca. 16%.
- Men hvis man kigger på det samlede procent: (34,134*2)+13,591+2,14=83%
- Fordi vi kan se, at 16% ligger efter intervallet ved 80, bliver det derfor 16/100 = 0,16 * 1000 = 160.

Opgave 3.7

We visited a field with plants. We measured the height of each plant and built the histogram below. We then chose one plant at random. What is the probability that the plant is under 30 cm tall?



Løsning

- Vi kan I vores tilfælde se, at når vi snakker om en plante under 30 cm, så er det med 30 inkluderet.
- Vi kan se på x-aksen at vi har 7 muligheder, men 4 udfald som vi ønsker at sandsynliggøre.
- Derfor kan vi sige, at vores sandsynlighed ender med at være 4/7.

Opgave 3.8

M&M sweets are of varying colors and the different colors occur in different proportions. The table below gives the probability that a randomly chosen M&M has each colour, but the value for orange candies is missing.

Color	Brown	Red	Yellow	Green	Orange
Probability	0.2	0.3	0.2	0.1	?

You draw an M&M at random from a packet. What is the probability that you get either a green one or an orange one?

Løsning

- Så vi skal huske I vores tilfælde at den nederste række er svarende til sandsynligheden.
- Vi kan i vores tilfælde se at sandsynligheden som vi kender for at slå en orange er 1/5=0,2.
- Fordi vi nu kender sandsynligheden allerede for den grønne og nu for den orange, kan vi bare lægge dem sammen og beregne den samlede sandsynlighed for begge
- I tilfældet, kan det ses at sandsynligheden for dem begge er 0,1+0,2=0,3.

Opgave 3.9

What percent of cases are likely to be between 85 and 93 in a normal distribution with mean 87 and variance 4?

a. 83.85% d. 69.02%

b. 30.72% e. none of these

c. 49.87%

Løsning

- VI kan se, at middelværdien er 87 og at variansen er 4.
- For at finde standardafvigelsen ved vi at kvadratroden af variansen er vores standardafvigelse, derfor bliver standardafvigelsen 2.
- Men hvis man lægger middelværdien sammen med vores standardafvigelse, kan vi se at det bliver 89 og hvis der tilføjes yderligere 2 på så er det 91. Det samme additionsprincip gælder for 91, som giver os 91+2=93.
- Nu har vi dannet vores intervaller, langt så godt.
- Vi ved, at fra den ene standardafvigelse og gennem middelværdien til den anden standardafvigelse er 68,318%.
- Men hvis der tilføjes yderligere 13,154 mellem intervallet for 89 og 91 - og derved 2,14 mellem 91 og 93. Så kan følgende procent opnås.
- 68,314+13,154+2,14 = 83,608.

Opgave 3.10

A survey found that one out of five Americans say he or she has visited a doctor in any given month. If 10 people are selected at random, find the probability that exactly 3 will have visited a doctor last month.

Løsning

- $\frac{10!}{3!\cdot(10-3)!}\cdot\left(\frac{1}{5}\right)^3\cdot\left(\frac{4}{5}\right)^7=\frac{393216}{1953125}\approx 0{,}2013266\cdot 100\%=20{,}132\%$
- Der er 20,132% sandsynlighed for at der var nøjagtig 3 der besøgte lægen sidste måned

Opgave 3.11

Suppose a loaded die has the following model:

Face	1	2	3	4	5	6
Probability	0.3	0.1	0.1	0.1	0.1	0.3

If this die is thrown and the top face shows an odd number,

a. What is the probability that the die shows a four?

b. What is the probability that the die shows a 1?

Løsning

A. S vi kan se, at vi bliver spurgt om en terningekast med udfald med ulige numre. Derefter spørger de, hvad sandsynligheden er at hvis sandsynligheden viser 4. I vores tilfælde er det 0/6, eller bare 0 fordi 4 er ikke en del af ulige udfald.

B. Fordi vi allerede har fået at vide, at vores muligheder er ulige tal, derfor har vi 3 muligheder. Fordi vi har dobbelt sandsynlighed for at slå en præcist udfald, kan vi derfor sige at vores sandsynlighed bliver 0,3*2 fordi der er dobbelt sandsynlighed for at slå en 1'er.

Lektion 4

Opgave 4.1

1. 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

   Which one of these 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.

Løsning A

Step 1:

- Vi starter allerførst med at opskrive dataet i vektoren som hedder email.
- Derefter anvendes shapiro.test til at kunne finde ud af om vores middelværdi egentligt er signifikant indenfor de kritiske områder eller ej?

```
1  email <- c(5460,5900,6090,6310,7160,8440,9930)
2  Shapiro.test(email)
```

```
> email <- c(5460,5900,6090,6310,7160,8440,9930)
> shapiro.test(email)

    Shapiro-Wilk normality test

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

- Vi kan i vores tilfælde se, at vores p-værdi er ikke signifikant hvilket betyder at den ligger inden for de 99% confidenslevel som er acceptering af null-hypotesen.

Step 2:

- Nu anvendes t.test funktionen til at kunne udføre One-Sample T-test.

- Det skal lige understreges, at vi har bare navnet givet vores "assigning vektor" res.ttest, men den kan godt blive kaldt på noget andet med et andet navn.

```
1   res.ttest <- t.test(email,conf.level=0.99)
2   res.ttest
```

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

    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
```

- Vi kan se, at vores p-værdi er ikke signifikant her, men det skal vi bare kigge bort fra.
- Kigger vi overimod middelværdi af x, så kan det ses at det er 7041,429 som ligger inden for 4784,991 og 9297,866.
- Dette betyder, at vores ligning kommer til at se sådan ud: 4784,991 < 7041,429 < 9297,866.
- Dette betyder, at vi kan med 99% garanti sige at middelværdien af antal gange enhver ansat tjekker deres email er liggende i intervallet mellem 4784,991 og 9297,966.

Løsning B

Step 1:

- Vi kan se, at vi bliver spurgt om at sammenligne antal gange de professionelle tjekker deres email, sammen med de ansatte og derved se om der er nogen forskel.
- Vi ved allerede at de professionelle ansatte tjekker deres email 15 gange på en dag.
- Hvorimod så ved vi, at de ansatte tjekker deres email i løbet af et år.
- Fordi vi har fået oplyst en dag og årsforhold, kan vi derfor gange antallet af gange de professionelle ansatte tjekker deres med mail, med de antal dage der er på et år.
- Dette betyder, at vi har 15 x 365 = 5475.
- Hvis vi placerer vores resultat inde i det givende interval, får vi følgende: 4784,996 < 5475 < 9297,966. (Værdien passer ind!)
- Derfor kan vi med 99% garanti sige, at der er ingen signifikant forskel mellem den gennemsnitlige professionelle og de ansatte.

Opgave 4.2

2. A job advisor claims that the average salary for engineers is 24.000 euros/year. Ten engineers are randomly selected and the mean of their salaries is 23450 euros/year and a standard deviation of 400 euros/year. Is there evidence (at 95% confidence level) to reject the statement of the job advisor?

Løsning

- Vi kan se, at den gennemsnitlige løn for ingeniør i populationen er 24,000 og vi kan se at for sample gruppen er det 23450. Vi kan med åbne øjne, derfor sige at der er forskel i løn mellem en stor gruppe og en lille fraktion af den store gruppe.
- Derfor kan null-hypotesen afvises og alternativ hypotesen accepteres, og derved kan job-advisors påstand afvises.

Opgave 4.3

3. 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.

Løsning

- Vi starter allerførst med at finde normalfordelingen gennem shapiro.test.

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

```
> shapiro.test(towns)

    Shapiro-Wilk normality test

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

- Derefter anvendes t.test for One-Sample t.test og her undersøger vi om vores middelværdi ligger indenfor de kritiske intervalområder.

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

```
> 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
```

- Vi kan I tilfældet her se, at vores middelværdi ligger inden for de konfidensintervaller.
- Derfor kan vi med 99% garanti sige, at konfidensintervallet for børn født i de 7 byer er: 6505,022<7834,714<9164,407.

Opgave 4.4

4. A survey of 50 students found that the mean age of their bicycles was 5.6 years. Assuming the standard deviation of the population is 0.8 years, which of these statements is correct?

   A. Based on the sample of 50 students, we can be 90% confident that the mean age of all bicycles is between 5.297 and 5.903.
   B. Based on the sample of 50 students, we can be 99% confident that the mean age of all bicycles is between 5.297 and 5.903.
   C. A and B are correct.
   D. None of them is correct.

Løsning for A.

- Vi kender allerede på forhånd, at vores middelværdi for de 50 elevers cykler ligger på 5,6 år.
- Vi ved også, at i Inferential Statistik er det udefra samplet vi kan danne konklusioner om vores Population.
- Vi ved også, at selve Standardafvigelsen ligger på 0,8 og kvadrer vi denne tal med 2 får vi Variansen som er 0,64. (Der er ikke meget forskel i tallene).
- Vi kan placere vores sample mean mellem de to givede konfidensintervaller ($5{,}297 < x < 5{,}903$) og derved kan vi rent faktisk sige at der er 90% garanti for at middelværdien af cyklerne passer i de givede intervaller.

Løsning for B.

- Vi kan se, at det samme princip gælder her, men i stedet for 90% garanti, er det nu blevet til 99% sikkerhed.
- Fordi vores sample middelværdi passer ind i de givende intervaller, kan vi derfor med 99% sikkerhed sige at den gennemsnitlige alder af cyklerne er mellem 5,297 og 5,903.

Løsning for C

- Ude fra vores beregninger og forklaringer, kan vi sige at C'eren er den mest korrekte svar ud af de 4.

Opgave 4.5

5. The following data represent a sample of the assets (in millions of euros) of 30 motherboards computer hardware manufacturers in Europe. Find the 90% confidence interval of the mean.

12.23	16.56	4.39
2.89	1.24	2.17
13.19	9.16	1.42
73.25	1.91	14.64
11.59	6.69	1.06
8.74	3.17	18.13
7.92	4.78	16.85
40.22	2.42	21.58
5.01	1.47	12.24
2.27	12.77	2.76

Løsning

- Vi starter allerførst med at opskrive vores data og derefter anvendes shapiro wilk testen til at kunne find ud af om der er snak om en signfikant forskel eller ej.

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

```
> shapiro.test(motherboards)

    Shapiro-Wilk normality test

data:  motherboards
W = 0.63604, p-value = 2.071e-07
```

- Nu kan One-Sample T-test anvendes ved t.test kommandoen og derved kan vi indsætte vores middelværdi inde mellem de kritiske konfidensintervaller.

