LESSON 7. STATISTICS IN R – PART 1

Class content:

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

Statistisk Dataanalyse

LESSON 7. STATISTICS IN R – PART 1

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

- **I tilfældet, kan det ses at vi har allerede installeret R-Programmet i Computeren, så det er Opgaven som "hoppes" over.**
- **I tilfældet, kan det ses at vi skal finde Middelværdien, Medianen og Standardafvigelse.**
- **Vi starter allerførst med at finde Middelværdien ("Mean").**

Summen af Vektordata	$\mathbf{8+9+9+14+8+8+10+7+6+9+7+8+10+14+11+8+14+11=171}$
Middelværdien af Vektordata	$\frac{\mathbf{171}}{\mathbf{18}}=\frac{\mathbf{19}}{\mathbf{2}}=\mathbf{9,5}$

- **Nu skal vi finde Medianen af vores Vektordatasæt.**
- Vi kan i vores tilfælde se, at fordi vi har 18 talværdier, kan vi derfor trække 2 fra 18 som giver os 16. Derefter kan vi dividere 16 med 2 og dette giver os 8.
- Når vi har fundet vores 8-tal, kan vi derfor tælle 8 talværdier fra begyndelsen indtil vi lander de 2 tal i midten som er 6 og 9.
- Middelværdien af 6 og 9 er svarende til vores Median som er følgende: $\frac{6+9}{2}=\frac{15}{2}=7{,}5$

- **Nu findes Standardafvigelsen af Vektordatasættet.**

$\mathbf{9,5-8=1,5}$	$\mathbf{9,5-9=0,5}$	$\mathbf{9,5-9=0,5}$	$\mathbf{14-9,5=4,5}$	$\mathbf{9,5-8=1,5}$	$\mathbf{9,5-8=1,5}$	$\mathbf{10-9,5=0,5}$	$\mathbf{9,5-7=2,5}$	$\mathbf{9,5-6=3,5}$	$\mathbf{9,5-9=0,5}$

$9,5 - 7 = 2,5$	$9,5 - 8 = 1,5$	$10 - 9,5 = 0,5$	$14 - 9,5 = 4,5$	$11 - 9,5 = 1,5$	$9,5 - 8 = 1,5$	$14 - 9,5 = 4,5$	$11 - 9,5 = 1,5$

- Nu skal vi tage summen af kvadraterne for subtraktionerne.

$1,5^2 = 2,25$	$0,5^2 = 0,25$	$0,5^2 = 0,25$	$4,5^2 = 20,25$	$1,5^2 = 2,25$	$1,5^2 = 2,25$	$0,5^2 = 0,25$	$2,5^2 = 6,25$	$3,5^2 = 12,25$	$0,5^2 = 0,25$
$2,5^2 = 6,25$	$1,5^2 = 2,25$	$0,5^2 = 0,25$	$4,5^2 = 20,25$	$1,5^2 = 2,25$	$1,5^2 = 2,25$	$4,5^2 = 20,25$	$1,5^2 = 2,25$		
Summen	$2,25 + 0,25 + 0,25 + 20,25 + 2,25 + 2,25 + 0,25 + 6,25 + 12,25 + 0,25 + 6,25 + 2,25 + 0,25 + 20,25 + 2,25 + 2,25 + 20,25 = 100,25$								

- **Nu findes den endelige Standardafvigelse.**

$$sd = \sqrt[2]{\frac{100,25}{(18-1)}} \approx 2,428386$$

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.

- **Vi skal lave en dataframe, hvor vi samler alle data for Memory i en, alle data for Storage i en enhed og Display i en enhed. Følgende datasæt viser konceptet.**

Datacomput	**Memory <-**	**C(8,16,16,8,16)**
	Storage_Drive <-	**C(500,500,1000,240,500)**

	Display <-	**C(13,15,13,15,17)**

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.

