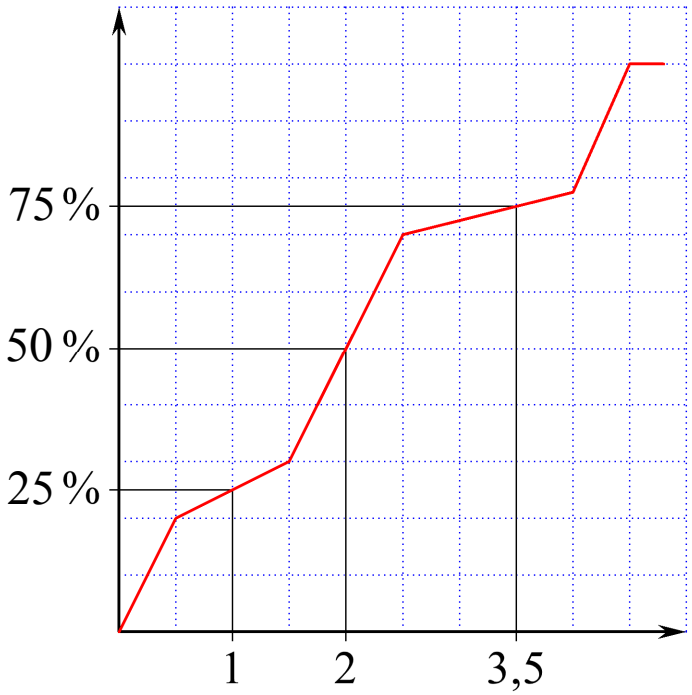
Nogle emner fra

Deskriptiv Statistik



2011 Karsten Juul

Indhold

Nyere hæfte:
http://mat1.dk/deskriptiv_statistik_for_c_niveau_i_hf.pdf

Nogle emner fra deskriptiv statistik

Hvad er deskriptiv statistik?

Den deskriptive statistik består af metoder til at få overblik over nogle indsamlede tal .

De indsamlede tal kaldes observationer .

I dette hæfte er der eksempler på nogle af metoderne fra den deskriptive statistik

UGRUPPEREDE OBSERVATIONER

Hyppigheder

På en skole har alle elever taget tid på hvor mange minutter det tager dem at komme fra deres hjem til skolen.

Rækken af observationer er uoverskuelig:

$$7\ 5\ \ 8\ 7\ 4\ 6\ 6\ 9\ \ \ldots\ \ 8\ 6\ 7\ 7\ 5$$

For at få overblik over observationerne tæller vi op hvor mange gange observationen 4 forekommer, hvor mange gange observationen 5 forekommer, osv.

Hyppigheden af en observation er antallet af gange den forekommer.

Observationerne ovenfor har følgende hyppigheder:

Observation:	4	5	6	7	8	9
Hyppighed:	16	20	24	26	28	25

Det samlede antal observationer

Af tabellen ser vi at der er 16 firtaller, 20 femtaller, osv.
Det samlede antal observationer er altså

$$16 + 20 + 24 + 26 + 28 + 25 \ = \ 139$$

Middeltallet

Middeltallet for nogle observationer er det vi plejer at kalde gennemsnittet.
Vi kan udregne middeltallet ved at lægge observationerne sammen og dividere resultatet med antallet af observationer.

Når vi lægger de 16 firtaller sammen får vi $4 \cdot 16$,
når vi lægger de 20 femtaller sammen får vi $5 \cdot 20$,
osv.

Middeltallet kan vi altså udregne sådan:

$$\frac{4 \cdot 16 + 5 \cdot 20 + 6 \cdot 24 + 7 \cdot 26 + 8 \cdot 28 + 9 \cdot 25}{139} \ = \ 6{,}75540 \ \approx \ 6{,}8$$

Frekvenser

Observationerne ovenfor er fra 1985. I 2000 var tallene sådan:

Observation:	4	5	6	7	8	9
Hyppighed:	13	16	10	9	16	20

I 1985 er der flere der har 4 minutter til skole end i 2000,
men der er også flere elever på skolen,
så tallene kan ikke uden videre sammenlignes.

I stedet kan vi spørge om der i 1985 var en større procentdel af eleverne der havde 4 minutter til skole.

Frekvensen af en observation er hyppigheden divideret med det samlede antal observationer.

I 1985 er frekvensen af 4 lig

$$\frac{16}{139} = 0{,}115108 \approx 12\%$$

At 4 har frekvensen 12%, betyder at firtallerne udgør 12% af observationerne.