```
res.ttest <- t.test(motherboards,conf.level=0.90)
res.ttest
```

```
> res.ttest <- t.test(motherboards,conf.level=0.90)
> res.ttest

    One Sample t-test

data:  motherboards
t = 4.2169, df = 29, p-value = 0.0002214
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
  6.621856 15.559478
sample estimates:
mean of x
 11.09067
```

- Vi kan I vores tilfælde se at vores populationsmiddelværdi kan placeres inde mellem de 2 intervaller 6,62 og 15,55.
- Derfor bliver intervallet: 6,62 < 11,09 < 15,55.
- Dette betyder, at vores null-hypotese er accepteret og at vi kan med 90% sige at vores middelværdi er placeret inde i de givende konfidensintervaller.

Lektion 5

Opgave 5.1

1. The average relative humidities (%) for two cities, are: 72.9 (city A) and 70.8 (city B), based on 25 measurements of relative humidity in each city. The standard deviations of these measurements are: 2.5 (city A) and 2.8 (city B). Based on the samples, can it be concluded with 95% confidence level that the relative humidity in the two cities is significantly different?

Løsning

- Vi kan se, at vi har to middelværdier som vi skal sammenligne med hinanden.
- Vi kan se, at by A har middelværdien 72,9 og by B har middelværdien 70.8.
- Vi kan i vores tilfælde se, at Standardafvigelse for by A = 2,5 og for by B = 2,8.
- Vi kan se, at Variansen for by A = 2,5*2,5=6,25 og for by B er det 7,84.
- Fordi vores Varians fortæller, hvordan vores data afviger fra hinanden og at vi kan se at variansen ligger tæt ved 5% sandsynlighed for at afvise null-hypotesen.
- Derfor kan vi sige, at nedbørsforholdet er signifikant forskellige blandt de to byer.

Opgave 5.2

2. The dean of a university claims that the average scores in Software Engineering education, of those students that were educated in public high schools is higher than the average scores of those students that were educated in private high schools. A sample of the 50 students from each group is randomly selected. The average scores of the students from public schools are 8.6 (out of 10) and with a standard deviation equal to 3.3. The average scores of the students from private schools are 7.9 (out of 10) and with a standard deviation equal to 3.3. Are we 90% confident that the statement of the dean is true?

Løsning

- Vi skal finde ud af i denne opgave, om der egentligt er en forskel mellem en kommunaldrevet skole og en privatskole ift. Karaktergennemsnit.
- Vi kan se, at karaktergennemsnittet for kommunaldrevet skole er 8,6, hvorimod karaktergennemsnit for privatskole er 7,9.
- Vi kan se, at vi begge skole har den samme Standardafvigelse og fordi Variansen fortæller hvor meget dataene afviger fra hinanden kan vi se at 3,3*3,3 = 10,89.
- Fordi vi kan se, at Variansen afviger med 10,89 så betyder det at der er 90% sandsynlighed for at påstand af dekanen er ikke sandt.

Opgave 5.3

3. We would like to know if the concentration of a compound in two brands of yogurt is different. We select 50 bottles of each type. The average concentration in one of the brands is 88.42 mg/L and in the other one is 80.61 mg/L. The standard deviations of the populations are 5.62 mg/L and 4.83 mg/L, respectively. Can we be 95% confident that there is a significant difference among the two brands? What about 99% confident?

Løsning for 95

- Vi kan se, at vi skal sammenligne to Yoghurt Mærker med hinanden. Derfor er det Independent T-test.
- Vi kan i vores tilfælde se, at selve den gennemsnitlige koncentration af Brand A er 88,42.
- Hvorimod for Brand B er den gennemsnitlige koncentration svarende til 80,61.
- I vores tilfælde, kan det ses at selve standardafvigelsen for Brand A er 5,62.
- Herunder er Variansen for Brand A svarende til: 5,62*5,62=31,58.
- Hvorimod standardafvigelsen for Brand B er 4,83.
- Herunder kan det ses at Variansen for Brand B er svarende til 4,83*4,83=23,32.
- Vi kan her se, at vores Varians tal er meget forskellige for de 2 Brands og derved kan vi sige med 95% at der er en signifikant forskel mellem de to brands.

Løsning for 99

- Vi kan implementere den samme princip for selve sandsynligheden med 99%.
- Fordi vi har så store Varians tal, kan vi derfor sige at der er 99% sikkerhed for at der er forskel mellem de to brands af Yoghurt.

Opgave 5.4

4. We want to know whether or not a certain training program is able to increase the maximum long jump of athletes. We recruit a simple random sample of 20 long jump athletes and measure each of their maximum long jump. Then, we have each athlete use the training program for one month and then measure their maximum long jump again at the end of the month. These are the results (below). Does the training program have any effect on the maximum long jump? (use level of significance = 0.05)

Athlete	Maximum long jump before training program	Maximum long jump after training program
1	3.7	4.0
2	3.3	3.7
3	3.2	3.2
4	4.0	3.7
5	4.2	4.7
6	4.2	4.3
7	4.7	4.7
8	3.7	4.0
9	5.0	5.0
10	4.5	4.8
11	4.0	4.2
12	3.0	3.3
13	2.7	2.8
14	3.2	3.0
15	3.2	3.0
16	4.7	4.7
17	4.0	4.3
18	4.2	4.5
19	4.2	4.5
20	3.8	4.0

Løsning

- Vi anvender t.test funktionen sammen med paired til at kunne find ud af om de er signifikante eller ej.

```
athlete <-
data.frame(before=c(3.7,3.3,3.2,4.0,4.2,4.2,4.7,3.7,5.0,4.5,4.0,3.0,2.7
,3.2,3.2,4.7,4.0,4.2,4.2,3.8),after=c(4.0,3.7,3.2,3.7,4.7,4.3,4.7,4.0,5
.0,4.8,4.2,3.3,2.8,3.0,3.0,4.7,4.3,4.5,4.5,4.0))
res.ttest <- t.test(athlete$before,athlete$after,paired=TRUE)
res.ttest
```

```
> res.ttest <- t.test(athlete$before,athlete$after,paired=TRUE)
> res.ttest

    Paired t-test

data:  athlete$before and athlete$after
t = -2.997, df = 19, p-value = 0.007411
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -0.24626407 -0.04373593
sample estimates:
mean difference
         -0.145
```

- Sjovt nok, kan vi se at vores p-værdi ligger under 0,05. Dette betyder, at der er en signifikant forskel i længden før og efter.
- Men sjovt nok, kan det ses at R viser os en enkelt middelværdi som ligger midt i et interval, men husk at vi beskæftiger os med at sammenligne middelværdier af to grupper.
- Derfor er det nødvendigt at understrege, at der skal være middelværdi af x og y og ikke kun af x. Dette er grunden til at vi ikke tager resultat af x i betragtning.

Lektion 6

Opgave 6.1

1. We want to investigate if there is an effect of the type of fertilizer applied to apple trees and the production of apples. We randomly select 15 trees and randomly assign them to one of three groups (5 trees per group). We perform a test in which we apply one type of fertilizer (fertilizer 1, 2 or 3) to each group. The data are shown. At a α=0.05, can it be concluded that there is a significant difference in the production of apples depending on which fertilizer is used? Which fertilizer/fertilizers causes a higher/lower production than the other/others?

Fertilizer 1	Fertilizer 2	Fertilizer 3
10	6	5
12	8	9
9	3	12
15	0	8
13	2	4

Løsning

- Vi starter allerførst med at først med at opskrive alle dataerne i form af målinger.
- Derefter opdeler vi dataet, så vi kan behandle dem.
- Til sidst samler vi dem sammen i en dataframe ramme og derefter undersøger om de er læst korrekt.

```
1   fertilizer <-
    data.frame(f1=c(10,12,9,15,13),f2=c(6,8,3,0,2),f3=c(5,9,12,8,4))
2   mean(fertilizer$f1)
3   mean(fertilizer$f2)
4   mean(fertilizer$f3)
```

```
> fertilizer <-
data.frame(f1=c(10,12,9,15,13),f2=c(6,8,3,0,2),f3=c(5,9,12,8,4))
> mean(fertilizer$f1)
[1] 11.8
> mean(fertilizer$f2)
[1] 3.8
> mean(fertilizer$f3)
[1] 7.6
```

- Vi kan se, at for kunstgødning 1 = 11,8, kunstgødning 2 = 3,8, og kunstgødning 3 = 7,6.
- Vi kan dermed se, at de forskellige kunstgødning 1 adskiller sig fra kunstgødning 2 og 3 og at kunstgødning 1 er den som producerer mest.

Opgave 6.2

2. 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	5	6
8	9	10
11	2	5
17	0	9
16	1	8
15	3	6

Løsning

- Vi starter allerførst med at opskrive dataet.
- Derefter opdeles dataet, hvor tingene sættes inde i et dataframe altså i en ramme.
- Derefter anvendes *str()* funktionen til at kunne ses om tingene er læst som faktorer.
- Til sidst anvendes *aov* funktionen til at kunne udføre One-Way ANOVA, hvor vi efterfølgende ser overlappelse gennem LSD-Testen.

```r
data <- c(12,8,11,17,16,15,5,9,2,0,1,3,6,10,5,9,8,6)
treatment <- c(rep("m",6),rep("e",6),rep("d",6))
methods <- data.frame(data,treatment)
str(methods)
methods$treatment <- as.factor(methods$treatment)
res.aov <- aov(data~treatment,data=methods)
summary(res.aov)
print(LSD.test(res.aov,"treatment"))
```

```
> data <- c(12,8,11,17,16,15,5,9,2,0,1,3,6,10,5,9,8,6)
> treatment <- c(rep("m",6),rep("e",6),rep("d",6))
> methods <- data.frame(data,treatment)
> str(methods)
'data.frame':   18 obs. of  2 variables:
 $ data     : num  12 8 11 17 16 15 5 9 2 0 ...
 $ treatment: chr  "m" "m" "m" "m" ...
> methods$treatment <- as.factor(methods$treatment)
> res.aov <- aov(data~treatment,data=methods)
> summary(res.aov)
            Df Sum Sq Mean Sq F value   Pr(>F)
treatment    2  293.4  146.72   16.74 0.000151 ***
Residuals   15  131.5    8.77
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> print(LSD.test(res.aov,"treatment"))
$statistics
   MSerror Df     Mean       CV t.value      LSD
  8.766667 15 7.944444 37.26951 2.13145 3.643608

$parameters
        test p.ajusted    name.t ntr alpha
  Fisher-LSD      none treatment   3  0.05

$means
       data      std r        LCL       UCL Min Max   Q25  Q50   Q75
d  7.333333 1.966384 6  4.7569132  9.909753   5  10  6.00  7.0  8.75
e  3.333333 3.265986 6  0.7569132  5.909753   0   9  1.25  2.5  4.50
m 13.166667 3.430258 6 10.5902466 15.743087   8  17 11.25 13.5 15.75

$comparison
NULL

$groups
       data groups
m 13.166667      a
d  7.333333      b
e  3.333333      c

attr(,"class")
[1] "group"
```

- Vi kan I tilfældet se, at der bliver dannet grupper hvor medikation, kost og træning er alle forskellige fra hinanden og derved findes der en signifikant forskel i metoder for at reducere blodtrykket.

Opgave 6.3

3. In the table below, there are randomly selected scores for eight amateur basketball teams in each of five Danish regions, for a particular weekend. Is there sufficient evidence to support that there is a difference in mean scores by region? If yes, which region/s got the highest scores and which one the lowest?

Region Hovedstaden	Region Sjælland	Region Syddanmark	Region Midtjylland	Region Nordjylland
68	78	89	62	57
75	79	87	74	65
95	65	75	71	78
85	67	65	70	88
84	60	84	72	67
88	79	92	72	77
85	57	84	64	72
75	74	72	75	69

Løsning

- Vi danner vektor for enhver Region, men vi anvender *mean()* funktionen til at finde forskellen i score blandt de 5 regioner.

```
1   Hovedstaden <- c(68,75,95,85,84,88,85,75)
2   mean(Hovedstaden)
3   Sjælland <- c(78,79,65,67,60,79,57,74)
4   mean(Sjælland)
5   Syddanmark <- c(89,87,75,65,84,92,84,72)
6   mean(Syddanmark)
7   Midtjylland <- c(62,74,71,70,72,72,64,75)
8   mean(Midtjylland)
9   Nordjylland <- c(57,65,78,88,67,77,72,69)
10  mean(Nordjylland)
```

```
> Hovedstaden <- c(68,75,95,85,84,88,85,75)
> mean(Hovedstaden)
[1] 81.875
> Sjælland <- c(78,79,65,67,60,79,57,74)
> mean(Sjælland)
[1] 69.875
> Syddanmark <- c(89,87,75,65,84,92,84,72)
> mean(Syddanmark)
[1] 81
> Midtjylland <- c(62,74,71,70,72,72,64,75)
> mean(Midtjylland)
[1] 70
> Nordjylland <- c(57,65,78,88,67,77,72,69)
> mean(Nordjylland)
[1] 71.625
```

- Vi kan se, at der er en signifikant forskel blandt scorene hos de 5 regioner i basketball.
- Vi kan se i tilfældet, at Region Hovedstaden, Syddanmark er dem som har scoret mest.
- Region Sjælland er den Region som har scoret mindst i Basketball, hvorimod Nordjylland og Midtjylland er placeret midterst i rankeringslisten.

Lektion 7

Opgave 7.1 og Opgave 7.2

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

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

II. Median

III. Standard deviation

b) Construct a histogram for the data

c) Construct a boxplot for the data

Løsning for Mean

- Vi starter alleførst med assigne vores data til en vektor, ligesom vi har gjort de andre gange.
- Vi kan nu udregne middelværdien for vektoren

```
1  console <- c(8,9,9,14,8,8,10,7,6,9,7,8,10,14,11,8,14,11)
2  mean(console)
3  median(console)
4  sd(console)
5  hist(x=console)
6  boxplot(console)
```

```
> console <- c(8,9,9,14,8,8,10,7,6,9,7,8,10,14,11,8,14,11)
> mean(console)
[1] 9.5
```

Løsning for Median

- Vi skal bruge mediankommandoen til at kunne udregne medianen.

