Note om deskriptiv statistik

Indhold

Fra ugrupperede observationer til det udvidede kvartilsæt og boksplottet

Beregning af middelværdi og spredning samt bestemmelse af typetal

I denne note lærer man at bestemme kvartilsættet og det udvidede kvartilsæt ud fra et sæt af observationer. Man lærer også at tegne et boksplot, samt at bestemme middelværdien og spredningen. Derudover lærer man at bestemme variationsbredden, kvartilbredden, middelværdien og typetallet. Formelsamlingen (side 24 og side 25) viser tydeligt, hvad man skal gøre, men her følger nogle konkrete eksempler. Side 24 og 25 fra formelsamlingen er vedlagt dette dokument på siderne 4 og 5.

Eksempel. Det er vigtigt, at du forstår de begreber, der præsenteres. Tag dig tid til at se på siderne 4 og 5 løbende.

Du har deltaget i en konkurrence, og du har fået disse strafpoint.

4, 0, 8, 1, 8, 1, 8, 2, 3, 0, 9, 5, 6, 12, 8, 1, 2, 8, 8, 1, 8, 5, 7

1. For at bestemme kvartilsættet: Start med at sortere tallene ved at sætte tallene i rækkefølge med de mindste tal først.

   0, 0, 1, 1, 1, 1, 2, 2, 3, 4, 5, 5, 6, 7, 8, 8, 8, 8, 8, 8, 8, 9, 12

2. Bestem medianen, m, som er det midterste tal i listen.
   Tæl hvor mange tal, der er. Man tæller 23 tal, hvilket er et ulige tal. Derfor skal man ifølge formel (97) blot bestemme det midterste tal. Der må være 11 tal på hver side af medianen, som i dette tilfælde derfor er 5.

   0, 0, 1, 1, 1, 1, 2, 2, 3, 4, 5, 5, 6, 7, 8, 8, 8, 8, 8, 8, 8, 9, 12

3. Nu skal vi bestemme nedre kvartil, Q_1. Det ifølge formel (98) medianen for den nederste halvdel af observationerne. Vi trækker lige de tal ud her:

   0, 0, 1, 1, 1, 1, 2, 2, 3, 4, 5

   Der er 11 tal, så vi finder igen det midterste tal, som er 1.

   0, 0, 1, 1, 1, 1, 2, 2, 3, 4, 5

4. Øvre kvartil, Q_3, bestemmes tilsvarende for den øverste halvdel af observationerne, formel (99). Øvre kvartil er derfor 8.

   6, 7, 8, 8, 8, 8, 8, 8, 8, 9, 12

5. For at bestemme det udvidede kvartilsæt, skal man også bestemme dem mindste observation, min. Det er 0. Se formel (100).
6. Vi skal også bestemme den største observation, max. Det er 12. Se formel (101).
7. Benyt formel (104) til at skrive kvartilsættet. Man finder, at det er $(1, 5, 8)$.

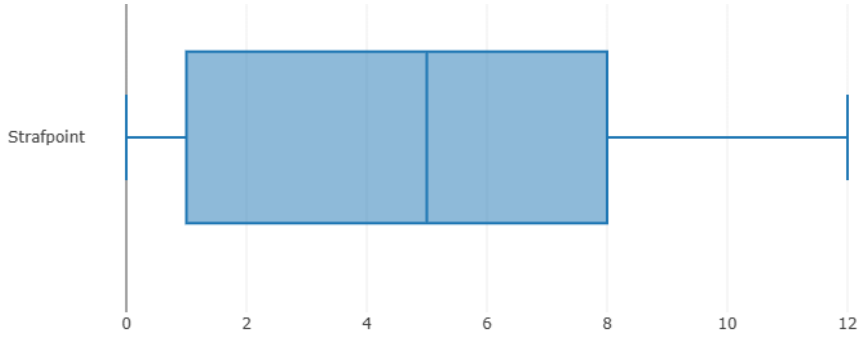
8. Benyt formel (105) til at bestemme det udvidede kvartilsæt. Man finder, at det er $(0, 1, 5, 8, 12)$

- Kvartilsættet inddeler observationerne i "kvarte" – altså i fjerdele.
- Når vi har opskrevet det udvidede kvartilsæt, kan vi nemt aflæse at:
  - Nedre kvartil: 25% af strafpointene var 1 eller færre.
  - Øvre kvartil: 75% af strafpointene var 8 eller færre.
  - Medianen: 50% af strafpointene var 5 eller færre.
  - Halvdelen af strafpointene er mellem 1 og 8.
- Kvartilsættet skjuler detaljer, som afsløres bedre, når man ser på et søjlediagram.