Vi udregner også de andre frekvenser:

Observation:	4	5	6	7	8	9
1985 Frekvens i %:	12	14	17	19	20	18
2000 Frekvens i %:	15	19	12	11	19	24

Ekempel: Vi ser at de der har 5 minutter eller mindre til skole, udgør

$$12\% + 14\% = 26\% \quad \text{i 1985}$$

$$15\% + 19\% = 34\% \quad \text{i 2000.}$$

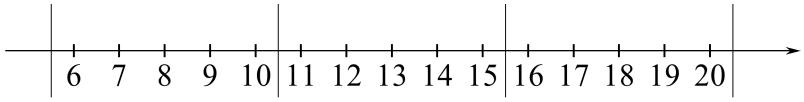
GRUPPEREDE OBSERVATIONER

Gruppering af observationer

I 2010 er forholdene helt ændret:

obs.	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
%	2	6	6	3	7	6	4	8	7	10	9	10	12	7	3

Dette er uoverskueligt.
Vi kan gøre det mere overskueligt ved at gruppere observationerne i intervaller.
Vi kan f.eks. gruppere i følgende tre intervaller:



Interval	5,5 - 10,5	10,5 - 15,5	15,5 - 20,5
Frekvens	24 %	35 %	41 %

Denne tabel er fremkommet sådan:

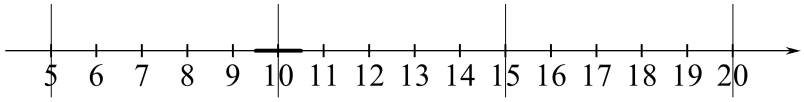
I intervallet 5,5 - 10,5 ligger de observationer der er 6, 7, 8, 9 eller 10.
Disses frekvenser lægger vi sammen:

2 % + 6 % + 6 % + 3 % + 7 % = 24 %

Dette betyder at 24 % af observationerne ligger i intervallet 5,5 - 10,5 .

Hvorfor lader vi det andet interval begynde ved 10,5 i stedet for 10 ?

Hvis vi lader andet interval begynde ved 10, så ser det sådan ud:



Hvis nogle elever f.eks. har 9,8 eller 10,4 minutter til skole,
så har vi skrevet at de har 10 minutter til skole.

I tabellen står at 7 % af observationer er 10.
Dette betyder at 7 % af tidene ligger mellem 9,5 og 10,5 .

Når vi deler ved 10 i stedet for 10,5 ,
så vil nogle af de 7 % ligge første interval, og andre af dem i andet interval.

Vi har valgt 10,5 i stedet for 10 for at opnå at alle 7 % kommer i samme interval.

Hvis vi deler ved 10 er der bl.a. følgende muligheder:

1. De 7 % fordeles mellem første og andet interval så vi tæller dem med som 3,5 % i hvert af de to intervaller.
2. Alle 7 % tæller med i første interval.
   Dvs. i intervallerne 5-10, 10-15, 15-20 hører det højre endepunkt med til intervallet.
3. Alle 7 % tæller med i andet interval.
   Dvs. i intervallerne 5-10, 10-15, 15-20 hører det venstre endepunkt med til intervallet.

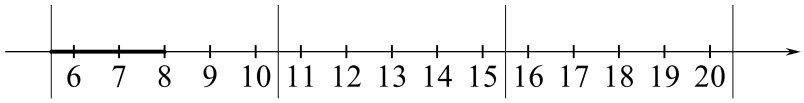
I det danske gymnasium er der tradition for at bruge 2. mulighed.

Nspire bruger 3. mulighed.

Fejlen er lige stor i de to metoder, men går hver sin vej.

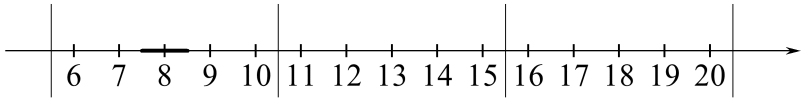
Observationerne er jævnt fordelt i hvert interval

I tabellen hvor observationerne er grupperet ved vi kun at 24 % ligger i første interval. Vi ved ikke om der ligger flest i venstre del eller højre del eller et andet sted. Derfor regner vi som om observationerne er jævnt fordelt i intervallet.



Da observationerne er jævnt fordelt i første interval,
ligger halvdelen af de 24 % i første halvdel af intervallet.
Det betyder at

12 % af eleverne har 8 minutter eller mindre til skole.