```
1  console <- c(8,9,9,14,8,8,10,7,6,9,7,8,10,14,11,8,14,11)
2  mean(console)
```

```
> console <- c(8,9,9,14,8,8,10,7,6,9,7,8,10,14,11,8,14,11)
> mean(console)
[1] 9.5
> median(console)
[1] 9
> sd(console)
[1] 2.455486
> hist(x=console)
> boxplot(console)
```

Løsning for Standard Deviation

- VI kan se, at *sd()* funktionen anvendes for at finde Standard Deviationen.

```
1  console <- c(8,9,9,14,8,8,10,7,6,9,7,8,10,14,11,8,14,11)
2  sd(console)
```

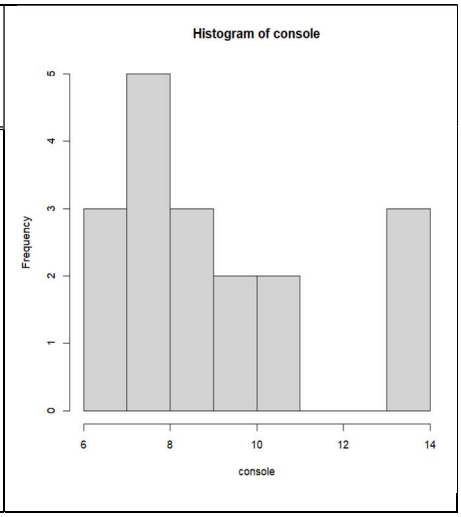
```
> console <- c(8,9,9,14,8,8,10,7,6,9,7,8,10,14,11,8,14,11)
> sd(console)
[1] 2.455486
```

Løsning for Histogram

- Vi skal lave Histogram og her anvendes *hist(x=dataframe)* funktionen.

```
1  console <-
   c(8,9,9,14,8,8,10,7,6,9,7,8,10,14,11,8,14,
   11)
2  hist(x=console)
```

```
> console <-
c(8,9,9,14,8,8,10,7,6,9,7,8,10,14,11,8,14,11)
> hist(x=console)
```

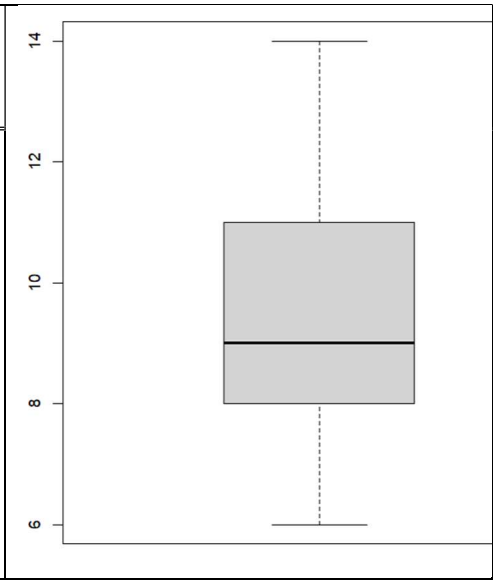


Løsning for Boxplot

- Vi anvender *boxplot(dataframe)* funktionen til at kunne danne Boksplotgrafen.

```
1  console <-
   c(8,9,9,14,8,8,10,7,6,9,7,8,10,14,1
   1,8,14,11)
2  boxplot(console)
```

```
> console <-
c(8,9,9,14,8,8,10,7,6,9,7,8,10,14,11,8,14,
11)
> boxplot(console)
```



Opgave 7.3

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.

Step 2: Calculate the mean, median and standard deviation for the variables "memory", "storage", and "display".

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.

Løsning for Step 1

- Vi kan se, at vi har oprettet 3 vektorer med navnet memory, drive og display.
- Derefter har vi sat dem sammen i en data.frame ramme.

```
1  memory <- c(8,16,16,8,16)
2  drive <- c(500,500,1000,240,500)
3  display <- c(13,15,13,15,17)
4  data.comput <- data.frame(memory,drive,display)
```

```
> drive <- c(500,500,1000,240,500)
> memory <- c(8,16,16,8,16)
> drive <- c(500,500,1000,240,500)
> display <- c(13,15,13,15,17)
> data.comput <- data.frame(memory,drive,display)
>
> View(data.comput)
```

Løsning for Step 2

Part 1: Mean

```
1  mean(data.comput$memory)
2  mean(data.comput$drive)
3  mean(data.comput$display)
```

```
> mean(data.comput$memory)
[1] 12.8
> mean(data.comput$drive)
[1] 548
> mean(data.comput$display)
[1] 14.6
```

Part 2: Median

```
1   median(data.comput$memory)
2   median(data.comput$drive)
3   median(data.comput$display)
```

```
> median(data.comput$memory)
[1] 16
> median(data.comput$drive)
[1] 500
> median(data.comput$display)
[1] 15
```

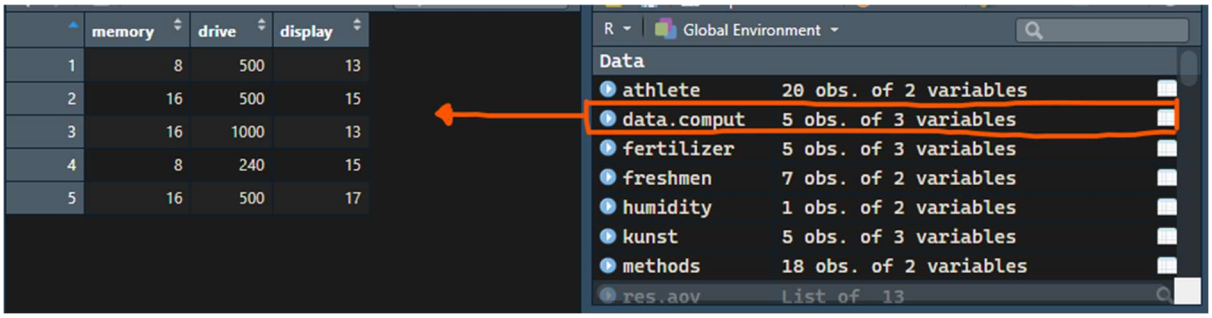
Part 3: Standard Deviation

```
1   sd(data.comput$memory)
2   sd(data.comput$drive)
3   sd(data.comput$display)
```

```
> sd(data.comput$memory)
[1] 4.38178
> sd(data.comput$drive)
[1] 276.6225
> sd(data.comput$display)
[1] 1.67332
```

Løsning for Step 3

- VI kan se på vores Working Directory at vores arbejde er gemt.
- Så det eneste man skal gøre er bare at trykke Save (Ctrl+S) og derved gemme.



Opgave 7.4

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?

b. Make a histogram using the default hist() function. How would you describe the data distribution?

c. Define the number of breaks and choose 5 breaks. The HELP tab can help you here.

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

Løsning for A

- Vi skal først importere filen fra Import-knappen, og derefter kan vi bruge filen i R.
- Vi kan se, at vi skal anvende summary() funktionen til at kunne løse opgaven.

```
1   summary(AirPassengers)
```

```
> summary(AirPassengers)
    Month               Passengers       ...3
 Length:144         Min.   :104.0   Mode:logical
 Class :character   1st Qu.:180.0   NA's:144
 Mode  :character   Median :265.5
                    Mean   :280.3
                    3rd Qu.:360.5
                    Max.   :622.0


      ...4
 Min.   :104.0
 1st Qu.:233.5
 Median :363.0
 Mean   :363.0
 3rd Qu.:492.5
 Max.   :622.0
 NA's   :142
```

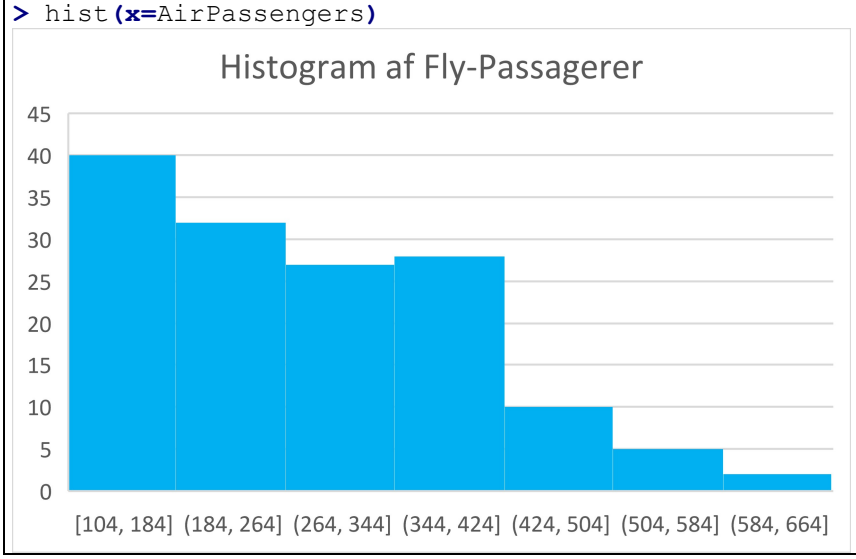
- VI kan I vores tilfælde se, at mindste værdien ved Passengers er 104 og største værdien er 622,0.

Løsning for B

- For at kunne løse denne opgave skal vi anvende *hist(x=AirPassengers)* og derved fås vores Histogram.
- Det skal understreges, at vores Histogram funktion virkede ikke, så vores Resultat er fra Excel.

```
1   hist(x=AirPassengers)
```

```
> hist(x=AirPassengers)
```



Løsning for C

- Vores Histogram funktion virker ikke og derfor kan vi ikke anvende *hist(x=AirPassengers, breaks=5).*

Løsning for D

- Vi har den samme historie, men for at gøre dette skal vi skrive *hist(x=AirPassenger, breaks=80).*
- Man kan bare beholde den sidste break kommando med histogrammet.

Opgave 7.5

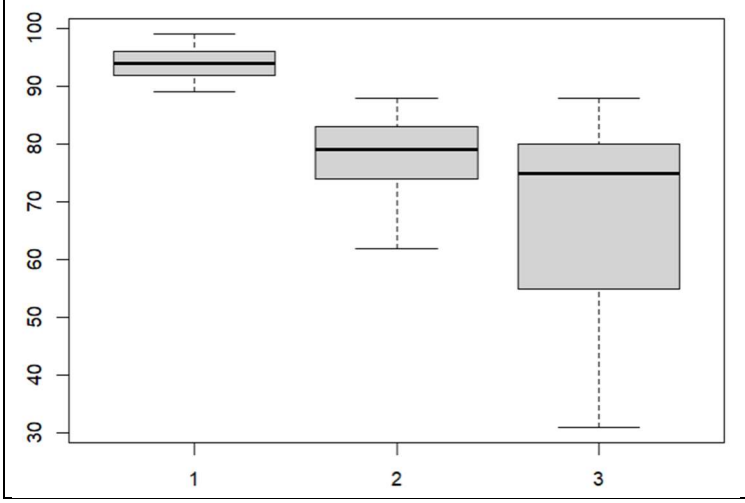
Import the dataset “basketball.csv”, which contains the scores obtained by three professional basketball players in the pre-season games. Make a boxplot for each of the players. When looking at the boxplots, who seems to be the best player? Can we be sure on this result?

Løsning

- Vi kan se, at vi skal anvende *boxplot()* funktionen til at kunne danne de 3 grafer for de 3 spillere i en ramme.

```
1   boxplot(basketball$Michael,basketball$Damon,basketball$Allen)
```

```
> boxplot(basketball$Michael,basketball$Damon,basketball$Allen)
```



- Vi kan ude fra vores boksplot se, at de forskellige grafer er placeret anderledes på baggrund af deres data.
- Vi kan se, at Michael som er Boksplot 1 har en større Maksimumsværdi sammenlignet med de andre grafer. Derfor kan vi sige Michael er den person som har scoret mest i Basketball sammenlignet med Damon og Allen.
- Hvorimod hvis man kigger på den mindste Minimumsværdi, så kan det ses at vi har Allen har scoret mindst i Basketball sammenlignet med Michael og Damon.

Lektion 8

Opgave 8.1

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?

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?

Løsning for A

- For, at kunne vide om dataet følger en normal distribution, ved vi fra præsentationen skal vi anvende *shapiro.test()* kommandoen.

