LESSON 2. DESCRIPTIVE STATISTICS – PART 2

1.3. Measures of:

- Percentile
- Decile and Quartile
- Outlier

1.4. Data representation

- Frequency distributions and histograms
- Box-plots
- Graphs: Time series, Pie graphs, Scatter plots

1.5. Shapes of frequency distributions

- Skewness
- Kurtosis
- Shapes of distributions

Statistisk Dataanalyse

Chapter 1 – Part 2: Assignments

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.

Classroom A	Classroom B
156	185
175	175
189	169
165	182
160	179
154	163
158	191
170	182
171	180
169	174
180	161
175	180
172	176
169	174
162	182
	173

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)

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

Statistisk Dataanalyse

- 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

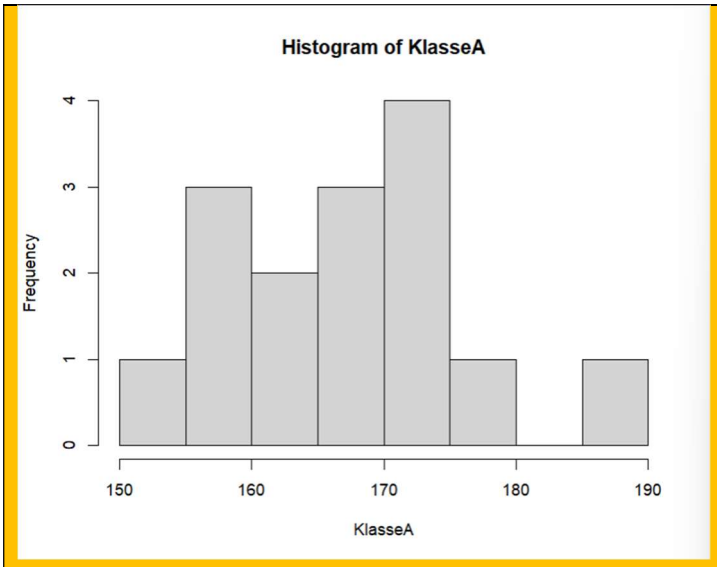
B. Nu skal vi lave en grouped frekvens tabel og dette er ved at danne intervaller fra A til B, og fra C til D osv. Her inde i de forskellige intervaller skal vi indsætte værdier på, hvor mange gange den samme værdi fra datasættet optræder flere gange.

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

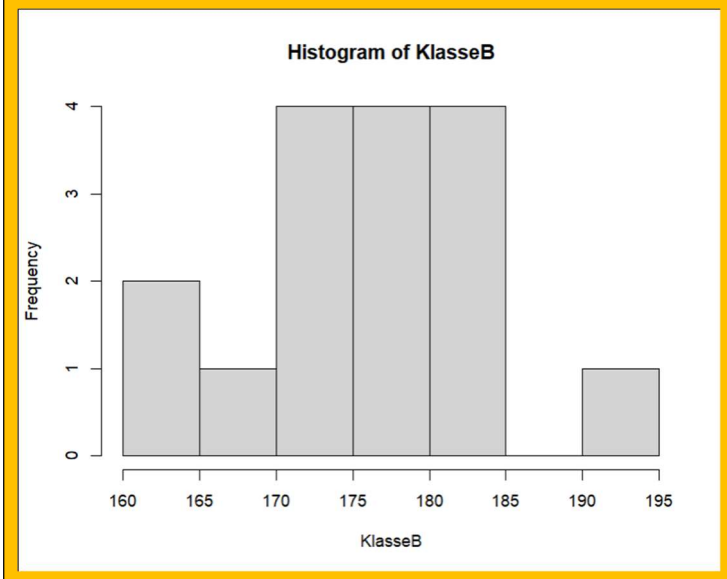
Grupperet Frekvens Tabel for Klasse B				
(160-165(	(165-170(	(170-175(	(175-180(	(180-191)
2	1	3	3	7

C. Vi skal lave Histogram og i tilfældet, har vi konstrueret det i R.

- Her har vi lavet det for Klasse A.