Da intervallet 7,5 - 8,5 har længde 1,
udgør det en femtedel af intervallet 5,5 - 10,5 .
En femtedel af 24 % er 4,8 % , så

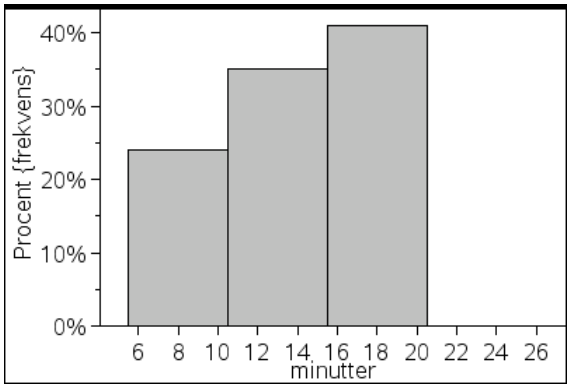
4,8 % af eleverne har ca. 8 minutter til skole,

men

0 % af eleverne har præcis 8 minutter til skole.

At 0 % af eleverne har præcis 8 minutter til skole, fortæller ikke noget om virkeligheden. Det er en egenskab ved de grupperede observationer.

Histogram



A minutter	B frekvens
6	2
7	6
8	6
9	3
10	7
11	6
12	4
13	8
14	7
15	10
16	9
17	10
18	12
19	7
20	3

(1) Figuren ovenfor hedder et histogram .
Histogrammet viser de grupperede observationer.

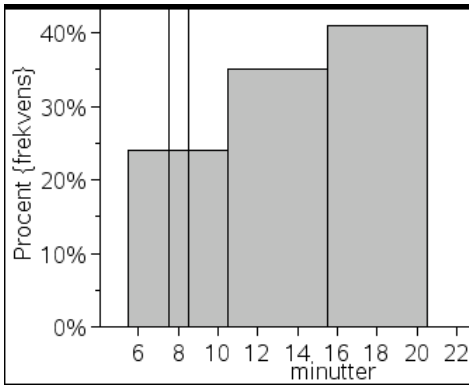
(2) Over hvert af de tre intervaller er der et rektangel der viser frekvensen.
Toppen af første rektangel er ud for 24 % på den lodrette akse.
Det betyder at arealet af første rektangel er 24 % .

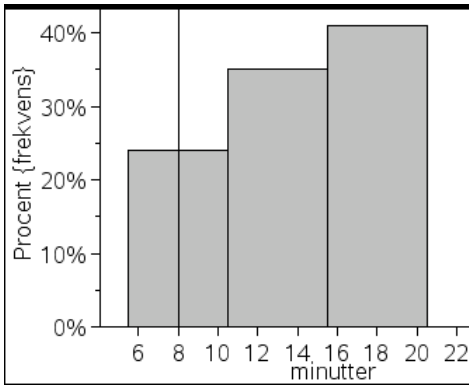
(3) Det er arealet der viser frekvensen.

Første halvdel af første rektangel har arealet 12 % .
Dette betyder at 12 % af eleverne har 8 minutter eller derunder til skole.

Nedenfor til venstre er vist den del af første rektangel som svarer til intervallet 7,5 - 8,5.
Denne del er en femtedel af rektanglet, så arealet er $\frac{24\%}{5} = 4{,}8\%$.
Dette betyder at 4,8 % af eleverne har ca. 8 minutter til skole.

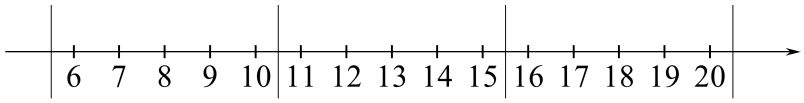
Nedenfor til højre er vist den del af rektanglet der svarer til præcis 8 minutter.
Arealet af denne del er 0 %, så 0 % af eleverne har præcis 8 minutter til skole.
Dette gælder på histogrammet. Det fortæller intet om virkeligheden.





Sumkurve

Vi ser stadig på følgende grupperede observationer:



Interval	5,5 - 10,5	10,5 - 15,5	15,5 - 20,5
Frekvens	24 %	35 %	41 %

Hvor mange procent af observationerne er 13 eller derunder?

Observationerne i første interval er under 13.
Disse udgør 24 %.

Observationerne i første halvdel af andet interval er også under 13.
Disse udgør halvdelen af 35 % , altså 17,5 % .

24 % + 17,5 % = 41,5 % , så

41,5 % af observationerne er 13 eller derunder.