```
1   potato <-
2   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)
3   sort(potato)
4   shapiro.test(potato)
```

```
> potato <-
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)
> sort(potato)
 [1] 14.5 17.2 18.1 18.5 18.5 19.4 20.5 21.5 22.1 23.2 24.1 24.5
> shapiro.test(potato)

    Shapiro-Wilk normality test

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

- Vi kan se ovenpå, at vi med vilje har anvendt *sort()* funktionen selvom der var ingen grund til at anvende det.
- Men grunden til at sorteringens funktion blev anvendt, er fordi vi ønsker at se om tallene er lige sorteret eller med andre orde (lige normalt fordelt).
- Vi kan se, at efterfølgende er shapiro.wilk.testen blevet anvendt og her kan det ses at vores p-værdi viser at vores data er ikke signifikant hvilket betyder at det befinder sig mellem de kritiske intervaller inden for null-hypotesen. Derfor kan vi acceptere at vores dataerne for de 12 kartoffel opvækst er lige fordelte.

Løsning for B

- I denne sammenhæng, har vi anvendt *t.test* funktionen fordi vi kan se at vi beskæftiger os med en One-Sample T.test hvor vi sammenligner populationens middelværdi med en specifik værdi.

```
1  res.ttest <- t.test(potato)
2  res.ttest
3
4
```

```
> res.ttest <- t.test(potato)
> res.ttest

    One Sample t-test

data:  potato
t = 23.133, df = 11, p-value = 1.116e-10
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 18.25544 22.09456
sample estimates:
mean of x
   20.175
```

- Vi kan udefra vores funktion se, at vores sample middelværdi er estimeret til at være 20,175.

Løsning for C

- Spørgsmålet i vores tilfælde hentyder til Confidence Interval.
- Ved at bruge den samme funktion, kan vi finde Confidence Intervallet.
- OBS: Vi har tilføjet *conf.level* som fortæller den justerede confidence procent.

```
1  res.ttest <- t.test(potato,conf.level=0.95)
2  res.ttest
3
4
```

```
> res.ttest <- t.test(potato,conf.level=0.95)
> res.ttest

    One Sample t-test

data:  potato
t = 23.133, df = 11, p-value = 1.116e-10
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 18.25544 22.09456
sample estimates:
mean of x
   20.175
```

- Her kan det ses, at vores confidence interval er endt med at være 18,25 < mu < 22,09 som passer med at vores sample middelværdi kan passe inde i intervallet.

Opgave 8.2 og 8.3

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

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

Løsning for 95

- Vi starter allerførst med, at danne en data.frame hvor på vi indskriver vores data i en ramme.
- Derefter anvender vi *t.test* funktionen, hvor i vi inkluderer dataet fra data.frame og hermed paired=TRUE og conf.level.

```
1 cholostrol <-
2 data.frame(week0=c(215,239,208,190,172,244),week6=c(184,160,201,188,16
3 9,219))
4 res.ttest <- t.test(cholostrol$week0,cholostrol$week6,paired=TRUE,
  conf.level=0.95)
  res.ttest
```

```
> cholostrol <-
data.frame(week0=c(215,239,208,190,172,244),week6=c(184,160,201,188,169,2
19))
> res.ttest <- t.test(cholostrol$week0,cholostrol$week6,paired=TRUE,
conf.level=0.95)
> 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
```

- Vi kan se, at vores p-værdi ligger over de 0,05=95% i signifikant niveau. Dette betyder, at der er ingen statistisk signifikant forskel.
- Vi kan dermed se, at vores sample middelværdi for begge dataer samplet er liggende på 24,5.
- Vores sample middelværdi ligger godt i konfidens intervallet -6,23 < 24,5 < 55,23.

Løsning for 90

- Vi kan se, at vi har brugt den samme funktion men denne gang har vi ændret vores conf.level som er 90% nu.

```
cholostrol <-
data.frame(week0=c(215,239,208,190,172,244),week6=c(184,160,201,188,169,219))
res.ttest <- t.test(cholostrol$week0,cholostrol$week6,paired=TRUE,
conf.level=0.90)
res.ttest
```

```
> cholostrol <-
data.frame(week0=c(215,239,208,190,172,244),week6=c(184,160,201,188,169,219))
> res.ttest <- t.test(cholostrol$week0,cholostrol$week6,paired=TRUE,
conf.level=0.90)
> 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
```

- Vi kan se, at vores middelværdi af sample passer godt inde i konfidens intervallet og derved kan vi sige med 90% sikkerhed at kolesterolniveauet ikke har ændret sig.
- Men vi kan se, at selve resultaterne fra både 95 og 99 er ikke rigtige og derved skal vi forvente at der er noget fejl som vi skal rette. Derfor har vi løst det i form af independent t-test uden brug af pair=TRUE.

Ny Løsning for 95

```
cholostrol <-
data.frame(week0=c(215,239,208,190,172,244),week6=c(184,160,201,188,169,219))
res.ttest <- t.test(cholostrol$week0,cholostrol$week6,paired=TRUE,
conf.level=0.95)
res.ttest
```

```
> res.ttest

    Welch Two Sample t-test

data:  cholostrol$week and cholostrol$uge
t = 1.7122, df = 9.3819, p-value = 0.1196
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.668773 56.668773
sample estimates:
mean of x mean of y
 211.3333  186.8333
```

- Vi kan se med 95% chance at der er sket en udvikling gennem de 6 uger og at der er en statistisk signifikant forskel.

Ny Løsning for 90

```
1  cholostrol <-
2  data.frame(week0=c(215,239,208,190,172,244),week6=c(184,1
3  60,201,188,169,219))
4  res.ttest <-
   t.test(cholostrol$week0,cholostrol$week6,paired=TRUE,
   conf.level=0.90)
   res.ttest
```

```
> cholostrol <-
data.frame(week0=c(215,239,208,190,172,244),week6=c(184,160,
201,188,169,219))
> res.ttest <-
t.test(cholostrol$week0,cholostrol$week6,paired=TRUE,
conf.level=0.90)
> 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
```

```
> res.ttest

    Welch
Two Sample
t-test

data:
cholostrol$
week and
cholostrol$
uge
t = 1.7122,
df =
9.3819, p-
value =
0.1196
alternative
hypothesis:
true
difference
in means is
not equal
to 0
90 percent
confidence
interval:
 -1.608414
50.608414
sample
estimates:
mean of x
mean of y
 211.3333
186.8333
```

- Vi kan se det samme her, hvor der er 90% chance for at der er sket en ændring i kolesterolniveauet gennem de 6 uger. Derfor er der en statistisk signifikant forskel i kolesterolniveauet fra uge 0 til uge 6.

Opgave 8.4

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?

Løsning for A

- Vi starter allerførst med at importere datasættet. Derefter kan vi anvende Independent t.test (uden) brug af paired.

```
1  res.ttest <- t.test(BrandsYoghurt$brandA,BrandsYoghurt$brandB)
2  res.ttest
3
4
```

```
> res.ttest <- t.test(BrandsYoghurt$brandA,BrandsYoghurt$brandB)
> res.ttest

    Welch Two Sample t-test

data:  BrandsYoghurt$brandA and BrandsYoghurt$brandB
t = -3.6857, df = 41.714, p-value = 0.0006516
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -25.422955  -7.430464
sample estimates:
mean of x mean of y
 50.72813  67.15484
```

Løsning for B

- Vi forventer egentligt, at vores data er med 95% sandsynlighed normalt ligeligt fordelt.

Løsning for C

- Vi kan se i Løsning A, at vores mean x og y ligger uden for konfidensintervallerne samt at vores p-værdi befinder sig under 0,05=95%.
- Derfor kan vi sige, at der er en signifikant forskel mellem de to Yoghurt Brands.

Opgave 8.5

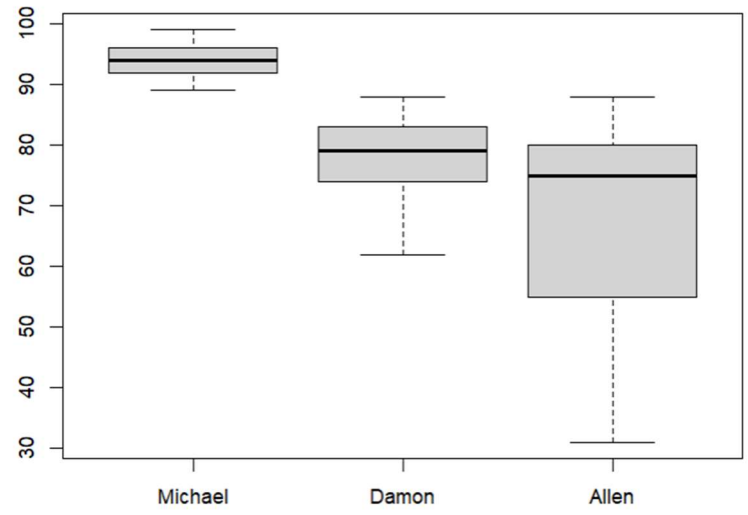
5) We want to compare the scores obtained by three professional basketball players. The data with all the scores obtained by the players in the pre-season games is available in the csv file "basketball.csv".
   a. Construct a boxplot for each of the players, to better visualize the data.
   b. When using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

Løsning for A

- Vi skal først starte med, at importere selve basketball filen.
- Vi skal bruge *boxplot()* funktionen til at kunne danne en ordentligt graf.

```
1   boxplot(basketball)
```

```
> boxplot(basketball)
```



Løsning for B

- Bare ved at kigge på selve boksplottene kan vi dermed se at der er en signifikant forskel ved scorerne blandt de 3 basketball spillere.
- Vi kan dog godt anvende *mean()* funktionen til at beregne middelværdierne for de enkelte basketball spillere.

```
1   mean(basketball$Michael)
    mean(basketball$Damon)
    mean(basketball$Allen)
```

```
> mean(basketball$Michael)
[1] 94.2381
> mean(basketball$Damon)
[1] 78.28571
> mean(basketball$Allen)
[1] 65.7619
```

- Vi kan i vores tilfælde se, at selve middelværdierne for Damon og Allen er ikke særlig langt fra hinanden.
- Men hvis man kigger overimod Michael, så kan det ses at hans Middelværdier mere længere og højere oppe i distancen sammenlignet med de andre.

Løsning for C

- Ude fra de sidste løsninger fra Løsning A og Løsning B, kan vi sige, at der er en forskel blandt de 3 basketball spillere i score og at Michael er den bedste udefra de 3.

Opgave 8.6

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.

Løsning for A

- For at kunne være 99% sikker på at antallet af gange en medarbejder tjekker sin mail om året, skal vi bare anvende *mean()* funktionen og derved placere middelværdien inde i de givende intervaller.

```
1  email <- c(5460,5900,6090,6310,7160,8440,9930)
2  mean(email)
3
4
```

```
> email <- c(5460,5900,6090,6310,7160,8440,9930)
> mean(email)
[1] 7041.429
```

- Vi kan se, at vores Middelværdi kan placeres inde i mellem 4785 og 9298.

Løsning for B

- Vi kan se, at vores professionel tjekker deres email 15 gange på en dag.
- Men vores 7 ansatte tjekker deres email pr. år.
- Derfor kan vi gange vores 15 med de 365 dage, som er 5735. Denne værdi passer inde mellem de to givede intervaller givet i Løsning A.
- 4785 < 5735 < 9298 og her kan vi med 99% sikkerhed sige at antallet af gange de ansatte af virksomheden tjekker deres email er ikke signifikant forskelligt fra den gennemsnitlige professionel.

Løsning for C

- Vi kan dermed sige, at vores svarmulighed C er rigtigt, eftersom beregninger fra A og B er regnet rigtige.

Opgave 8.7

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.

Løsning

- Vi skal finde konfidensintervallet og dette er kun muligt gennem t.test funktionen.
- Her har vi skrevet hvilken konfidensniveau vi ønsker at beskæftige os med.

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

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

- Vi kan se, at vores konfidensinterval ender med at være 6505,022 < 7834,714 < 9164,407.