- Her har vi lavet det for Klasse B.



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

Percentile	Score
95th	140
80th	120
65th	101
50th	94
35th	91
20th	87
5th	80

a. Symmetrical bell-shaped
b. Skewed left (negatively skewed)
c. **Skewed right (positively skewed)**
d. Impossible to tell from the above

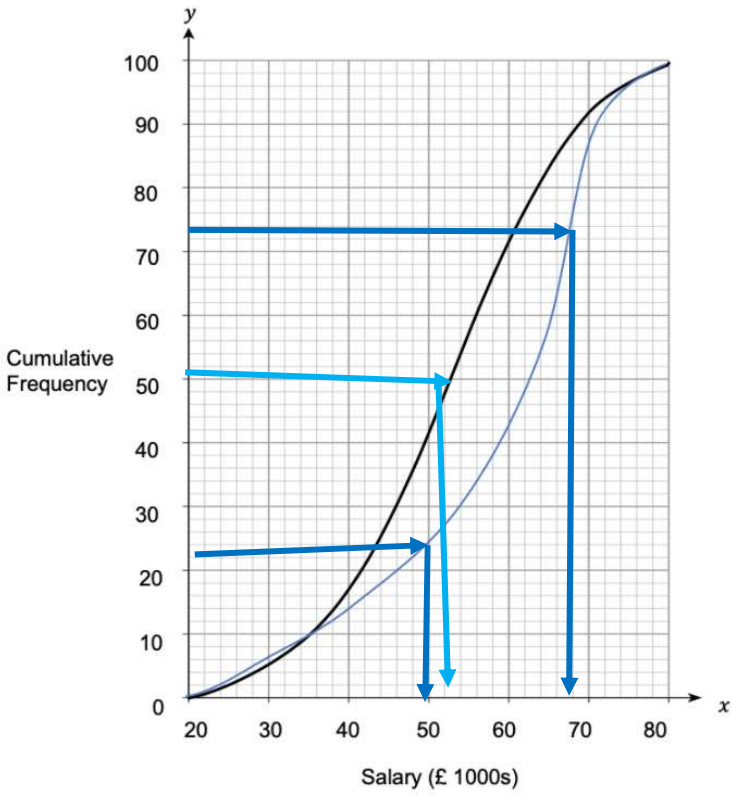
- Vi kan I vores tilfælde se, at vores værdier starter fra øverst og går ned ad. Dette fortæller, at der er snak om en positiv højre skævhed.
- En anden måde, at finde det ud af på er ved at finde middelværdien og medianen af scoren på. I tilfældet er det fundet på følgende måde:

$$Median = \frac{101 + 94}{2} = \frac{195}{2} = 97{,}5$$

$$Middelværdi = \frac{140 + 120 + 101 + 94 + 91 + 87 + 80}{7} = \frac{713}{7} \approx 101{,}8571$$

- Vi kan se, at vores middelværdi er større end medianen (mean>median), derfor er der snak om højre positiv skævhed.

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

i. I tilfældet, kan det ses at vi har skal finde kvartilbredden af den finske Bank og derved afgøre om det er 60000.

$$IQR = 69 - 50 = 19$$

- Falsk, vi kan se at vores kvartilbredde er ikke 60 men 19 og derved kan vi sige at påstand er falsk.

ii. Vi fra 50 ved andenaksen og derved rammer den sorte linje (Welsh) og derved går ned mod førsteaksen og ser om vi rammer 62.