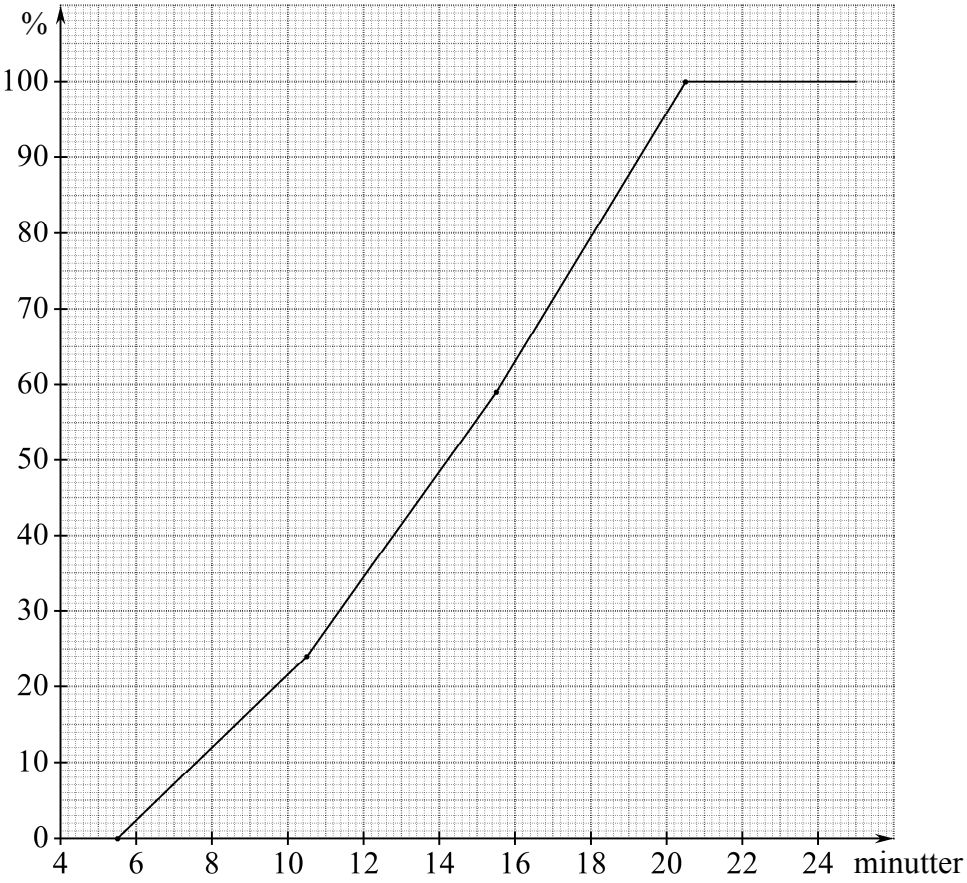
Sådan aflæser vi på en sumkurve

Svaret på spørgsmålet

"Hvor mange procent af observationerne er 13 eller derunder?"

kan vi nemt aflæse på sumkurven nedenfor:

Vi begynder på den vandrette akse ved 13, går lodret op til kurven, og går vandret ind på den lodrette akse. Her aflæser vi svaret 41,5 % .



Sådan gør vi for at få tegnet sumkurven.

Først ser vi på tabellen der viser de grupperede observationer,
og finder ud af at

0 % af observationerne er 5,5 eller mindre
24 % af observationerne er 10,5 eller mindre
59 % af observationerne er 15,5 eller mindre
100 % af observationerne er 20,5 eller mindre

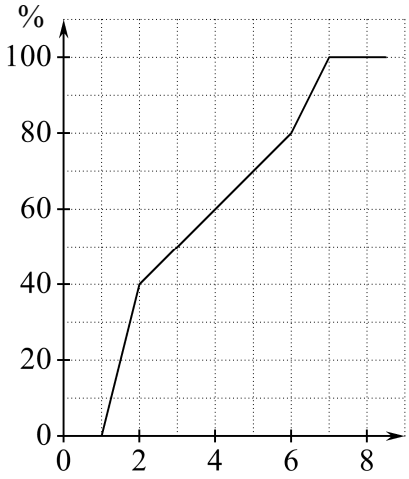
Disse tal bruger vi til at afsætte fire punkter i koordinatsystemet.
Det er de 4 prikker.

Mellem to af disse punkter er grafen lineær
da observationerne er jævnt fordelt i hvert af de tre intervaller.

Kurvens knæk

Ved 10,5 på den vandrette akse har kurven et knæk.
Dette fortæller ikke noget om virkeligheden.
Det fortæller kun at vi har valgt at begynde på et nyt interval her.