Boksplot

Du vi har bestemt det udvidede kvartilsæt, kan vi tegne et boksplot, se formel (106).

- Prøv at forstå den nedenstående figur, og tegn selv et boksplot på et stykke papir.
- Tænk over de enkelte kvartiler, og hvad selve boksen repræsenterer.



Variationsbredden

Benyt formel (102) til at bestemme variationsbredden. Her er den $12 - 0 = 12$.

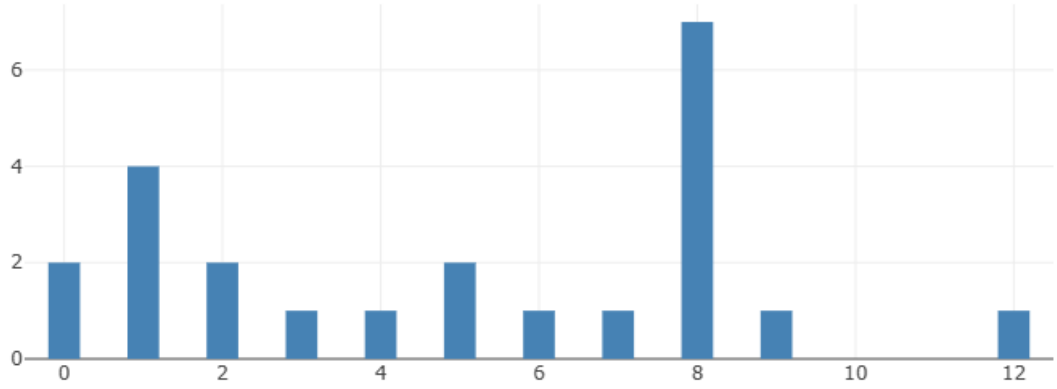
Kvartilbredden

Benyt formel (103) til at bestemme kvartilbredden. Her er den $8 - 1 = 7$.

Søjlediagram

Benyt formel (96) til at tegne et søjlediagram.

- Prøv at forstå den nedenstående figur, og tegn selv et søjlediagram på et stykke papir.
- Der er flere detaljer i søjlediagrammet end i boksplottet. Her ser man tydeligt, hvor mange gange der blev givet 1 strafpoint.
  Diagrammet i formelsamlingen er lidt anderledes mht. akserne. Brug formelsamlingens måde at tegne på til din figur.



Middelværdien

Benyt formel (107). Vi lægger alle observationernes værdier sammen, og derefter dividerer vi med antallet af observationer.

$$\bar{x} = \frac{0+0+1+1+1+1+2+2+3+4+5+5+6+7+8+8+8+8+8+8+8+9+12}{23} = 5$$

- Overbevis dig selv om, at du forstår, hvordan formlen er blevet benyttet.

Spredningen

Benyt formel (108) til at beregne spredningen.

$$\sigma = \sqrt{\frac{(0-5)^2+(0-5)^2+(1-5)^2+(1-5)^2+\cdots+(9-5)^2+(12-5)^2}{23}} = 3{,}43$$

- Overbevis dig selv om, at du forstår, hvordan formlen er blevet benyttet.
- Vi ser senere på en lidt anden formel for spredningen for en stikprøve, hvor man dividerer med $n-1$.

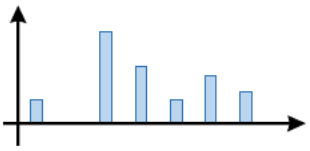
Typetallet

Tallet 8 er det tal med den største hyppighed i datasættet. Derfor er typetallet 8.

24 DESKRIPTIV STATISTIK

Ugrupperede observationer

Søjlediagram (96)



Median m (97)

Observationerne sorteres efter størrelse.

Hvis antallet af observationer er ulige, så er m værdien af den midterste observation.

$m = 5$

2 3 4 5 5 8 8 8 11

Hvis antallet af observationer er lige, så er m gennemsnittet af de to midterste observationer.

$m = \frac{8+5}{2} = 6{,}5$

2 3 4 5 5 8 8 8 11 13

Nedre kvartil Q_1 (98)

Q_1 er medianen for de observationer, der ligger til venstre for medianen m.