- Falsk, vi kan dermed se, at vi rammer 51 og dette siger at vores påstand er forkert.

iii. For at kunne udregne vores Outlier, skal følgende formel anvendes som findes nedenfor.

$$Q3 + 1{,}5 \cdot IQR$$
$$Q1 - 1{,}5 \cdot IQR$$

- Vi indsætter vores værdier ude fra grafen, inde i vores formel fra foroven.

$$60 + 1{,}5 \cdot 15 = 82{,}5$$
$$45 - 1{,}5 \cdot 15 = 22{,}5$$

- Sandt, I vores tilfælde indsætter vi 78 mellem intervallet (82,5<78<22,5) og derved kan vi se at værdierne passer med hinanden.

iv. Vi skal her finde Variationsbredden for begge Banker.

- Eftersom vi kan se, at Største og Mindsteværdien lander i det samme værdiområde i Grafen, kan vi skrive følgende.

$$80 - 20 = 60$$

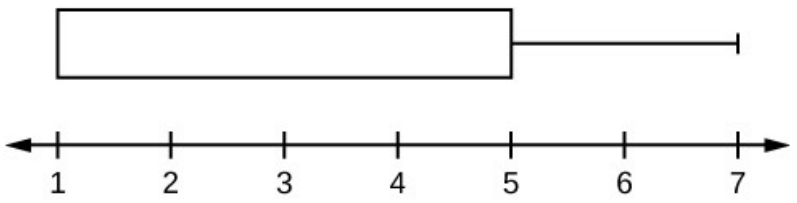
- Sandt, vi kan se at Variationsbredden for begge Banker er den samme.

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.

- Vi kan se, at vores middelværdi (mean) er 8 og vores median er 7.
- Fordi vi har (mean>median), så er der snak om positiv højre skævhed.

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.

- Vi kan se, at vores boksplot viser os at den mindsteværdi er 1 og vores største værdi er 7. Fordi vores 1. kvartil er det samme som vores 0. kvartil, og at vores 3. kvartil er 5.
- Så kan vi derfor sige, at svaret C er den rigtige svar passende til boksplot-billedet.

6. A teacher gives a 20-point test to 10 students. Find the 25th percentile rank of a score of 12.
18, 15, 12, 6, 8, 2, 3, 5, 20, 10

- Vi starter allerførst med at sortere selve datasættet.

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

- Nu anvender vi følgende formel:

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

- Vi indsætter 25'ende procentil inde på k's plads og derefter indsætter vi antal data i rækkefølgen på n's plads.

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

- Vi kan se, at 2,75 er positionen mellem 3 og 5, og derved kan det siges at vi finder midrange for de værdier.

$$\frac{3+5}{2} = 4$$

- Derfor kan det siges, at den 25th procentil er 4, i datasættet mellem 3 og 5.

7. Find Q_1, Q_2, and Q_3 for the data set.
15, 13, 6, 5, 12, 50, 22, 18

- Vi starter allerførst med at sortere selve datasættet.

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

- Vi starter allerførst med at finde medianen Q2.
- Fordi vi har to værdier i midten 13,15, skal disse to værdier findes midrange af.

$$Q2 = \frac{13+15}{2} = 14$$

- Nu skal vi finde 1. Kvartil Q1, og dette er muligt gennem 6 og 12.
- Vi bruger den samme udregningsmetode.

$$Q1 = \frac{6+12}{2} = 9$$

- Vi ønsker, at finde 3. Kvartil Q3 og dette er muligt gennem 18 og 22.
- Vi bruger den samme udregningsmetode.

$$Q3 = \frac{18 + 22}{2} = 20$$

8. 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.

- Vi starter med, at sortere datasættet.

22,23,63,74,78,78,89,91,91,94

- Nu skal vi udregne medianen og her kan det ses vi har lige antal værdier.
- Derfor tager vi de midterste værdier som 78 og 78 og finder summen og dividerer med 2.

$$Median = \frac{78 + 78}{2} = 78$$

- Her kan det ses, at vores Median er større end Middelværdien (mean).
- Dette vil sige, at vores (median>mean) og derfor er det venstre negativ skævhed.