- **Vi starter allerførst med at finde Middelværdien for de 3 Systemer.**

Summen af Memory	$\mathbf{8 + 16 + 16 + 8 + 16 = 64}$
Middelværdi af Memory	$\frac{\mathbf{64}}{\mathbf{5}} = \mathbf{12,8}$
Summen af Storage	$\mathbf{500 + 500 + 1000 + 240 + 500 = 2740}$
Middelværdi af Storage	$\frac{\mathbf{2740}}{\mathbf{5}} = \mathbf{548}$
Summen af Display	$\mathbf{13 + 15 + 13 + 15 + 17 = 73}$
Middelværdi af Display	$\frac{\mathbf{73}}{\mathbf{5}} = \mathbf{14,6}$

Statistisk Dataanalyse

- **Nu skal vi finde Medianen for de adskillige Systemer.**

Memory	$8, 16, 16, 8, 16$	16
Storage	$500,500,1000,240,500$	1000
Display	$13,15,13,15,17$	13

- **Nu skal vi finde Standardafvigelsen for de 3 Systemer.**

Memory	$12,8 - 8 \approx 4,8$	$16 - 12,8 \approx 3,2$	$16 - 12,8 \approx 3,2$	$12,8 - 8 \approx 4,8$	$16 - 12,8 \approx 3,2$
Storage	$548 - 500 = 48$	$548 - 500 = 48$	$1000 - 548 = 452$	$548 - 240 = 308$	$548 - 500 = 48$
Display	$14,6 - 13 \approx 1,6$	$15 - 14,6 \approx 0,4$	$14,6 - 13 \approx 1,6$	$15 - 14,6 \approx 0,4$	$17 - 14,6 \approx 2,4$

- Nu skal vi bare lige finde summen og dividere de relevante værdier for at finde den endelige Standardafvigelse

Memory	$4,8^2 = 23,04$	$3,2^2 \approx 10,24$	$3,2^2 \approx 10,24$	$4,8^2 = 23,04$	$3,2^2 \approx 10,24$
Storage	$48^2 = 2304$	$48^2 = 2304$	$452^2 = 204304$	$308^2 = 94864$	$48^2 = 2304$
Display	$1,6^2 \approx 2,56$	$0,4^2 \approx 0,16$	$1,6^2 \approx 2,56$	$0,4^2 \approx 0,16$	$2,4^2 = 5,76$

- Nu findes den endelige Standardafvigelse.

Memory's Standardafvigelse	$23{,}04 + 10{,}24 + 10{,}24 + 23{,}04 + 10{,}24 = 76{,}8$	$\sqrt[2]{\frac{76{,}8}{4}} \approx 4{,}38178$
Storage's Standardafvigelse	$2304 + 2304 + 204304 + 94864 + 2304 = 306080$	$\sqrt[2]{\frac{306080}{4}} = 2 \cdot \sqrt{19130} \approx 276{,}6225$
Display's Standardafvigelse	$2{,}56 + 0{,}16 + 2{,}56 + 0{,}16 + 5{,}76 = 11{,}2$	$\sqrt[2]{\frac{11{,}2}{4}} \approx 1{,}67332$

- **Den sidste Opgave med, at gemme Data kan "hoppes".**

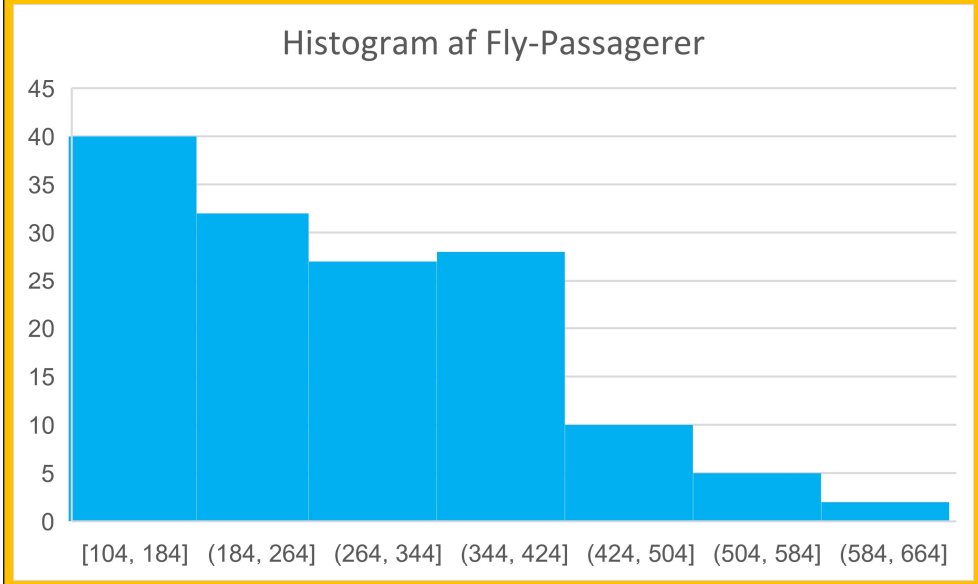
4) Import the dataset "Air_passengers.xlsx", which contains data on the number of passengers that have flew in a specific airplane per month. Now do the following:
   a. Summarize the data. What is the minimum and maximum number of passengers who flew in this airplane?
   b. Make a histogram using the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