Q_1 m

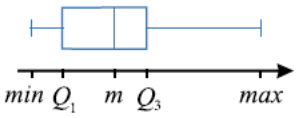
2 3 4 5 5 8 8 8 11

Øvre kvartil Q_3 (99)

Q_3 er medianen for de observationer, der ligger til højre for medianen m.

m Q_3

2 3 4 5 5 8 8 8 11 13

Observationen *min*	(100)	*min*: mindste observation
Observationen *max*	(101)	*max*: største observation
Variationsbredde *VB*	(102)	$VB = max - min$
Kvartilbredde *KB*	(103)	$KB = Q_3 - Q_1$
Kvartilsæt	(104)	(Q_1, m, Q_3)
Udvidet kvartilsæt	(105)	(min, Q_1, m, Q_3, max)
Boksplot	(106)	
Middelværdi (gennemsnit) $\overline{x}$ for observationssættet $x_1, x_2, \ldots, x_n$	(107)	$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$
Spredning σ for observationssættet $x_1, x_2, \ldots, x_n$	(108)	$\sigma = \sqrt{\frac{(x_1 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n}}$

Opgave 1

En familie har haft en del hamstere og har registreret deres levealder i måneder:

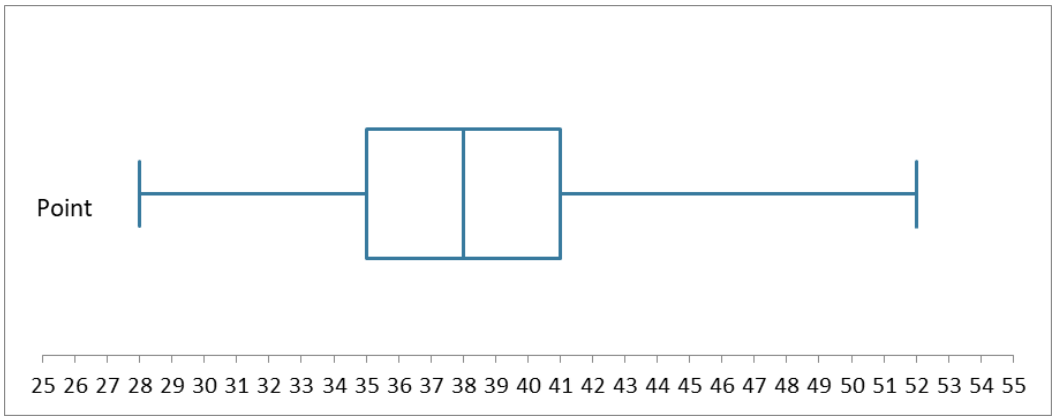
21, 24, 25, 8, 26, 23, 16, 18, 25, 23

a) Bestem medianen.
b) Bestem middelværdien for hamsternes alder.
c) Tegn et boksplot.

Opgave 2

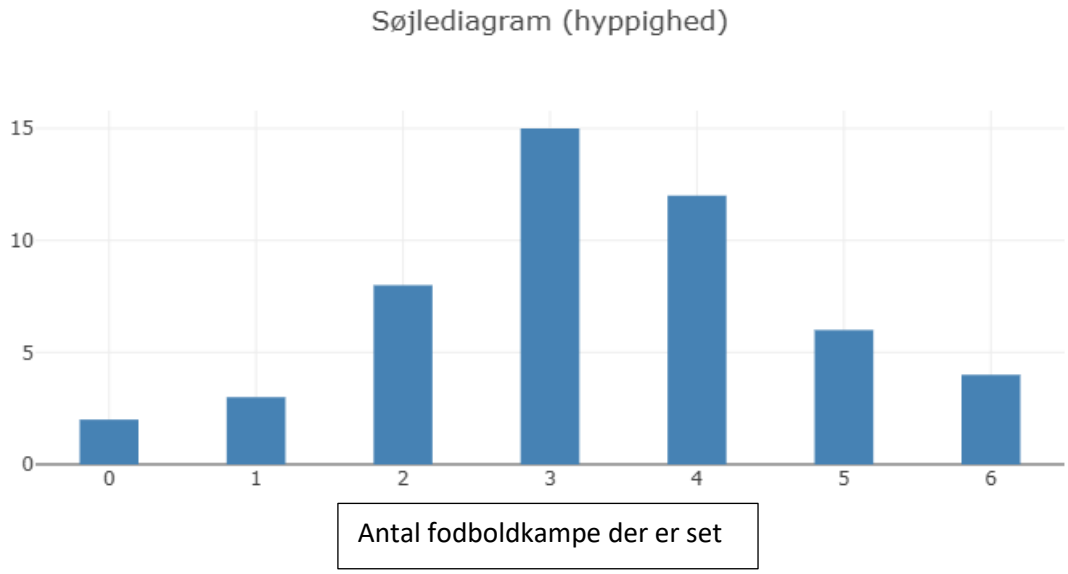
Se på dette boksplot, som viser point for en lille fodboldklub gennem flere sæsoner.