Opgave 8.8

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

Løsning

- Vi skal anvende One-Way ANOVA metoden herhenne.
- Vi skal først opskrive vores dataer i en hel vektor under navnet måling.
- Derefter inddeler vi dataet gennem rep-funktionen i lige grupper.
- Derefter danner vi en data.frame og opsamler dataet ligeligt.
- Derefter ser på vi om variablerne er korrekt genkendt og derefter anvendes as.factor metoden.

```
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)
```

```
> measure <- c(12,8,11,17,16,15,14,9,2,5,7,4,6,10,5,9,8,6)
Warning messages:
1: In doTryCatch(return(expr), name, parentenv, handler) :
  display list redraw incomplete
2: In doTryCatch(return(expr), name, parentenv, handler) :
  invalid graphics state
3: In doTryCatch(return(expr), name, parentenv, handler) :
  invalid graphics state
> 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" ...
> blodtryk$treatment <- as.factor(blodtryk$treatment)
```

- Efter variabler er genkendt korrekt anvendes *aov()* funktionen til at kunne udføre One-Way ANOVA.
- Det skal understreges, at vi anvender LSD-testen til at kunne se, hvor forskellen egentligt ligger blandt metoderne til at formindske blodtrykket.

```
1   res.aov <- aov(measure~treatment,data=blodtryk)
2   summary(res.aov)
3   print(LSD.test(res.aov,"treatment"))
```

```
> 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
> print(LSD.test(res.aov,"treatment"))
$statistics
   MSerror Df     Mean       CV t.value      LSD
  11.26667 15 9.111111 36.84057 2.13145 4.130591

$parameters
        test p.ajusted    name.t ntr alpha
  Fisher-LSD      none treatment   3  0.05

$means
    measure      std r       LCL       UCL Min Max   Q25  Q50   Q75
d  7.333333 1.966384 6  4.412565 10.254102   5  10  6.00  7.0  8.75
e  6.833333 4.262237 6  3.912565  9.754102   2  14  4.25  6.0  8.50
m 13.166667 3.430258 6 10.245898 16.087435   8  17 11.25 13.5 15.75

$comparison
NULL

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

attr(,"class")
[1] "group"
```

- Vi kan hermed se at medikation er den mest effektive metode imod højt blodtryk, da den har den højeste middelværdi. Vi kan også se under groups, at medikationen adskiller sig fra træning og kost, hvorimod træning og kost har den samme effekt på blodtrykket.
- Dermed kan der konkluderes og siges, at der er en forskel mellem de 3 metoder hvor medikation er den mest effektive.

Lektion 9

Opgave 9.1

1. The director of a university department wishes to see whether there is a difference in the knowledge of students of a course, depending on the teaching method used (1 or 2), and professor that teaches the course (A or B). Four students from each professor and teaching method are randomly selected and they are asked to attend to a single exam. Is there a significant effect of the professor and the teaching method on the student´s results? Which professor and/or method provided the best results?

	Professor A	Professor B
Method 1	20, 25, 22, 29	30, 32, 35, 29
Method 2	15, 18, 22, 21	21, 27, 18, 15

Løsning

- Vi starter allerførst med, at opskrive vores data således at der er dannes en dataframe hvor der skrive hvor mange gange metoden optræder på den vandrette række.
- Derefter opskrives, hvor mange professor optræder ved den enkelte felt.
- Til sidst opskrives selve data, nemlig scoren.

```
1  uni <-
   data.frame(method=c("1","1","1","1","1","1","1","1","2","2","2","2","2"
   ,"2","2","2"),professor=c("a","a","a","a","b","b","b","b","a","a","a","
   a","b","b","b","b"),score=c(20,25,22,29,30,32,35,29,15,18,22,21,21,27,1
   8,15))
```

```
> uni <-
data.frame(method=c("1","1","1","1","1","1","1","1","2","2","2","2","2","
2","2","2"),professor=c("a","a","a","a","b","b","b","b","a","a","a","a","
b","b","b","b"),score=c(20,25,22,29,30,32,35,29,15,18,22,21,21,27,18,15))
```

- Oven på kan vi se at vores metode og professor er signifikante i forhold til læring.
- Men interaktionen med læring er ikke signifikant.
- Nu skal vi til at analysere om variablerne om de er læst rigtige i forhold til deres typer.

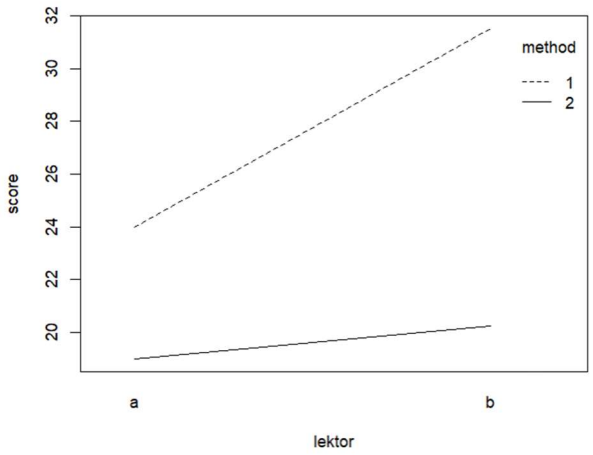
```
1  uni$method <- as.factor(uni$method)
2  uni$professor <- as.factor(uni$professor)
```

```
> uni$method <- as.factor(uni$method)
> uni$professor <- as.factor(uni$professor)
```

- Nu er alle tingene læst rigtigt inde i R og nu skal vi implementere metoden for One-Way ANOVA.

```
1 Two_way_anova_uni <-
  aov(score~method+professor+method*professor,data=uni)
2 summary(Two_way_anova_uni)
3 interaction.plot(uni$professor,uni$method,uni$score,xlab="lektor",ylab
  ="score",trace.label="method")
```

```
> Two_way_anova_uni <-
aov(score~method+professor+method*professor,data=uni)
> summary(Two_way_anova_uni)
                 Df Sum Sq
method            1 264.06
professor         1  76.56
method:professor  1  39.06
Residuals        12 175.75
                 Mean Sq
method            264.06
professor          76.56
method:professor   39.06
Residuals          14.65
                 F value
method            18.030
professor          5.228
method:professor   2.667
Residuals
                  Pr(>F)
method           0.00114 **
professor        0.04120 *
method:professor 0.12838
Residuals
---
Signif. codes:
  0 '***' 0.001 '**' 0.01
  '*' 0.05 '.' 0.1 ' ' 1
>
interaction.plot(uni$professor,uni$method,uni$score,xlab="lektor",ylab="s
core",trace.label="method")
```



- Vi kan hermed se, at den bedste måde at lære på er selvfølgelig gennem metode 1 og i bedste kombination med professor b.

Opgave 9.2

2. A gardening company is testing new ways to improve plant growth. Plants are randomly selected and exposed to a combination of two factors, a "Light" in two different strengths and a plant food supplement with two different mineral supplements. After a number of weeks, the plants are measured for growth and the results (in cm) are the following. Which combination of light and supplement provides the best results?

	Supp. 1	Supp. 2
Light 1	26.7 25.2	28.6 29.3
Light 2	32.3 32.8	26.1 24.2

Løsning

- Vi bruger den samme metode som vi gjorde i sidste opgave.
- Men denne gang ændrer vi bare på navngivningen.

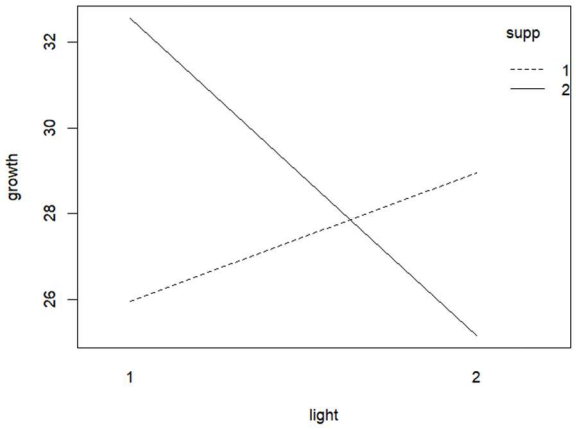
```
plant <-
data.frame(light=c("1","1","1","1","2","2","2","2"),supp=c("1","1","2",
"2","1","1","2","2"),growth=c(26.7,25.2,28.6,29.3,32.3,32.8,26.1,24.2))
plant$light <- as.factor(plant$light)
plant$supp <- as.factor(plant$supp)
```

```
> plant <-
data.frame(light=c("1","1","1","1","2","2","2","2"),supp=c("1","1","2","2
","1","1","2","2"),growth=c(26.7,25.2,28.6,29.3,32.3,32.8,26.1,24.2))
> plant$light <- as.factor(plant$light)
> plant$supp <- as.factor(plant$supp)
```

- Nu, hvor variablerne er læst rigtigt inde i selve R - kan vi derfor begynde på at anvende One Way ANOVA-Testen til at kunne beregne hvad der er den bedste forhold for plantevækst.

```
Two_way_anova_plant <- aov(growth~light+supp+light*supp,data=plant)
summary(Two_way_anova_plant)
interaction.plot(plant$supp,plant$light,plant$growth,xlab="light",ylab
="growth",trace.label="supp")
```

```
> Two_way_anova_plant <- aov(growth~light+supp+light*supp,data=plant)
> summary(Two_way_anova_plant)
             Df Sum Sq Mean Sq
light         1   3.92    3.92
supp          1   9.68    9.68
light:supp    1  54.08   54.08
Residuals     4   3.30    0.82
             F value  Pr(>F)
light          4.752 0.09477 .
supp          11.733 0.02665 *
light:supp    65.552 0.00126 **
Residuals
---
Signif. codes:
  0 '***' 0.001 '**' 0.01
  '*' 0.05 '.' 0.1 ' ' 1
>
interaction.plot(plant$supp,plant$light,plant$growth,xlab="light",ylab="g
rowth",trace.label="supp")
```



Opgave 9.3

3. Two types of paint (A and B), were tested to see how many months they lasted before it began to peel. They were tested in two climatic conditions to study the effects of climate on the paint. Each group contained five test panels. At α=0.01, analyze the data shown. Which paint lasts longer and in which climate?

	Climate 1	Climate 2
Paint A	60, 53, 58, 62, 57	58, 66, 68, 76, 80
Paint B	36, 41, 54, 65, 53	58, 63, 79, 55, 66

Løsning

- Vi skal gentage den samme step som vi har gjort med de forrige opgaver.
- Vi starter allerførst med, at opskrive dataerne i data.frame og derefter skal variablerne indlæses rigtigt.

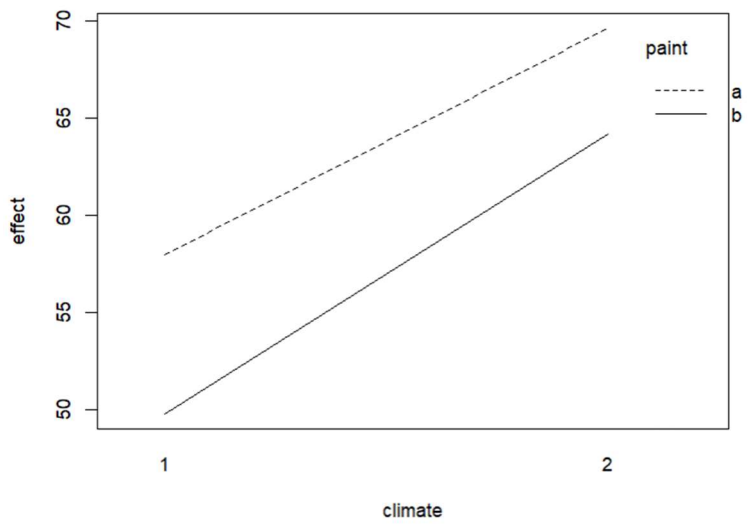
```
1 types <-
  data.frame(paint=c("a","a","a","a","a","a","a","a","a","a","b","b","b",
  "b","b","b","b","b","b","b"),climate=c("1","1","1","1","1","2","2","2",
  "2","2","1","1","1","1","1","2","2","2","2","2"),effect=c(60,53,58,62,5
  7,58,66,68,76,80,36,41,54,65,53,58,63,79,55,66))
2 types$paint <- as.factor(types$paint)
3 types$climate <- as.factor(types$climate)
```

```
> types <-
data.frame(paint=c("a","a","a","a","a","a","a","a","a","a","b","b","b","b
","b","b","b","b","b","b"),climate=c("1","1","1","1","1","2","2","2","2",
"2","1","1","1","1","1","2","2","2","2","2"),effect=c(60,53,58,62,57,58,6
6,68,76,80,36,41,54,65,53,58,63,79,55,66))
> types$paint <- as.factor(types$paint)
> types$climate <- as.factor(types$climate)
```

- Efter variablerne er blevet indlæst rigtigt, kan vi derefter anvende Two-Way ANOVA-testen til at regne ud forholdet.

```
1 Two_way_anova_types <-
2 aov(effect~paint+climate+paint*climate,data=types)
3 summary(Two_way_anova_types)
4 interaction.plot(types$climate,types$paint,types$effect,xlab="climate"
  ,ylab="effect",trace.label="paint")
```

```
> Two_way_anova_types <-
aov(effect~paint+climate+paint*climate,data=types)
> summary(Two_way_anova_types)
              Df Sum Sq Mean Sq
paint          1  231.2   231.2
climate        1  845.0   845.0
paint:climate  1    9.8     9.8
Residuals     16 1218.8    76.2
              F value  Pr(>F)
paint           3.035 0.10067
climate        11.093 0.00424 **
paint:climate   0.129 0.72452
Residuals
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05
  '.' 0.1 ' ' 1
>
interaction.plot(types$climate,types$paint,types$effect,xlab="climate",yl
ab="effect",trace.label="paint")
```



- Vi kan I vores tilfælde se, at de bedste udtørringsforholde er ved brug af paint a og klima 2 som giver de bedste udtørringer for malingen.