A. Spørgsmålet som vi bliver stillet her, er om vi kan finde Størsteværdien og Mindsteværdien?

- Det er meget simpelt i vores tilfælde, fordi vi kan bruge =min og =maks funktioner inde på Excel for at finde frem til Størsteværdien og Mindsteværdien af Passagerer. Husk, at vi bruger Excel på Dansk!
- I vores tilfælde er Mindsteværdien af Hyppigheden = 104, og Størsteværdien = 622.

B. Vi kan se, at vi skal lave et Histogram, hvor vi skal beskrive skævhed gennem Distributionen.

- På Excel, kan man under Indsæt-Menuen finde Histogram ikonet hvilket gør at vi kan danne følgende Diagram.



- Ude fra diagrammet, kan vi se at der er snak om positiv højre skævhed og dette er fordi vores middelværdi er større end medianen på datasættet.

C. Vi bliver her spurgt om, hvordan det er muligt at ændre på søjler i diagrammet og dette kan gøres på Excel.

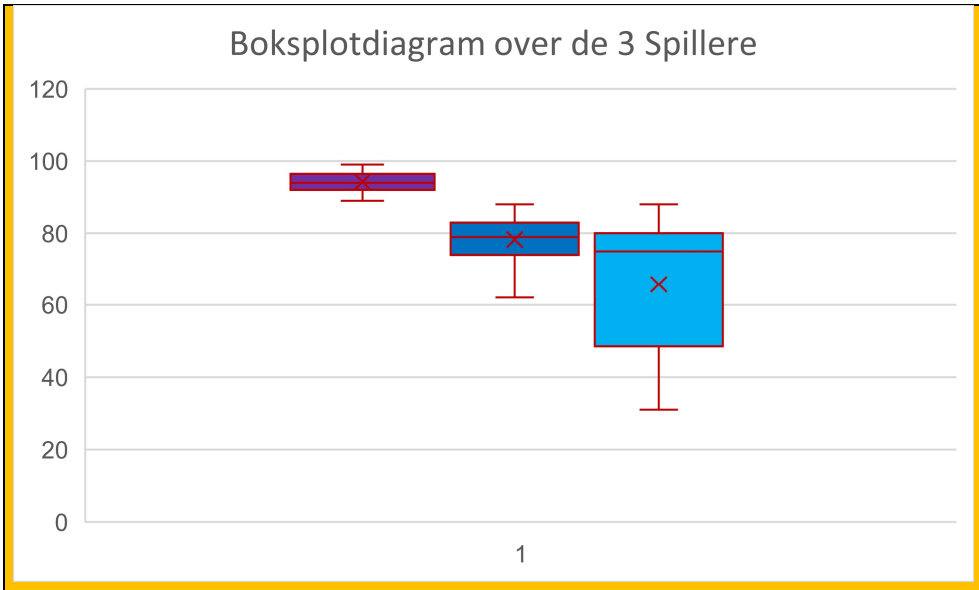
D. Det samme handler om opdeling til 20 søjler.

- Vi springer opgaven over, da jeg ikke lige helt kender på Excel men som kan findes på Youtube. Men fordi R-Program er med i pensummet, skippes denne over som Håndopgave.

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?

- For at kunne løse denne opgave på den ”håndfulde-måde”, kan vi gå inde under Indsæt-Menuen og derved finde Boksplot ikonet under Histogram. Derefter kan Boksplotgrafen dannes som følgende:

Statistisk Dataanalyse



- Vi kan se, at den Blå-Boksplot har den øverste Størsteværdi sammenlignet med de andre Boksplotgrafer. Men fordi denne Boksplotgraf tilhører Michael, kan vi derfor sige at Michael er ham som er den bedste spiller blandt de 3 nævnte.