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

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

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

- Når vi snakker om ungrouped, så snakker vi om alle værdier/variabler/højder som ikke er talt med i score!

```
#Opgave 1
classroomA <- c(156,175,189,165,160,154,158,170,171,169,180,175,172,169,162)
classroomB <- c(185,175,169,182,179,163,191,182,180,174,161,180,176,174,182,173)
```

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

Når vi snakker om en grupperet frekvenstabel, så snakker vi om at inddrage de relevante værdier som ikke står 0 ved i scoren!

```
#Opgave 1b
//Vi skal i tilfældet her sortere de brugte højder ud!
sortedheights <- c(154,156,158,160,161,162,163,165,169,169,169,170,171,172,173,174,174,175,175,175,176,179,180,180,180,182,182,182,185,189,191
#Opgave 1a
//Vi skal sætte tingene på en ugrupperet rækkefølge, men det betyder ikke sorteret men omvendt en sorteret rækkefølge hvor indsætter også de talværdier som ikke e
totalclass <- data.frame(height=c('154','155','156','157','158','159','160','161','162','163','164','165','166','167','168','169','170','171','172','173','174','1
totalclass$height <- as.factor(totalclass$height)
```

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

Her skal vi lave et histogram i R, og til det skal vi på forhånd have defineret vores sorteret / relevante værdier i R. Derefter kan vi bruge histograms-kommandoen og indsætte numerisk værdi i.

```
#Opgave 1c
newdata <- data.frame(intervals=c("[154,159]","[160,169]","[170-179]","[180,191]"),vardier=c(3,8,10,9)
hist(x=newdata$vardier)
```

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

Vi skal starte med at indskrive værdierne inde i R og derved kan vi danne et histogram.

Først starter vi med at skrive den nedenstående tabel inde i R.

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

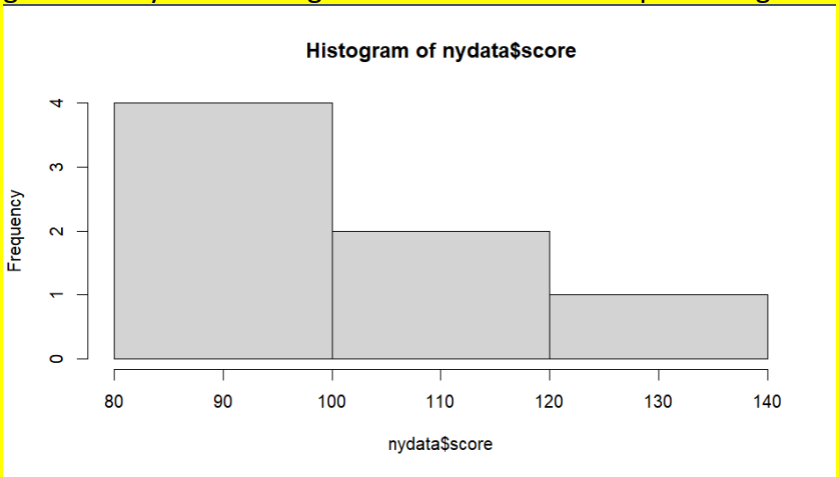
a. Symmetrical bell-shaped
b. Skewed left (negatively skewed)
c. Skewed right (positively skewed)
d. Impossible to tell from the above

Vi skal starte med at indskrive værdierne inde i R og derved kan vi danne et histogram.

Først starter vi med at skrive den nedenstående tabel inde i R.

```
#Opgave 2
nydata <- data.frame(procentil=c('95','80','65','50','35','20','5'),score=c(140,120,101,94,91,87,80))
hist(x=nydata$score)
```

Nu skal vi starte med at danne et histogram. Her kan det ses at den samme kommando som tidligere skal anvendes. Men skal indsættes et nyt defineret arraylist med numeriske række! Eftersom vi har den numeriske score i talværdier, har vi derfor forbundet vores score med vores generelle nydata i histogrammet. Derved har vi opnået følgende resultat på skærmen!



Vi bliver bedt om at finde ud af om grafen er symmetrisk klokke-formet, negativ skævt eller positiv skævt - eller om det er umuligt at fortælle foroven.

