

LESSON 1. DESCRIPTIVE STATISTICS – PART 1

1.1. Statistics

- What is statistics?
- Definitions: Variable, data, population and sample
- Descriptive and Inferential statistics

1.2. Variables and Types of Data

Qualitative, quantitative, categorical, ordinal, discrete, continuous

1.3. Measures of:

- Central Tendency (Location):
 - Mean
 - Median
 - Mode
 - Midrange
- Variation (Dispersion)
 - Range
 - Variance
 - Standard Deviation
 - Coefficient of Variation

Chapter 1: 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. For each of the classroom, calculate the following:
 - a. Median
 - b. Mean
 - c. Mode
 - d. Midrange

Display the results on a table.

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. Vi starter med at definere klasserne gennem "assign <- " metoden. Assign kan opfattes som defineringsmetoden 😊
 - Derefter skrives median, hvor de definerede værdier som classroomA og classroomB indsættes hvert for sig.

```
#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)
#Opgave 1a
median(classroomA)
median(classroomB)
```

- Herefter får vi følgende resultater fra kommandolinjen 😊

```
> #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)
> #Opgave 1a
> median(classroomA)
[1] 169
> median(classroomB)
[1] 177.5
```

- b. Mean i tilfældet her vil betyde, at vi finder middelværdien og derfor bruger vi den samme metode som median, men hvor vi bare i stedet skriver mean().

```
#Opgave 1b
mean(classroomA)
mean(classroomB)
```

```
> #Opgave 1b
> mean(classroomA)
[1] 168.3333
> mean(classroomB)
[1] 176.625
```

- c. Vi bruger den samme metode, men forskellen her er at vi ikke har nogen direkte funktion til at finde typetal. Derfor den bedste metode ville være, at sorte og derved tælle selv. I tilfældet bruges sort() funktionen på den samme måde.

```
#Opgave 1c
sort(classroomA)
"Fordel rækken på begge sider!"
sort(classroomB)
"Fordel rækken på begge sider!"
```

- Følgende resultat kommer ved kommandolinjen.

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

- Det kan ses, at vi har at gøre med bimodal typeptal i classroom A fordi vi har to typetalsværdier som er 169 og 175.
- Hvorimod classroom B er også bimodal, fordi vi har typetals værdier som 180 og 182.

- d. Vi kan se, at vi skal finde midrange. Her kan vi bare tage størsteværdien og mindsteværdien, fra den sorterede liste i Opgave 1.c og derved ligge sammen og finde middelværdien af de to.

$$classroomA = \frac{189 + 154}{2} = \frac{343}{2} = 171,5$$

$$classroomB = \frac{161 + 191}{2} = 176$$

2. Find the mean of the following data:
20, 26, 40, 36, 23, 42, 35, 24, 30

- I tilfældet, kan det ses at vi har gjort det på følgende måde:

```
#Opgave 2
newdata <- c(20, 26, 40, 36, 23, 42, 35, 24, 30)
mean(newdata)
```

- Følgende resultater kommer fra kommandolinjen.

```
> #Opgave 2
> newdata <- c(20, 26, 40, 36, 23, 42, 35, 24, 30)
> mean(newdata)
[1] 30.66667
```

3. Find the median of the following measurements:
713, 300, 618, 595, 311, 401, and 292

- I tilfældet, kan det ses at vi har gjort det på følgende måde:

```
#Opgave 3
newdata1 <- c(713, 300, 618, 595, 311, 401, 292)
median(newdata1)
```

- Følgende resultater kommer fra kommandolinjen.

```
> #Opgave 3
> newdata1 <- c(713, 300, 618, 595, 311, 401, 292)
> median(newdata1)
[1] 401
```

4. Find the median of the following measurements:
684, 764, 656, 702, 856, 1133, 1132, 1303

- I tilfældet, kan det ses at vi har gjort det på følgende måde:

```
#Opgave 4
newdata2 <- c(684, 764, 656, 702, 856, 1133, 1132, 1303)
median(newdata2)
```

- Følgende resultater kommer fra kommandolinjen.

```
> newdata2 <- c(684, 764, 656, 702, 856, 1133, 1132, 1303)
> median(newdata2)
[1] 810
```

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

- I tilfældet har vi løst det på følgende måde gennem sorteringsmetoden.

```
#Opgave 5
newdata3 <- c(8, 9, 9, 14, 8, 8, 10, 7, 6, 9, 7, 8, 10, 14, 11, 8, 14, 11)
sort(newdata3)
```