Lektion 10

Opgave 10.1

1. The number of hours of study of the students of a course and the final grade of the students (out of 100), is shown in the table. Calculate the correlation coefficient, and determine whether the correlation is significant. Obtain the regression line.

Hours_of_study	Grade
74	87
59	63
45	50
29	39
20.8	21
19.1	28
13.4	14
8.5	15

Løsning

- Vi starter allerførst med at opskrive dataet i en dataframe.
- Derefter laver vi en korrelationstest, hvor vi tegner regressionen på scatterplottet.

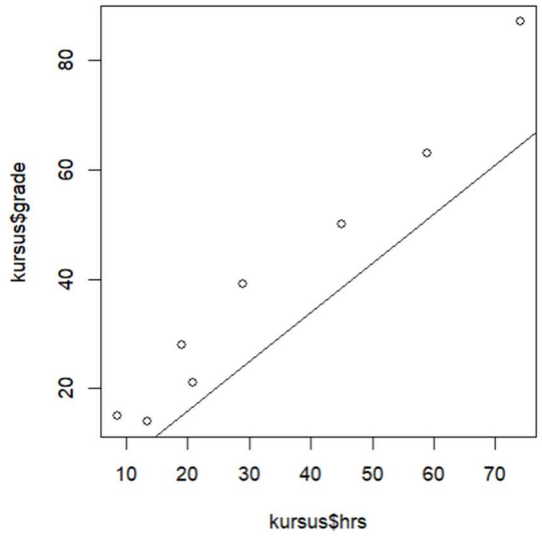
```
kursus <-
data.frame(hrs=c(74,59,45,29,20.8,19.1,13.4,8.5),grade=c(87,63,50,39,21,28,14,15))
cor.test(kursus$hrs,kursus$grade)
plot(kursus$hrs,kursus$grade)
regression_kursus <- lm(hrs~grade,data=kursus)
abline(regression_kursus)
```

```
> kursus <-
data.frame(hrs=c(74,59,45,29,20.8,19.1,13.4,8.5),grade=c(87,63,50,39,21,28,14,15))
Warning messages:
1: In doTryCatch(return(expr), name, parentenv, handler) :
  display list redraw incomplete
2: In doTryCatch(return(expr), name, parentenv, handler) :
  invalid graphics state
3: In doTryCatch(return(expr), name, parentenv, handler) :
  invalid graphics state
> cor.test(kursus$hrs,kursus$grade)

    Pearson's product-moment correlation

data:  kursus$hrs and kursus$grade
t = 15.515, df = 6, p-value = 4.537e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9313803 0.9978695
sample estimates:
      cor
0.9877649

> plot(kursus$hrs,kursus$grade)
> regression_kursus <- lm(hrs~grade,data=kursus)
> abline(regression_kursus)
```



- Vi kan ude fra vores resultater se, at vi har at gøre med en stærk positiv lineære sammenhæng.
- Men vi kan også se, at vores lineære sammenhæng er i den forstand statistisk signifikant fordi vi kan se at vores p-værdi ligger under de 0,05=95%.

Opgave 10.2

2. A researcher carried out an experiment to investigate the relationship between alcohol consumption and blood concentration. The experiment included 5 participants. These were the results:

Participant	Alcohol consumption, number of glasses	Blood alcohol concentration, parts per 1000
1	1	10
2	2	8
3	3	12
4	4	16
5	5	20

   a. Is there a significant relationship between the alcohol consumption, and the concentration of alcohol in blood?
   b. What is the equation of the regression line?
   c. What is the % of the variance in blood alcohol concentration that can be explained by the alcohol consumption?
   d. We want to predict the blood alcohol concentration of a person that has consumed 4.2 glasses. What is the predicted value of blood alcohol concentration and the prediction interval?

Løsning for A

- Bare, at vi løber vores øjne igennem dataet, kan vi se, at der er signifikant forhold mellem dataerne efterrsom den højre ende følger med de høje dataer i concentrationen mod den nedre ende.
- Vi kan se i vores tilfælde, at p-værdien ligger under 0,05=95%.

```
1  alcohol <-
2  data.frame(glasses=c(1,2,3,4,5),concentration=c(10,8,12,16,20))
3  cor.test(alcohol$glasses,alcohol$concentration)
4
```

```
> alcohol <-
data.frame(glasses=c(1,2,3,4,5),concentration=c(10,8,12,16,20))
> cor.test(alcohol$glasses,alcohol$concentration)

    Pearson's product-moment correlation

data:  alcohol$glasses and alcohol$concentration
t = 4.0415, df = 3, p-value = 0.02726
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1950535 0.9947434
sample estimates:
      cor
0.919145
```

Løsning for B

- For at kunne finde frem til regressionsligningen, kan vi anvende *summary()* funktionen efter at have opskrevet ligningen for alkohol konsumptionen.

```
1  plot(alcohol$glasses,alcohol$concentration)
2  regression_alcohol <- lm(concentration~glasses,data=alcohol)
3  abline(regression_alcohol)
4  summary(regression_alcohol)
```

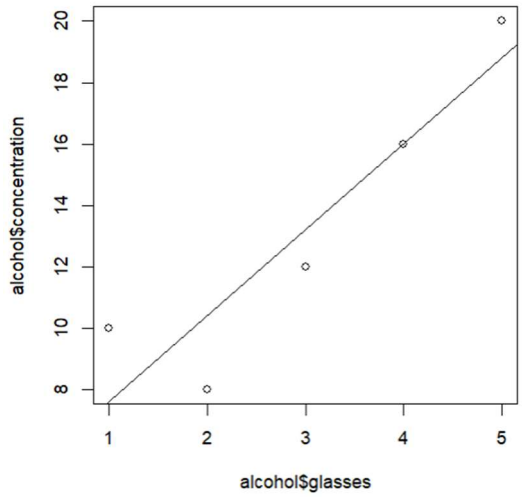
```
> plot(alcohol$glasses,alcohol$concentration)
> regression_alcohol <- lm(concentration~glasses,data=alcohol)
> abline(regression_alcohol)
> summary(regression_alcohol)

Call:
lm(formula = concentration ~ glasses, data = alcohol)

Residuals:
         1          2          3          4          5
 2.400e+00 -2.400e+00 -1.200e+00 -6.661e-16  1.200e+00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.8000     2.2978   2.089   0.1279
glasses       2.8000     0.6928   4.041   0.0273 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.191 on 3 degrees of freedom
Multiple R-squared:  0.8448,    Adjusted R-squared:  0.7931
F-statistic: 16.33 on 1 and 3 DF,  p-value: 0.02726
```



- Vi kan se at vores funktionsforskrift bliver: 4,8+2,8*x.

Løsning for C

- Vi bliver teknisk set spurgt om vores regression egentligt passer, og hvor meget dataerne afviger fra grafen.
- Vi kan se, at ved at gange korrelationskoefficenten med sig selv fås følgende facit: 0,91*0,91=0,82.
- Vores Determinationskoefficent bliver: 0,82*100 = 82% som fortæller at der er small residuals nemlig at x ikke siger 100% perfekt om y.

Løsning for D

- Nu kan vi efterfølgende anvende vores *predict()* funktion til at kunne finde ud af alkohol koncentration i blodet gennem glas.

```
    1    predict(regression_alcohol,data.frame(glasses=4.2),interval="pre
         dict")
```

```
> predict(regression_alcohol,data.frame(glasses=4.2),interval="predict")
    fit      lwr      upr
1 16.56 8.476837 24.64316
```

- Vi kan se under fit-kolonnen at når vi drikker 4,2 genstande alkohol så ender vi med at have en alkoholkoncentration på 16,56.

Opgave 10.3

3. Businesses often use linear regression to understand the relationship between advertising spending and revenue. For example, they might fit a simple linear regression model using advertising spending as the predictor variable and revenue as the response variable. Calculate the value of the correlation coefficient between advertising spending and revenue based on the following data. What is the predicted revenue if a business spends 50 million DKK?

Business	Advertising spending, in million DKK	Revenue, in million DKK
1	43	228
2	48	320
3	56	235
4	61	243
5	67	341
6	70	352

Løsning

- Vi bruger den samme procedure til at kunne løse opgaverne.
- Vi starter allerførst med at finde korrelationskoefficenten.
- Derefter anvender vi regressionsligningen i R og tegner grafen gennem abline.

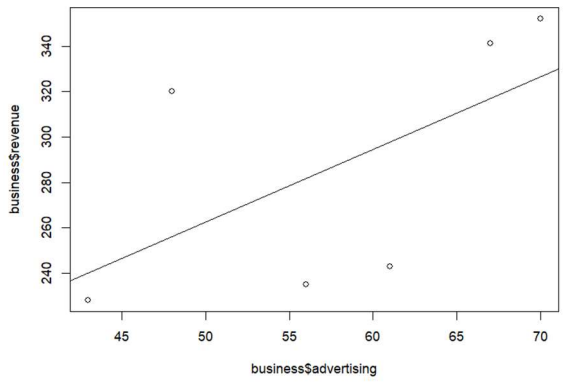
```
1
2
3
4
business <-
data.frame(advertising=c(43,48,56,61,67,70),revenue=c(228,320,235,
243,341,352))
cor.test(business$advertising,business$revenue)
plot(business$advertising,business$revenue)
regression_business <- lm(revenue~advertising,data=business)
abline(regression_business)
```

```
> kursus <-
data.frame(hrs=c(74,59,45,29,20.8,19
.1,13.4,8.5),grade=c(87,63,50,39,21,
28,14,15))
Warning messages:
1: In doTryCatch(return(expr), name,
parentenv, handler) :
  display list redraw incomplete
2: In doTryCatch(return(expr), name,
parentenv, handler) :
  invalid graphics state
3: In doTryCatch(return(expr), name,
parentenv, handler) :
  invalid graphics state
> cor.test(kursus$hrs,kursus$grade)

    Pearson's product-moment
correlation

data:  kursus$hrs and kursus$grade
t = 15.515, df = 6, p-value =
4.537e-06
alternative hypothesis: true
correlation is not equal to 0
95 percent confidence interval:
 0.9313803 0.9978695
sample estimates:
      cor
0.9877649

> plot(kursus$hrs,kursus$grade)
> regression_kursus <-
lm(hrs~grade,data=kursus)
> abline(regression_kursus)
```




```
> business <-
data.frame(advertising=c(43,48,56,
61,67,70),revenue=c(228,320,235,24
3,341,352))
>
cor.test(business$advertising,busi
ness$revenue)

    Pearson's product-moment
correlation

data:  business$advertising and
business$revenue
t = 1.4663, df = 4, p-value =
0.2164
alternative hypothesis: true
correlation is not equal to 0
95 percent confidence interval:
 -0.4235098  0.9479547
sample estimates:
      cor
0.5912752

>
plot(business$advertising,business
$revenue)
> regression_business <-
lm(revenue~advertising,data=busine
ss)
> abline(regression_business)
```

- Vi kan se, at vi har en svag positiv korrelationsværdi mellem revenue og advertising.
- Men nu skal vi beregne hvor meget revenue der bliver dannet hvis vi invensterer 50 millioner i reklamering.

```
1  predict(regression_business,
2  data.frame(advertising=50),interval="predict")
```

```
>
predict(regression_business,data.frame(advertising=50),interval="predict"
)
       fit      lwr      upr
1 262.5708 101.4464 423.6951
```

- Vi kan se, at når vi anvender *predict()* funktionen, så kan vi se at ved at investerer 50 millioner i reklamering ender vi med at danne en revenue på 262,57 ved fit.

Opgave 10.4

4. In a time series, we find a significant relationship between the increase in the number of people who are exercising in Denmark and the increase in the number of people who are committing crimes in US. Comment on whether there is causation.

Løsning

- Vi kan se at fordi der er snak om en signifikant forhold mellem to variabler, så kan vi se at der er snak om en korrelation og ikke en kausation.
- Grunden til at der er ikke snak om kausation, er fordi hvis man havde nævnt noget om ændring i det variabel så ville der blive resulteret en ændring i det andet variabel - som er kausation.