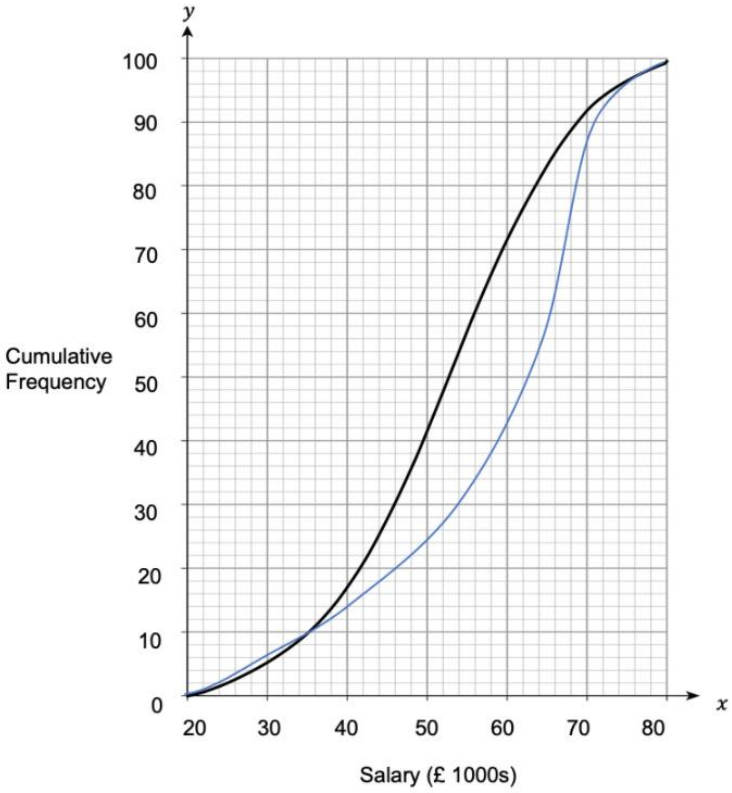
Vi kan ude fra histogrammet se, at der er en positiv skævhed fordi grafen vender mod højre, eller RIGHT/TRUE/POSITIVE!
Men til eksamen kan det være en fordel at beregne middelværdien og medianen ud, da det også har et forhold med hinanden i skævhed. I tilfældet, kan vi for neden se at vores middelværdi er 101,85 og median er 94 - som siger at middelværdien er højere end medianen. Dette betyder, at vi har graf som også har med positiv skævhed at gøre!

```
mean(nydata$score)
median(nydata$score)
```

```
> mean(nydata$score)
[1] 101.8571
>
> median(nydata$score)
[1] 94
```

Derfor kan vi sige, at vores svar til sidst bliver c - markeret med grønt!

1. 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 - Kvartilbredden af data for Banken af Finland er 60000

Fordi vi ønsker at finde kvartilbredden, skal vi bare starte med at trække 75%-25% og derved finder vi kvartilbredden mellem Q3 og Q1. Ude fra den blå graf, kan vi se at vi har 69 henne ved x-aksen og 51 ved x-aksen.

```
#Opgave 2i
69-51
```

II - Medianen for Welsh Baken er 62000 Pund

Fordi vi leder efter medianen, så kan det ses det er 50% ved Q2. Derfor kigger vi langs x-aksen ved den sorte graf sammen lignet med 50 ved y-aksen! - (Samme princip/metode som tidligere).

```
#Opgave 2ii
62
```

Vi har bare skrevet svaret i R!

III - 78000 er en Outlier for Welsh Banken

I tilfældet her, kan det ses at vores Outlier kan skrives på en simpel måde i R.

```
#Opgave 2iii
45-1.5*(18)
61+1.5*(18)
```

```
> #Opgave 2iii
> 45-1.5*(18)
[1] 18
> 61+1.5*(18)
[1] 88
>
```

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

Vi kan se, at vores middelværdi er større end vores median, derfor er det positiv skævhed!

a. negatively skewed
b. not skewed
c. positively skewed
d. symmetrical
e. bimodal.

3. **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.

Kigget på billedet ovenpå, så kan det ses at vi har en boksplot. Den mindsteværdi er 1, men det fortæller også at nummer 1 er med i en liste som var med til at danne boksplottet. Men det er ikke præcist svar, fordi det er meget subjektivt!
Den anden svar b, er overhovedet ikke rigtigt men vi kan se at c'eren er den mest rigtige svar og det er fordi boksplotten er lavet ude fra de nævnte datasæt 1,1,5,5,7. Det betyder, at mindsteværdi er 1 og den størsteværdi er 7. Vi kan se, at den midterste værdi er 5 som er medianen!
Det kan ses, at d og e er forkert fordi halvdelen af data falder ikke mellem de to tal og kvartilbredden er ikke 6.