a) Aflæs det udvidede kvartilsæt.
b) Er det korrekt, at holdet i en fjerdedel af sæsonerne har fået mellem 38 og 41 point?



Opgave 3

I en undersøgelse har man spurgt 50 elever om, hvor mange fodboldkampe de har set i fjernsynet i den seneste måned. Figuren viser et søjlediagram over resultatet.



a) Bestem middelværdien for antallet af fodboldkampe, der er set.
b) Bestem medianen for antallet af fodboldkampe, der er set.
c) Hvor mange procent har set højst 2 fodboldkampe?

Opgave 4

En elev har fået følgende karakterer ved studentereksamen:

10, 4, 00, 4, 7, 02, 12, 7, 02, 4, 4, 7, 02, 7, 10

a) Bestem middelværdien for karaktererne.
b) Bestem det udvidede kvartilsæt.
c) Tegn et boksplot.
d) Tegn et søjlediagram.
e) Bestem kvartilbredden.
f) Hvor stor en procentdel af karaktererne (antal) udgøres af karaktererne 00 og 02?
g) Hvor stor en procentdel af karaktererne (antal) udgøres af karaktererne 4 og 7?
h) Bestem spredningen for karaktererne – skriv op hvordan det beregnes.

Kvartilsættet og boksplot samt outlier

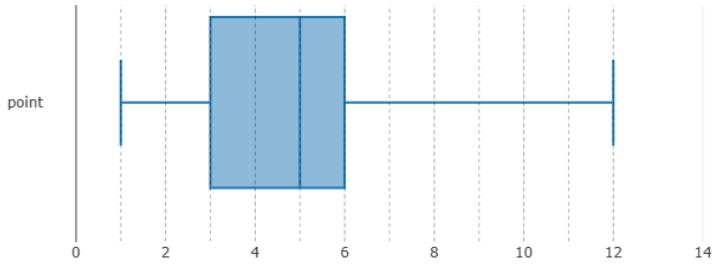
Vi har allerede set på boksplottet, og vi har set hvordan det fremstilles fra et udvidet kvartilsæt. Her gentages det i kompakt form.

Man starter med en tallinje. Ud fra det udvidede kvartilsæt sætter man:

- En streg ved værdien for den mindste observation - min .
- En streg ved nedre kvartil, Q_1
- En streg ved medianen, m.
- En streg ved øvre kvartil, Q_3.
- En streg ved værdien for den største observation - max .
- Tegn en boks rundt om mellem Q_1 og Q_3.

Eksempel

Det udvidede kvartilsæt (1, 3, 5, 6, 12) giver nedenstående boksplot.



Outlier

Har man et sæt af observationer, kan der være observationer, som ligger langt væk fra den typiske observation. Dette er en såkaldt *outlier*. Der kan være mere end en outlier i et datasæt. Man bestemmer, om en observation er en outlier med denne metode:

> 'Observation der ligger mere end halvanden kvartilbredde under nedre kvartil eller mere end halvanden kvartilbredde over øvre kvartil.'
>
> Derfor skal man beregne $Q_1 - 1{,}5 \cdot KB$ og $Q_3 + 1{,}5 \cdot KB$, hvor KB er kvartilbredden givet ved $KB = Q_3 - Q_1$.

Eksempel på om vi har en outlier

Vi har observationerne 1, 2, 3, 3, 3, 4, 4, 5, 6, 7, 12. Er 12 en outlier?

Vi bestemmer det udvidede kvartilsæt til at være (1, 3, 4, 6, 12).