- Følgende resultater kommer fra kommandolinjen, hvor man kan se typetallet.

```
> #Opgave 5
> newdata3 <- c(8, 9, 9, 14, 8, 8, 10, 7, 6, 9, 7, 8, 10, 14, 11, 8, 14, 11)
> sort(newdata3)
[1] 6 7 7 8 8 8 8 8 9 9 9 10 10 11 11 14 14 14
```

- Vi kan se, at vi at gøre med en multimodal typetal, som er: 7, 8, 9, 10, 11, 14

6. Find the mode of the following measurements:
110, 731, 1031, 84, 20, 118, 1162, 1977, 103, 752

- Det er gjort på følgende måde.

```
#Opgave 6
newdata4 <- c(110, 731, 1031, 84, 20, 118, 1162, 1977, 103, 752)
sort(newdata4)
```

- Vi får følgende resultater.

```
> #Opgave 6
> newdata4 <- c(110, 731, 1031, 84, 20, 118, 1162, 1977, 103, 752)
> sort(newdata4)
[1] 20 84 103 110 118 731 752 1031 1162 1977
```

7. Find the midrange of these data:
2, 3, 6, 8, 4, 1

- Vi har skrevet op på følgende måde i R.

```
#Opgave 7
newdata5 <- c(2, 3, 6, 8, 4, 1)
sort(newdata5)
//Fordel rækkerne
(8+1)/2
```

- Nu kan vi se følgende resultat:

```

> #Opgave 7
> newdata5 <- c(2,3,6,8,4,1)
> sort(newdata5)
[1] 1 2 3 4 6 8
> //Fordel rækkerne
Error: unexpected '/' in "/"
> (8+1)/2
[1] 4.5

```

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

- Vi har løst det på følgende måde:

```

#Opgave 8
"Occupation ("blue collar", "white collar", "unemployed") //Nominal and Categorical"
"Highest attain education ("low", "medium", "high") //Ordinal"
"Montly salary //Discrete or continous"
"Civil status ("single", "married", "window") //Discrete"
"//Number of children //Continous"

```

1. 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.

- A.) Det er ikke rigtigt.
- B.) Det er rigtigt.
- C.) Det er ikke rigtigt, men omvendt.
- D.) Det er ikke, rigtigt men det er den håndfulde gruppe som er taget af de 100.

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

Brand A	Brand B
10	35
60	45
50	30
30	35
40	40
20	25

a. Følgende måde er brugt i R.

```
#Opgave 10
BrandA <- c(10,60,50,30,40,20)
BrandB <- c(35,45,30,35,40,25)
#Opgave 10 a
range(BrandA)
60-10
range(BrandB)
45-25
```

- Følgende resultat fås ved kommandoen.

```
> #Opgave 10
> BrandA <- c(10,60,50,30,40,20)
> BrandB <- c(35,45,30,35,40,25)
> #Opgave 10 a
> range(BrandA)
[1] 10 60
> 60-10
[1] 50
> range(BrandB)
[1] 25 45
> 45-25
[1] 20
```

b. Følgende måde er brugt i R.

```
#Opgave 10 b
var(BrandA)
var(BrandB)
```

- Følgende resultat fås ved kommandoen.

```
> #Opgave 10 b
> var(BrandA)
[1] 350
> var(BrandB)
[1] 50
```

c. Følgende måde er brugt i R.

```
#Opgave 10 c
sd(BrandA)
sd(BrandB)
```

- Følgende resultat fås ved kommandoen.

```
> #Opgave 10 c
> sd(BrandA)
[1] 18.70829
> sd(BrandB)
[1] 7.071068
```

3. If the variance of a distribution is 9, the standard deviation is:
- 3
 - 6
 - 9
 - 81
 - impossible to determine without knowing n.

- Vi kan I tilfældet se, at selve standardafvigelse bliver følgende:

$$\sqrt[2]{9} = 3$$

4. 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.

- Man skal tænke på følgende måde:

10-5 = 5, 9-5 = 4, 8-5 = 3, 7-5 = 2, 6 - 5 = 1, 5-5 = 0

0-5 = -5, -1+5 = -4, -2 +5 = 3, -3+5 = 2, -4+5 = 1, -5+5 = 0