4. **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**

Vi skal finde den percentile rankering af scoren 12. Vi skal finde en "selvisk-metode", hvor værdierne under 12 skal bruges!
Vi starter med, at sortere datasættet, hvor sort funktionen bruges og derved tages middelværdien fra mindsteværdien og frem til værdien som er under 12, nemlig 10!
I tilfældet, kan det ses at vi har beregnet middelværdien for værdierne under 12!

```
#Opgave 4
percentilerank <- c(18,15,12,68,2,3,5,20,10)
sort(percentilerank)
mean(2+3+5+10)
```

```
> #Opgave 4
> percentilerank <- c(18,15,12,68,2,3,5,20,10)
> sort(percentilerank)
[1]  2  3  5 10 12 15 18 20 68
> mean(2+3+5+10)
[1] 20
>
```

1. **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

Vi har her i tilfældet isoleret vores liste og derved kan vi se at vi har 10 scores i alt.
$2, 3, 5, 6, 8, 10, 12, 15, 18, 20$
Vi har i opgaven fået at vide, at vi skal finde **en værdi som, er korresponderende til 25'ende procenttal**. Vi siger med andre orde, at vi skal gennem udregning finde en værdi fra 25'ende procenttal.

$$0{,}25 = \frac{x}{10}$$
$$2{,}5 \cdot 10 = \frac{x}{\cancel{10}} \cdot \cancel{10}$$
$$2{,}5 \cdot 10 = 2{,}5$$

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

Vi starter med at opskrive værdi inde i R, og derefter sortere vi listen.

```
#Opgave 5
Q <- c(15,13,6,5,12,50,22,18)
Q_sorteret <- c( 5, 6, 12, 13, 15, 18, 22, 50)
median(Q_sorteret)
```

```
#1.Kvartil
6+12
18/2
#3.Kvartil
18+22
40/2
```

```
> #1.Kvartil
> 6+12
[1] 18
> 18/2
[1] 9
> #3.Kvartil
> 18+22
[1] 40
> 40/2
[1] 20
>
```

Oven på kan det ses, at vi har fået vores kvartilsæt og derved kan det ses at vores 1. Kvartil er svarende til 18. Vores 3. Kvartil er svarende til 20.

3. **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.

I tilfældet her kan det ses at vores svar er negativ fordi vores typetal(mode) er større end middelværdien som siger at der findes negativ skævhed på en potentiel dannelse af grafen.

4. **A percentile score of 40 indicates that a person:**
   . answered 40% of the questions correctly on the test.
   b. knows 40% of the material covered by the examination.

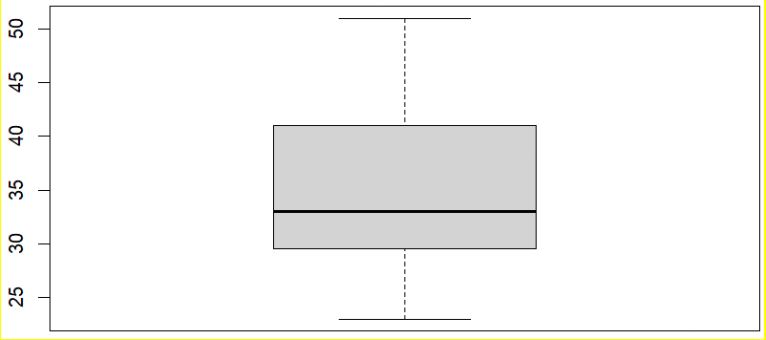
c. has earned a score equal to or better than 40 persons in his class.
d. has earned a score equal to or better than 40% of the persons in his class.

I tilfældet kan det ses her at vi skal implementere den selviske koncept. Det betyder, at vis nakker om den percentile score af 40 indikerer en person som er bedre end 40% af personerne fra hans klasse!

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

I tilfældet her, kan det ses at vi har startet med at definere datalisten og derefter har vi anvendt boxplot-funktionen til at konstruere en boksplot.

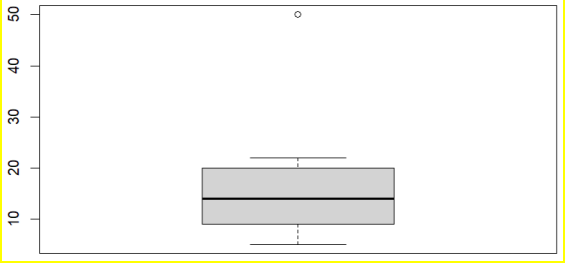
```
#Opgave 9
boksplot <- c(33,38,43,30,29,40,51,27,42,23,31)
boxplot(boksplot)
```



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

Vi har startet med at lave en dataliste og derefter har vi konstrueret en boksplot.

```
#Opgave 10
neudata <- c(15,13,6,5,12,50,22,18)
boxplot(neudata)
```



7. **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. Du kan godt konstruere et histogram, men du skal definere værdierne og derved kan du konstruere et histogram. Men udfordringen er, at du skal have numeriske

værdier for at kunne konstruere et histogram i din x-akse for at kunne konstruere et histogram. Du har dog vel ikke sådan i tilfældet her….

b. I have considered my reply, and that is of course the experience which I have had through the whole semester!