Kvartilbredden er derfor $Q_3 - Q_1 = 6 - 3 = 3$

Ligger 12 'halvanden kvartilbredde over øvre kvartil"? Vi regner:

$$Q_3 + 1{,}5 \cdot KB = Q_3 + 1{,}5 \cdot (Q_3 - Q_1) = 6 + 1{,}5 \cdot (6 - 3) = 6 + 4{,}5 = 10{,}5$$

Ja! Da 12 er større end 10,5 er 12 en outlier.

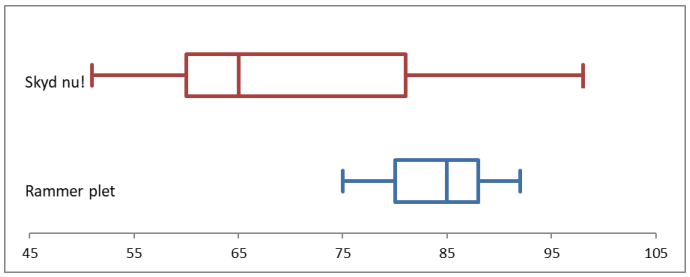
Hvad fortæller kvartilsættet og boksplottet os?

Kvartilsættet inddeler observationerne i fire dele. Her er nedre kvartil, median og øvre kvartil markeret.

2, 2, 3, **3**, 4, 5, 5, **6**, 7, 8, 8, **8**, 9, 12, 12

Boksplottet er en visuel fremstilling af det udvidede kvartilsæt, som giver os et hurtigt overblik over observationerne.

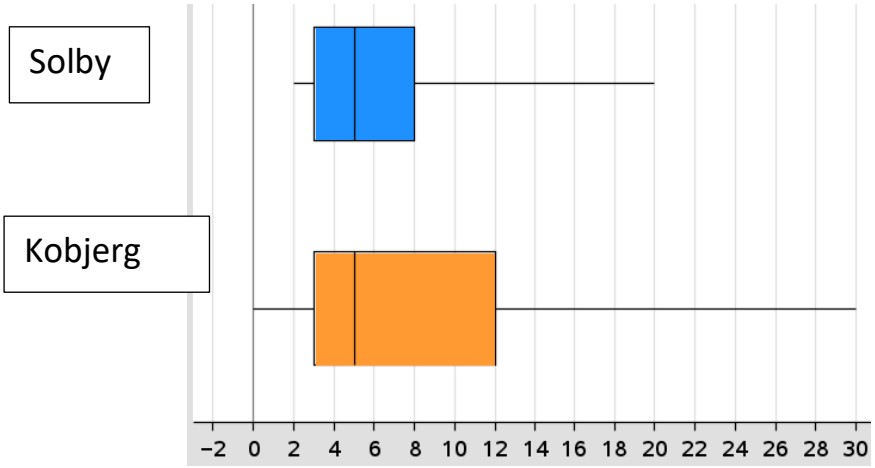
Specielt er boksplot nyttige, når man vil sammenligne datasæt. Vi ser på et eksempel, hvor man har observationer fra point i en skydekonkurrence for holdet "Rammer Plet" og holdet "Skyd nu!"



- Medianen er mindst for 'Skyd nu', så deres niveau er lavest.
- Man ser også nemt, at holdet 'Skyd nu!' har skytten med den højeste og den laveste pointscore.
- Variationsbredden er størst for 'Skyd nu'. Holdet har skytten der får flest og færrest point.
- Vi ser også, at kvartilbredden er væsentligt større for 'Skyd nu!' end for 'Rammer plet'. Variationen i antal point er derfor mindre for 'Rammer Plet'.
- 50% af skytterne fra "Skyd nu!" får mindre end 65 point.
    - Vi kan også sige, at det er mellem 51 og 65 point.
- 25% af skytterne fra "Skyd nu" får mellem 88 og 93 point.

Opgave 5

Du undersøger alderen på folks cykler i to byer Solby og Kobjerg. Her er resultatet af dine observationer vist i boksplot



a) Skriv de udvidede kvartilsæt op for begge boksplot.
b) Sammenlign de to boksplot.

   Inddrag ord som nedre kvartil, median, øvre kvartil, kvartilbredde og variationsbredde.
   Tænk på hvad fx øvre kvartil fortæller noget om (i procent!)

c) Afgør om der er en outlier i observationerne for Kobjerg.

   Man skal beregne:
   $Q_1 - \frac{3}{2} \cdot KB$ og $Q_3 + \frac{3}{2} \cdot KB$, hvor KB er kvartilbredden.