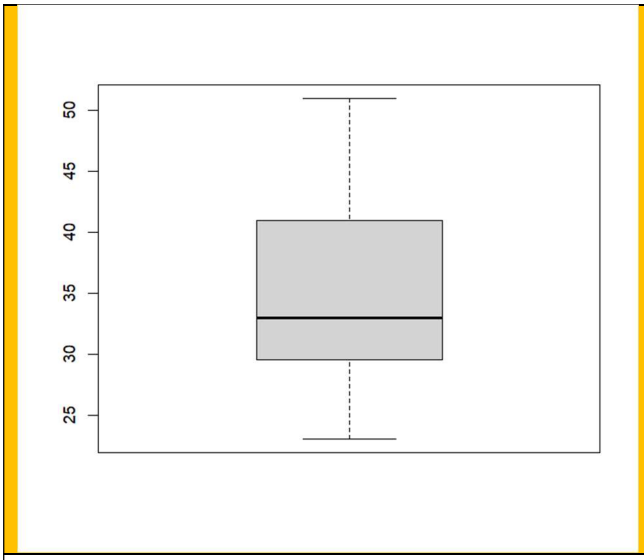
9. A percentile score of 40 indicates that a person:
   a. answered 40% of the questions correctly on the test.
   b. knows 40% of the material covered by the examination.
   c. has earned a score equal to or better than 40 persons in his class.
   d. **has earned a score equal to or better than 40% of the persons in his class.**

- Vi burger det selviske princip fra Præsentationen, hvor vi siger hvor mange procent bedre man har scoret end de andre.
- Derfor kan det ses, at 40 indikerer at den vedkommende person har scoret lig med eller bedre end de 40% fra vedkommendes klasse.

10. Construct two boxplots (“Type 1” and “Type 2”) for the data.
    33, 38, 43, 30, 29, 40, 51, 27, 42, 23, 31

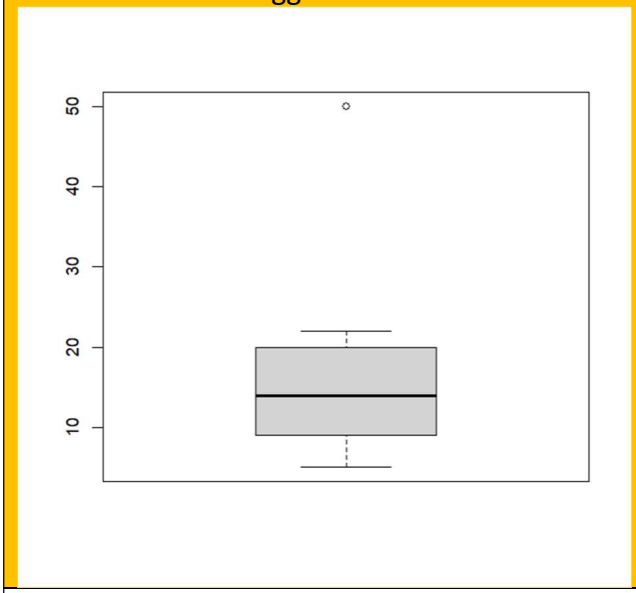
- Til denne opgave kan man anvende R eller Geogebra.
- Jeg har lavet boksplot for denne del af opgaven og den næste del i R.

Statistisk Dataanalyse



11. In this data set:
    15, 13, 6, 5, 12, 50, 22, 18
    Is there any outlier?

- Vi burger R programmet til at konstruere en Boksplotgraf.
- Det skal understreges, at vi kan se en enkelt udeliggende værdi fra resten af boksplotgrafen.
- Denne udeliggende værdi er kendt som en Outlier, som kan bekræftes 😊



12. 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

A. Ja, det kan man godt men man skal definere de forskellige faktoriske værdier om til numeriske værdier.
B. Ja, det kan vi godt og jeg har defineret de adskillige bogstaver som følgende.

A=1, B=2, O=3, AB=4

Statistisk Dataanalyse

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

- Vi har lavet følgende Boksplot udefra datasættet i R.