Opgave 10.5

5. Available videogaming statistics have estimated that there are 3.1 billion gamers across the globe. The number of gamers from 2015 to 2023 is shown in the table. What will be the number of gamers across the globe in the year 2040?

Year	Global_players
2015	2.03
2016	2.17
2017	2.33
2018	2.49
2019	2.64
2020	2.81
2021	2.96
2022	3.09
2023	3.22

Løsning

- Vi starter med at finde korrelationsværdien for at se hvor godt sammenhængen er mellem spillernes vækst og årstallet.
- Derefter kan vi danne en graf gennem plot og abline funktionen.

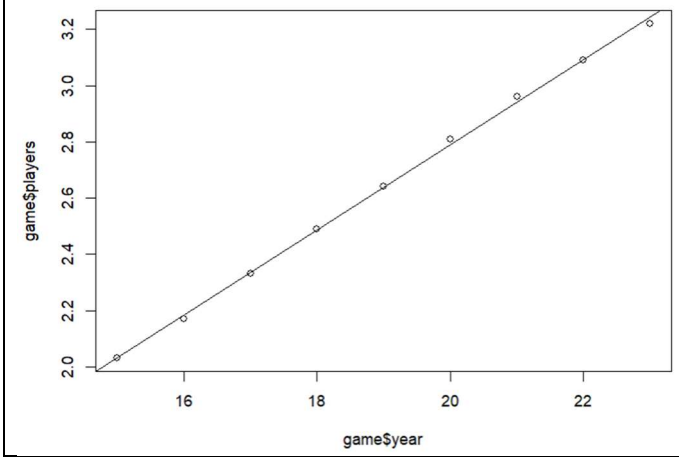
```
1 game <-
2 data.frame(year=c(15,16,17,18,19,20,21,22,23),players=c(2.03,2.17,2.33
  ,2.49,2.64,2.81,2.96,3.09,3.22))
3 cor.test(game$year,game$players)
4 plot(game$year,game$players)
5 regression_game <- lm(players~year,data=game)
6 abline(regression_game)
```

```
> game <-
data.frame(year=c(15,16,17,18,19,20,21,22,23),players=c(2.03,2.17,2.33,2.
49,2.64,2.81,2.96,3.09,3.22))
> cor.test(game$year,game$players)

    Pearson's product-moment
    correlation

data:  game$year and game$players
t = 77.977, df = 7, p-value
= 1.501e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9971538 0.9998839
sample estimates:
      cor
0.9994249

> plot(game$year,game$players)
> regression_game <- lm(players~year,data=game)
> abline(regression_game)
```



- Nu skal vi til anvende vores predict funktion til at udregne hvor mange spillere der bliver i året 2040.

```
1   summary(regression_game)
2   predict(regression_game,data.frame(year=40),interval="predict")
```

```
> summary(regression_game)

Call:
lm(formula = players ~ year, data = game)

Residuals:
      Min        1Q    Median
-0.024444 -0.004444 -0.001111
       3Q       Max
 0.003889  0.020556

Coefficients:
             Estimate Std. Error
(Intercept) -0.243889   0.037295
year         0.151667   0.001945
            t value Pr(>|t|)
(Intercept)  -6.539 0.000322 ***
year         77.977  1.5e-11 ***
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*'
  0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01507 on 7 degrees of freedom
Multiple R-squared:  0.9989,    Adjusted R-squared:  0.9987
F-statistic:  6080 on 1 and 7 DF,  p-value: 1.501e-11

> predict(regression_game,data.frame(year=40),interval="predict")
       fit      lwr      upr
1 5.822778 5.719151 5.926405
```

- Vi kan se, at der i året 2040 ender med at være 5,82 billioner af spillere indenfor videogame området.

Lektion 11

Opgave 11.1

We want to estimate how much Danish citizens have spent in vacations this year.

We randomly select 5000 residents of Frederiksberg and ask them (in a survey), how much money they spent in vacations this year.

Is this a good study design?

Løsning

- Denne "study-design" er ikke god, fordi vi kan se at 5000 beboer i Frederiksberg er ikke en god repræsentativitet til resten af Danmark, da det kun kommer fra København.
- En forslag til denne Study Design ville dog være at inkludere andre folk fra andre byer i Danmark, så det ikke blev begrænset til en bydel.

Opgave 11.2

We want to study whether taking a specific medication has an effect on a certain disease.

We know that the medication may work differently in males and in females.

We design the following study:

- We select 12 subjects (6 males and 6 females) with the disease.
- We form two groups: Group 1 (which includes the 6 males) and Group 2 (which includes the 6 females).
- We give the medication to Group 1, and we do not give the medication to Group 2.
- We evaluate the disease in the 12 subjects and compare the results between the two groups.

Is this a good study design?

Løsning

- Denne "Study Design" er ikke god, eftersom vi har kun givet medikation til mændene i Gruppe 1 og ingen medikation til kvinderne i Gruppe 2.
- På den måde, får vi ingen bred viden om der også sker en påvirkning af medikation hos kvinder. Derfor kunne det være en god idé at sætte 3 mænd og 3 kvinder i Gruppe 1 og 3 mænd og 3 kvinder i Gruppe 2. Derefter kan man afprøve forsøget og fået en lige forståelse for medikationens effekt.

Opgave 11.3

You want to find out which brand of running shoes is more popular among amateur marathon runners in Denmark (Brand 1 or Brand 2).

You talk to 2 friends of yours, who are amateur runners, and ask them which shoes do they use. Both of them use Brand 1.

You conclude that Brand 1 is more popular than Brand 2. In fact, you conclude that 100% of the amateur marathon runners in Denmark, use Brand 1.

Is this a good study design?

Løsning

- Denne "Study Design" er ikke god fordi vedkommende kun spørger sine venner og ikke andre i nærheden. Derfor kan denne konklusion virke som upålidelig.
- En anbefaling ville være at snakke med en større sample gruppe, udenfor venskabsgruppen så det ville være muligt at opnå en mere fair og konklusivt resultat omkring Brand 1 og Brand 2.

Opgave 11.4

There are 10 contestants and 2 judges. Each contestant prepares a cake.

The 10 cakes are randomly placed on the table, so that they judges cannot see which contestant backed each cake.

The two judges start tasting the cakes, one by one, from right to left.

The judges then choose a "Star Baker" for the week, and a contestant is also eliminated.

Is this a good study design?

Løsning

- Denne "Study Design" er GOD fordi vi har en større sample gruppe, samtidig med at undersøgelsen er gjort på en fair og tilfældig måde som gør at vores resultat virker til at være retfærdig og mere konklusivt.
- Man kan godt tage smagsprøve fra de andre kager i betragtning, men dette er noget som man mest ser bort fra i Study-Design opgaver!

Lektion 12

Opgave 12.1

1) You are a researcher interested in social factors that influence heart disease. You survey 15 towns and gather data on the percentage of people in each town who smoke, the percentage of people in each town who bike to work, and the percentage of people in each town who have heart disease.

Town	Heart.disease	Smoking	Biking
A	2.9	69.4	2.8
B	3.1	65.7	13.8
C	4.1	54.4	9.1
D	6.4	65.1	2.2
E	6.7	55.9	25.1
F	6.8	51.8	11.0
G	8.6	53.1	26.3
H	8.6	62.8	16.0
I	9.6	48.8	17.6
J	12.1	35.3	14.4
K	15.9	4.8	29.3
L	14.2	2.0	10.0

a) What is the dependent variable and the independent variables in this study? What would you expect about the relationship between the dependent variable and each of the independent variables?
b) Determine the regression line for the model and the corresponding R^2.
c) Are all independent variables significant to the model? Consider a 95% confidence level.

Løsning for A

- Vi kan se, at fordi vi skal se hvordan sociale faktorer har medindflydelse i hjerte sygdom så kan vi se at Heart_disease er vores afhængige variable da den afhænger af smoking og biking som er uafhængige variabler.

Løsning for B

- Vi starter med at danne 3 vektorer for de 3 værdier undtagen Town.
- Derefter sætter vi dem sammen i en data.frame og laver regression på dem.
- Under *summary()* får vi givet en overblik over hvilke værdier er signifikante og herunder hvad forskriften bliver for regressionslinjen.

```
1  disease <-
2  c(2.9,3.1,4.1,6.4,6.7,6.8,8.6,8.6,9.6,12.1,15.9,14.2)
   smoking <-
   c(69.4,65.7,54.4,65.1,55.9,51.8,53.1,62.8,48.8,35.3,4.8,2.9)
   biking <-
   c(2.8,13.8,9.1,2.2,25.1,11.0,26.3,16.0,17.6,14.4,29.3,10.0)
   sygdom <- data.frame(disease,smoking,biking)
   regression_sygdom <- lm(disease~smoking+biking,data=sygdom)
   summary(regression_sygdom)
```

```
disease <- c(2.9,3.1,4.1,6.4,6.7,6.8,8.6,8.6,9.6,12.1,15.9,14.2)
Error in plot.new() : attempt to plot on null device
> smoking <-
c(69.4,65.7,54.4,65.1,55.9,51.8,53.1,62.8,48.8,35.3,4.8,2.9)
> biking <-
c(2.8,13.8,9.1,2.2,25.1,11.0,26.3,16.0,17.6,14.4,29.3,10.0)
> sygdom <- data.frame(disease,smoking,biking)
> regression_sygdom <- lm(disease~smoking+biking,data=sygdom)
> summary(regression_sygdom)

Call:
lm(formula = disease ~ smoking + biking, data = sygdom)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5788 -0.9587 -0.3520  1.4772  2.6226

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.33607    1.98081   7.237 4.88e-05 ***
smoking     -0.15540    0.02689  -5.778 0.000267 ***
biking       0.08754    0.06907   1.267 0.236828
---
Signif. codes:
0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.852 on 9 degrees of freedom
Multiple R-squared:  0.8384,    Adjusted R-squared:  0.8025
F-statistic: 23.35 on 2 and 9 DF,  p-value: 0.000274
```

- Vi kan se, at vores funktionsforskrift fra Estimate bliver: f(x)=14,33-0,15*x
- Vi kan se, at 14,33 er a-værdien som er skæring med y-aksen og -0,15 er b-værdien som er hældningen.

Løsning for C

- Vi kan se fra Løsning B under summary at selve p-værdien for biking ligger over 0,05=95% konfidensniveau hvilket fortællerat en ændring i biking vil ikke medføre til store ændringer eller betydning for hjertesygdommen.

Opgave 12.2

A health insurance company was hired to provide a better overview of the healthcare expenses associated with hospitalization of patients in Denmark. The company has therefore collected data of 138 patients, who were admitted to different hospitals located in three different Danish regions. The data collected is found in healthcare.xlsx. A description of each variable is found in another tab of the same spreadsheet.

a) The employees from the health insurance company have hypothesized, from the beginning, that the treatment cost (TREATCOST) can be predicted by the number of days the patient has been admitted to the hospital (CAREDAYS). Build a simple linear regression model to investigate this relationship. How much of the variation in TREATCOST can be explained by the variation in CAREDAYS?

b) An employee raised the hypothesis that the treatment cost will also be affected by the region in which the patient was hospitalized. Develop a regression model where you include both CAREDAYS and REGION as independent variables. On average, how much will the treatment cost increase/decrease if the patient is hospitalized one day more? How much will the treatment cost increase/decrease if the patient was hospitalized in the region of Syddanmark in comparison to being hospitalized in the Capital region (Hovedstaden)?

c) Expand the analysis done in item a) and b) and include all the other independent variables in the analysis (remember to check if the variables are correctly recognized in R). Which variables are significant to the model (95% confidence level)?

Løsning for A

- For, at kunne regne variationen af TREATCOST i CAREDAYS, skal vi i tilfældet lave regression på de to variabler og derefter kan vi anvende *summary ()* til at kunne finde den justerede variation.

```
1 regression_Healthcare <- lm(TREATCOST~CAREDAYS,data=Healthcare)
2 summary(regression_Healthcare)
```

```
> regression_Healthcare <- lm(TREATCOST~CAREDAYS,data=Healthcare)
> summary(regression_Healthcare)

Call:
lm(formula = TREATCOST ~ CAREDAYS, data = Healthcare)

Residuals:
   Min     1Q Median     3Q    Max
-88247 -16660  -3700  12419 221222

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     6463       5927   1.091    0.277
CAREDAYS       16572       1169  14.174   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 35460 on 136 degrees of freedom
Multiple R-squared:  0.5963,    Adjusted R-squared:  0.5933
F-statistic: 200.9 on 1 and 136 DF,  p-value: < 2.2e-16
```

- Vi kan se, at vores justerede R^2 / Determination er 0,59 som i procent er 59,33%.