Opgave 6

Benyt datasættet fra 'Elevundersøgelsen på Falkonergården' til at vise og sammenligne boksplot for pigers og drenges skostørrelse.

a) Bestem det udvidede kvartilsæt.
b) Bestem middelværdien.
c) Tegn tre søjlediagrammer.
   Et for drenge, et for pigerne og et samlet for alle elever.
d) Afgør hvor stor en procentdel af drengene der bruger mindst størrelse 44.
e) Afgør hvor stor en procentdel af pigerne der bruger højst skostørrelsen 40.

Ekstra: Vælg selv et af de andre observationssæt fra elevundersøgelsen og sammenlign igen boksplot for piger og drenge.

Tabel med hyppigheder og frekvenser

Her arbejder vi stadig med ugrupperede observationer, som fx kan være, hvor mange timer en person har arbejdet de seneste 14 arbejdsdage:

5, 12, 8, 9, 6, 4, 8, 3, 10, 7, 8, 9, 5, 8

Med deskriptiv statistisk benytter vi tabeller, deskriptorer og diagrammer til at få et bedre overblik over listen af tal.

Før vi starter på en opgave, defineres (igen) nogle begreber.

- Stikprøve. Den fremkommer ud fra en population, og består af elementer, som fx kan være arbejdstid. Fx kan hele den danske befolkning være en population, og stiller man 1000 personer et spørgsmål om deres arbejdstid, har man udtaget en stikprøve. Observationerne kalder man for et datasæt eller et observationssæt.
- Ikke-grupperet datasæt er observationer med enkeltstående tal eller andre typer of observationer.
- De ikke grupperede observationer kan beskrives med statistiske deskriptorer.
- Observationssættets størrelse er antallet af observationer. (Det er 14 i eksemplet med arbejdstid fra tidligere.)
- Hyppigheden $h(x)$ af en observation x er det antal gange en observation optræder i et datasæt.
- Frekvensen $f(x)$ er den procentdel, som hyppigheden udgør af hele observationssættets størrelse.
- Den kumulerede hyppighed $H(x)$ er summen af hyppighederne af alle de observationer, der er mindre end eller lig med x.
- Den kumulerede frekvens $F(x)$ er summen af frekvenserne af alle de observationer, der er mindre end eller lig med x.

Her er et eksempel på en tabel med 22 elevers karakterer i historie ved studentereksamen. Kan du forstå denne tabel? Tænk tilbage til forløbet om binomialfordeling.

- Hvor mange elever fik karakteren 7?
- Hvor mange elever fik karakteren 10 eller 12?
- Hvor stor en andel (procent) af eleverne fik en karakter på højst 4?
- Hvor stor en andel (procent) af eleverne fik en karakter på 02 eller 7?

karakter	hyppighed	kumuleret hyppighed	frekvens	kumuleret frekvens
00	3	3	0,136	0,136
02	2	5	0,091	0,227
4	5	10	0,227	0,455
7	6	16	0,273	0,727
10	2	18	0,091	0,818
12	4	22	0,182	1

Opgave 7

I tabellen i filen 'Elevundersøgelsen på Falkonergården' skal man arbejde med kolonnen 'skostørrelser'. Hvis man er pige, så vælger man observationerne for D (drenge), og hvis man er en dreng, så vælger man observationerne med D (pige).

En tabel som denne skal altså udfyldes 'i hånden', her for P med start ved skostørrelse 35.

skostørrelse	hyppighed	kumuleret hyppighed	frekvens	kumuleret frekvens
35				
36				
..				
..				
..				
..				

- Bestem det udvidede kvartilsæt.
- Tegn et søjlediagram med hyppigheder.
- Tegn et søjlediagram med frekvenser.

Kopier en udfyldt tabel fra den anden gruppe, så du har en tabel for pigers og drenges skostørrelser.

  - Tegn boksplot for begge kvartilsæt på samme figur (over hinanden).
  - Sammenlign de to boksplot.

- For datasættet med drenge: Hvilket procentdel af drengene bruger en skostørrelse som er mindst 43 og højst 45?
- For datasættet med piger: Hvilket procentdel af pigerne bruger en skostørrelse som er enten 38 eller 39?

Opgave 8

I tabellen i filen 'Elevundersøgelsen på Falkonergården' skal man arbejde med en kolonne, som man tænker, giver mening i forhold til deskriptiv statistik.

Denne gang tænker man på pigerne og drengene som en samlet pulje.

- Udfyld en tabel med hyppigheder, kumulerede hyppigheder, frekvenser og kumulerede frekvenser.
- Formuler og besvar selv nogle spørgsmål i relation til tabellen og søjlediagrammet.
  Vis tabel og diagrammer samt spørgsmål til din makker – og lad makkeren besvare spørgsmålene.

Spredning

Spredning er ofte et begreb, som er lidt vanskeligt at forstå, men lad os se på et helt konkret eksempel her, hvor man gerne skal se, at spredningen er et mål for, hvor spredt datasættet er.

Vi betragter to lister med tal, der er 19 tal i begge lister.

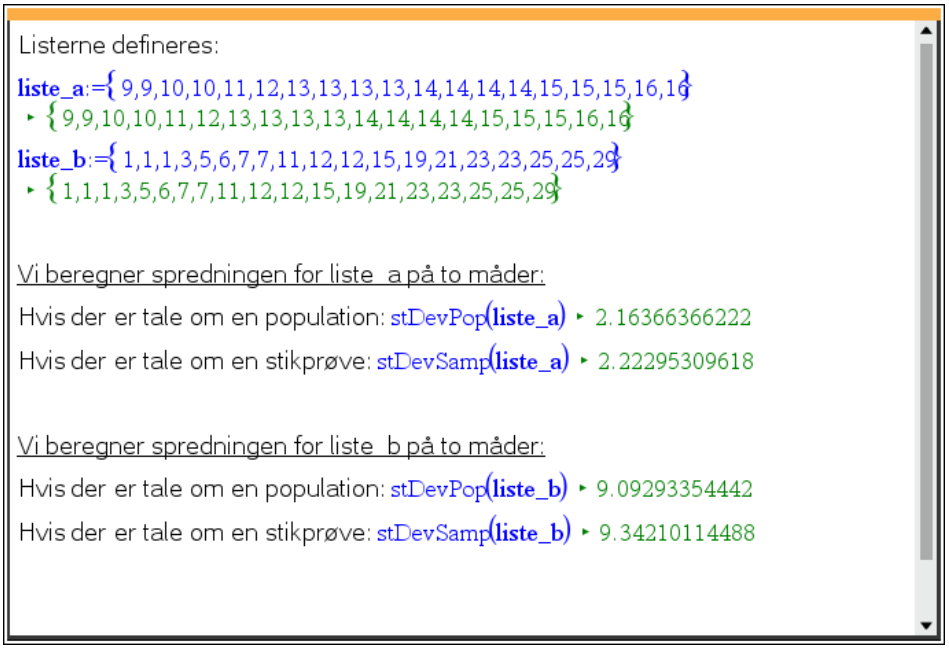
Liste A: 9, 9, 10, 10, 11, 12, 13, 13, 13, 13, 14, 14, 14, 14, 15, 15, 15, 16, 16

Liste B: 1, 1, 1, 3, 5, 6, 7, 7, 11, 12, 12, 15, 19, 21, 23, 23, 25, 25, 29

Overvejelse: Se på de to lister af tal. I hvilken liste tænker du, at tallene er "mest spredt ud"?

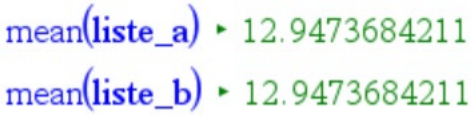
Det er ikke helt enkelt at se, og i matematik "ser" vi heller ikke bare, hvad spredningen er, vi definerer, hvad vi mener med spredningen og vi beregner spredningen.

Vi benytter først Nspire til beregningen.