- Dette fortæller at vores regression ikke kan sige alt om vores værdier og det betyder at den ikke er helt troværdig i forhold til at fortolke outputtet af TREATCOST.

Løsning for B

- Vi skal bare gøre det samme som sidst, men denne gang tilføjer vi en mere x-værdi som er REGION.

```
regression_Healthcare <-
lm(TREATCOST~CAREDAYS+REGION,data=Healthcare)
summary(regression_Healthcare)
```

```
> regression_Healthcare <- lm(TREATCOST~CAREDAYS,data=Healthcare)
> summary(regression_Healthcare)
```

```
> regression_Healthcare <- lm(TREATCOST~CAREDAYS+REGION,data=Healthcare)
> summary(regression_Healthcare)

Call:
lm(formula = TREATCOST ~ CAREDAYS + REGION, data = Healthcare)

Residuals:
   Min     1Q Median     3Q    Max
-83258 -15901  -4006  13782 227935

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    27323      11521   2.372   0.0191 *
CAREDAYS       14848       1413  10.508   <2e-16 ***
REGION2       -15238       8773  -1.737   0.0847 .
REGION3       -20064       9543  -2.102   0.0374 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35140 on 134 degrees of freedom
Multiple R-squared:  0.6095,    Adjusted R-squared:  0.6008
F-statistic: 69.72 on 3 and 134 DF,  p-value: < 2.2e-16
```

- Vi kan herhenne se, at vi har ikke fået oplyst hvad Region Hovedstad og Syddanmark er defineret som Region 2 for Hovedstaden og Region 3 for Syddanmark.
- Men for at kunne udregne forskellen i TREATCOST, er det vigtigt at understrege at der indtil videre er behandlet patienter i 138 dage som betyder at vi i x-værdien skal tilføje 139 som dagen ekstra.
- Vi starter med at lave regressionsligningen for Region 2.

$$27323 + 14848 \cdot x - 15238 \cdot z - 20064 \cdot v$$

- Når vi indsætter antal dage af behandling indenfor x og 2 inden ved Region 2 som er z.

$$27323 + 14848 \cdot 139 - 15238 \cdot 2 - 20064 \cdot 0 = 2060719$$

- Nu laver vi den samme metode for Region 3.

$$27323 + 14848 \cdot 139 - 15238 \cdot 0 - 20064 \cdot 3 = 2031003$$

- Vi kan hermed se, at for Region Hovedstad har en større TREATCOST end Region Syddanmark.
- Vi kan se, at TREATCOST falder med 20,063 kroner (set under Region 3 ved Estimate) når det omhandler om betaling i Syddanmark, hvorimod ved indlæggelse af ekstra dag på hospitalet vil det resultere 14,848 kroner ekstra set ved CAREDAYS under Estimate.

Løsning for C

- Vi skal gøre det samme som tidligere, hvor vi skal bare tilføje de resterende uafhængige variabler som har betydning for TREATCOST i alfabetisk røkkefølge.

```
regression_Healthcare <-
lm(TREATCOST~MEDICINE+LAB+XRAY+INHALATOR+STATUS+CAREDAYS+INTENSIVEDAYS
+AGE+SEX+INSURANCE+REGION,data=Healthcare)
summary(regression_Healthcare)
```

```
> regression_Healthcare <-
lm(TREATCOST~MEDICINE+LAB+XRAY+INHALATOR+STATUS+CAREDAYS+INTENSIVEDAYS+AG
E+SEX+INSURANCE+REGION,data=Healthcare)
> summary(regression_Healthcare)

Call:
lm(formula = TREATCOST ~ MEDICINE + LAB + XRAY + INHALATOR +
    STATUS + CAREDAYS + INTENSIVEDAYS + AGE + SEX + INSURANCE +
    REGION, data = Healthcare)

Residuals:
   Min     1Q Median     3Q    Max
-32251 -10466     77   6282  85095

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    9259.7321 17059.8659   0.543 0.588249
MEDICINE          2.3546     0.7592   3.101 0.002382 **
LAB               1.5984     0.4292   3.724 0.000295 ***
XRAY              1.5215     0.3246   4.688 7.11e-06 ***
INHALATOR         1.7714     0.2984   5.936 2.69e-08 ***
STATUS1       -3027.9156  7058.7724  -0.429 0.668692
CAREDAYS       5019.6769  1106.0858   4.538 1.31e-05 ***
INTENSIVEDAYS  2933.1817  2635.3474   1.113 0.267838
AGE             -49.3816   180.3819  -0.274 0.784720
SEX1           2218.6264  3292.7592   0.674 0.501689
INSURANCE     -7079.7291  5987.7084  -1.182 0.239300
REGION2        5778.8742  4953.5438   1.167 0.245587
REGION3        3486.2600  5444.1373   0.640 0.523104
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 18820 on 125 degrees of freedom
Multiple R-squared:  0.8954,    Adjusted R-squared:  0.8854
F-statistic:  89.2 on 12 and 125 DF,  p-value: < 2.2e-16
```

- Vi kan se, at alle de stjerner som står ved siden af variablerne, er de værdier som er signifikant og befinder sig under konfidensniveauet 0,05=95%.
- Dette betyder, at MEDICINE, LAB, XRAY, INHALATOR, CAREDAYS er signifikante.

Opgave 12.3

A real estate agent wants to better understand what are the factors that influence the price of houses sold in the region of greater Copenhagen. For that, she hires a group of statisticians and provides data on house age in years (X1), distance to public transportation in meters (X2), number of convenience stores (X3), house condition, where 1 = poor, 2 = medium, 3 =high (X4), and house price of unit area in 1,000 DKK (Y) for 413 houses. The data is available in the file "real_estate.txt"

a) Determine the appropriate regression model.

b) With the regression line equation you found in item a), predict the house price of unit area for a house that is 10 years old, is 1 km from public transport, has 1 convenience store close by and is in a high level condition.

c) Repeat item b) but using the predict() function in R. What is the estimated predicted house value?

Løsning

- Vi kan se, at vi skal anvende den samme metode som vi har gjort med de andre opgaver.
- Vi skal anvende regressionsmetode, hvori prisen (Y) er vores afhængige variabel.

```
1 | regression_real_estate <- lm(Y~X1+X2+X3+X4,data=real_estate)
2 | summary(regression_real_estate)
```

```
> regression_real_estate <- lm(Y~X1+X2+X3+X4,data=real_estate)
> summary(regression_real_estate)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4, data = real_estate)

Residuals:
    Min      1Q  Median      3Q     Max
-27.540  -3.304  -0.494   2.774  64.808

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.8976989  1.6576646  10.797  < 2e-16 ***
X1          -0.1204197  0.0300323  -4.010 7.23e-05 ***
X2          -0.0024376  0.0003643  -6.691 7.33e-11 ***
X3           0.5619746  0.1467081   3.831 0.000148 ***
X4          12.0496250  0.6320173  19.065  < 2e-16 ***
---
Signif. codes:
0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 6.74 on 409 degrees of freedom
Multiple R-squared:  0.757, Adjusted R-squared:  0.7546
F-statistic: 318.6 on 4 and 409 DF,  p-value: < 2.2e-16
```

- Vi kan se, at vi skal opskrive den rigtige regression ligning.

$$f(x) = 17{,}89 - 0{,}12 \cdot x - 0{,}0024 \cdot z + 0{,}56 \cdot v + 12{,}049 \cdot w$$

Løsning for B

- Vi kan se, at vi skal finde værdierne for prisen hvis der er nogle særlige parametre indsat.

$$f(x) = 17{,}89 - 0{,}12 \cdot 10 - 0{,}0024 \cdot 1000 + 0{,}56 \cdot 1 + 12{,}049 \cdot 3 = 50{,}997$$

Løsning for C

- Desværre virkede ikke vores tilfælde, men følgende er opskrivningen.

```
1  regression_real_estate <- lm(Y~X1+X2+X3+X4,data=real_estate)
2  summary(regression_real_estate)
   predict(regression_real_estate,data.frame(X=10,X2=1000,X3=1,X4=3))
```

```
> regression_real_estate <- lm(Y~X1+X2+X3+X4,data=real_estate)
> summary(regression_real_estate)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4, data = real_estate)

Residuals:
    Min      1Q  Median      3Q     Max
-27.540  -3.304  -0.494   2.774  64.808

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.8976989  1.6576646  10.797  < 2e-16 ***
X1          -0.1204197  0.0300323  -4.010 7.23e-05 ***
X2          -0.0024376  0.0003643  -6.691 7.33e-11 ***
X3           0.5619746  0.1467081   3.831 0.000148 ***
X4          12.0496250  0.6320173  19.065  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 6.74 on 409 degrees of freedom
Multiple R-squared:  0.757, Adjusted R-squared:  0.7546
F-statistic: 318.6 on 4 and 409 DF,  p-value: < 2.2e-16

> predict(regression_real_estate,data.frame(X=10,X2=1000,X3=1,X4=3))
Error in eval(predvars, data, env) : object 'X1' not found
```

- Vi kan se, at vores funktion virkede ikke i predict men facit burde være 51,586 Danske Kroner.

Opgave 12.4

A car dealer has collected data on all used cars he sold within last year. He makes a dataset called car.txt with the information collected for all 301 cars.

The dataset contains the following variables:

· Selling_Price: Price in which the car is being sold (in Euros)

· Original_Price: Price when the car was first bought (in Euros)

· Kms_Driven: Number of kilometers the car is driven

· Fuel_Type: Fuel type of car (Petrol/Diesel)

· Transmission: Gear transmission of the car (Automatic/Manual)

Based on the given data, use the appropriate regression model to predict the selling price of a car that was originally bought by 7500 Euros, it was driven for 8000 kilometers, uses petrol and has an automatic gear.

Løsning

- Her skal vi anvende regressionsligningen igen ved R og derved skal det understreges at Selling Price bliver vores afhængige variabel før tilde.

```
regression_car <- lm(Selling_Price~Original+Kms_Driven,data=car)
summary(regression_car)
predict(regression_car,data.frame(Original_Price=7500,Kms_Driven=8000)
)
```

```
> regression_car <- lm(Selling_Price~Original+Kms_Driven,data=car)
Error in eval(predvars, data, env) : object 'Original' not found
> summary(regression_car)
Error in summary(regression_car) : object 'regression_car' not found
> predict(regression_car,data.frame(Original_Price=7500,Kms_Driven=8000))
Error in predict(regression_car, data.frame(Original_Price = 7500,
Kms_Driven = 8000)) :
  object 'regression_car' not found
```

- Som I kan se virkede vores kommando-funktion desværre ikke.

Opgave 12.5

Using the same context from exercise 4, which statement is correct in relation to the model you developed to predict cars' selling price?

a) Increasing the original price of the car in 1 euro will result in a decrease of 469 euros in the car's selling price (considering all other variables are fixed).

b) Increasing the original price of the car in 1 euro will result in an increase of 1.59 euros in the car's selling price (considering all other variables are fixed).

c) The selling price of a car that is run by diesel is estimated to be 1619 Euros lower than a car run by petrol (considering all other variables are fixed).

d) The selling price of a car that has manual gear is estimated to be 1589 Euros lower than a car that has automatic gear (considering all other variables are fixed).

Løsning

- Vi skal bruge den samme regressionsmetode i R som tidligere, men denne gang skal vi afmærke hvilket svar er rigtigt i forhold til facitet i 5'eren.
- Fordi vores R ikke virker kan vi sige, at vores pris vil ligge på omkring 1589 kroner lavere end salgsprisen.

```
regression_car <- lm(Selling_Price~Original+Kms_Driven,data=car)
summary(regression_car)
predict(regression_car,data.frame(Original_Price=7500,Kms_Driven=8000)
)
```

```
> regression_car <- lm(Selling_Price~Original+Kms_Driven,data=car)
Error in eval(predvars, data, env) : object 'Original' not found
> summary(regression_car)
Error in summary(regression_car) : object 'regression_car' not found
> predict(regression_car,data.frame(Original_Price=7500,Kms_Driven=8000))
Error in predict(regression_car, data.frame(Original_Price = 7500,
Kms_Driven = 8000)) :
  object 'regression_car' not found
```