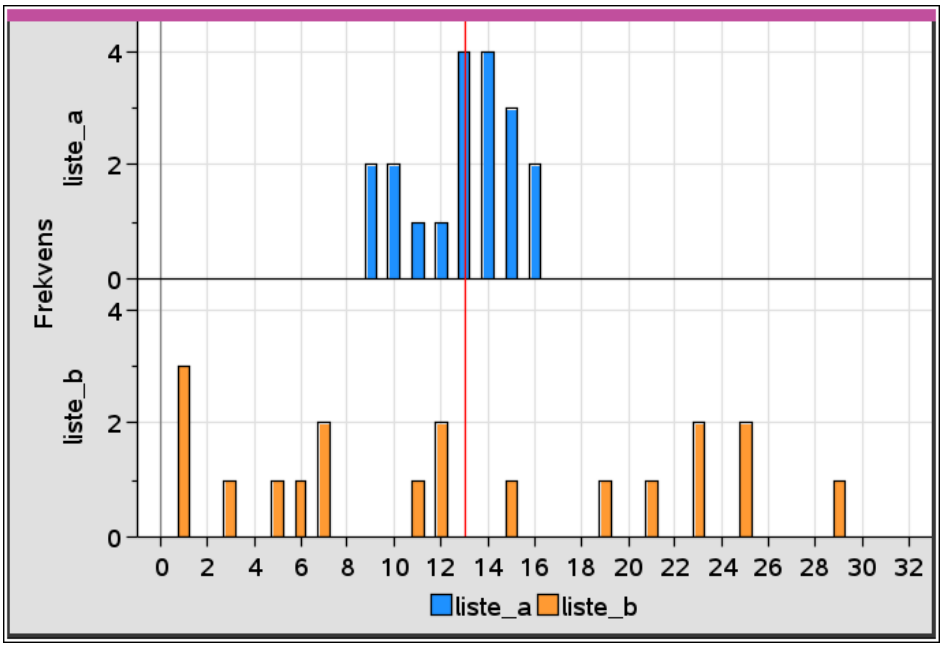


Vi kan nu konstatere, at spredningen er størst for Liste B, hvilket du måske også tænkte.

For at forstå begrebet spredning bedre, skal vi beregne middelværdien for tallene i de to lister. Her benytter vi igen Nspire, og vi konstaterer at middelværdierne for begge lister er 12,95.



Lad os nu illustrere tallene i listerne i et søjlediagram, hvor middelværdien er vist med en rød lodret streg.



Det man kan se her er, at tallene for Liste A er tæt på middelværdien, mens tallene for Liste B er længere væk fra middelværdien. Man ser, at tallene er spredt ud over et større talområde for liste B i forhold til Liste A.

Overvejelse: Giver det intuitivt mening (for dig), at spredningen er størst for Liste B, når man ser på søjlediagrammerne?

Dette er en måde at forstå spredningen på, men hvordan definerer vi spredningen i matematik mere stringent, og hvordan beregnes spredningen egentlig "bag" Nspire-kommandoen fra tidligere? Det ser vi på i det næste afsnit.

,

Beregningen af spredningen (formel) for en population eller en stikprøve

Spredningen blev tidligere beregnet på to forskellige måder i Nspire alt efter, om det er spredningen for en population eller en stikprøve. Man bemærker, at spredningen er størst, når vi beregner for en stikprøve. Grunden er, at man benytter to forskellige formler. Vi kommer her ikke nærmere ind på detaljerne i forskellen for de to formler. Formlernes numre i formelsamlingen er (154) og (154a).

Når beregningen foretages 'i hånden' bestemmer man først variansen, som er

den gennemsnitlige kvadratafstand til middelværdien

og som i vores tilfælde er (for spredningen af en population):

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_{18} - \bar{x})^2 + (x_{19} - \bar{x})^2}{n}$$

Når vi sætter ind i denne formel er $\bar{x} = 12{,}95$ og n er antal tal i listen så $n = 19$.

For $x_1, x_2, \ldots, x_{18}, x_{19}$ benytter vi Liste A, så derfor er fx $x_1 = 9$, $x_5 = 11$ og $x_{19} = 16$

$$\sigma^2 = \frac{(9 - 12{,}95)^2 + (9 - 12{,}95)^2 + \cdots + (16 - 12{,}95)^2 + (16 - 12{,}95)^2}{n}$$

Formel (154) fremkommer ved, at spredningen er kvadratroden af variansen. Her skrives formlen op på en måde, så den måske er lidt nemmere at forstå. Vi husker, at vi havde 19 tal i begge lister, og vi fokuserer her på Liste A.

$$\sigma = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_{18} - \bar{x})^2 + (x_{19} - \bar{x})^2}{n}}$$

$$= \sqrt{\frac{(9 - 12{,}95)^2 + (9 - 12{,}95)^2 + \cdots + (16 - 12{,}95)^2 + (16 - 12{,}95)^2}{19}}$$

For en stikprøve betegnes spredningen ikke med σ men med s. I brøken divideres ikke med n men med $n - 1$.

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_{18} - \bar{x})^2 + (x_{19} - \bar{x})^2}{n - 1}}$$