Eksempel der skal give indsigt i brugen af sumkurve

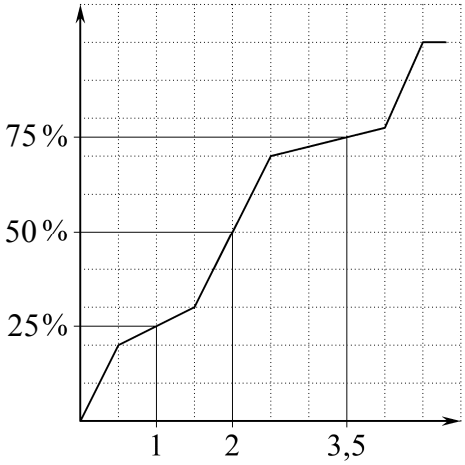


70 % af obs. er 5 eller mindre (aflæst).
100 % – 70 % = 30 % af obs. er større end 5 .

50 % af obs. er 3 eller mindre (aflæst).
70 % – 50 % = 20 % af obs. er mellem 3 og 5 .

65 % – 55 % = 10 % er mellem 3,5 og 4,5 .
2 % er mellem 3,9 og 4,1 .
0,2 % er mellem 3,99 og 4,01 .
0 % er præcis 4 .
1 % er ca. 4 , hvis vi vedtager at "ca. 4"
betyder "mellem 3,95 og 4,05" .

Median og kvartiler for grupperede observationer



På figuren har vi vist hvordan vi aflæser
de tre kvartiler.

Vi starter på den lodrette akse ved
25 %, 50 % og 75 %.

Vi går vandret ud til sumkurven
og lodret ned på den vandrette akse.
Vi ender i tallene 1 , 2 og 3,5 .

Disse tre tal kaldes kvartilerne fordi
de deler observationerne op i fire portioner
der hver indeholder 25 % af observationerne.

1 er den nedre kvartil .

2 er medianen .

3,5 er den øvre kvartil .

BOKSPLOT

For grupperede observationer aflæser vi kvartilerne på sumkurven .

For ugrupperede observationer finder vi kvartilerne sådan som det er vist nedenfor .

Median for ugrupperede observationer

En klasse har haft en prøve. De 17 elever fik følgende point:

(1) 52 69 70 20 47 71 48 27 27 62 15 48 23 52 49 39 36

Vi ordner disse tal efter størrelse så tallet til venstre er mindst:

(2) $\overline{15\ 20\ 23\ 27\ 27\ 36\ 39\ 47}\ 48\ \overline{48\ 49\ 52\ 52\ 62\ 69\ 70\ 71}$

Vi ser at det midterste af tallene er 48. Man siger at tallenes median er 48 .

Antag at der i stedet havde været et lige antal tal:

(3) $\overline{3\ 3\ 4\ 5}\ \overline{6\ 6\ 7\ 8}$

Da der er et lige antal tal, er der ikke et tal der står i midten. I stedet udregner vi gennemsnittet af de to midterste tal:

$$\frac{5+6}{2} = 5{,}5 \ .$$

Man siger at tallenes median er 5,5 .

Nedre og øvre kvartil for ugrupperede observationer

I linjen (2) står 8 tal til venstre for det midterste tal. Vi ser at disse 8 tals median er 27:

$\overline{15\ 20\ 23\ 27}\ \overline{27\ 36\ 39\ 47}\ 48\ \overline{48\ 49\ 52\ 52\ 62\ 69\ 70\ 71}$

Man siger at 27 er nedre kvartil for tallene i linjen (1).

I linjen (2) står 8 tal til højre for det midterste tal. Vi ser at disse 8 tals median er 57. Man siger at 57 er øvre kvartil for tallene i linjen (1).

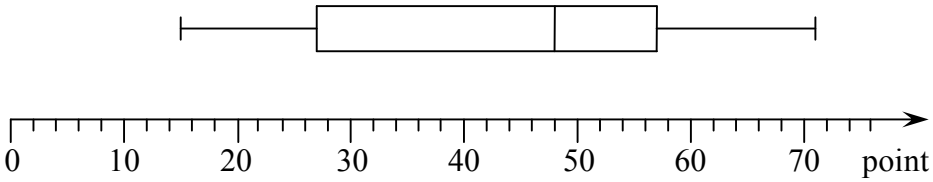
Ved kvartilerne for nogle tal forstås følgende tre tal:

nedre kvartil , median og øvre kvartil.

Kvartilerne for tallene i linjen (1) er altså de tre tal 27, 48 og 57 .

Boksplot

Figuren nedenfor viser kvartiler samt største- og mindsteværdi for tallene i linjen (1). En sådan figur kaldes et boksplot .



Kassen viser hvor den midterste halvdel af tallene ligger.

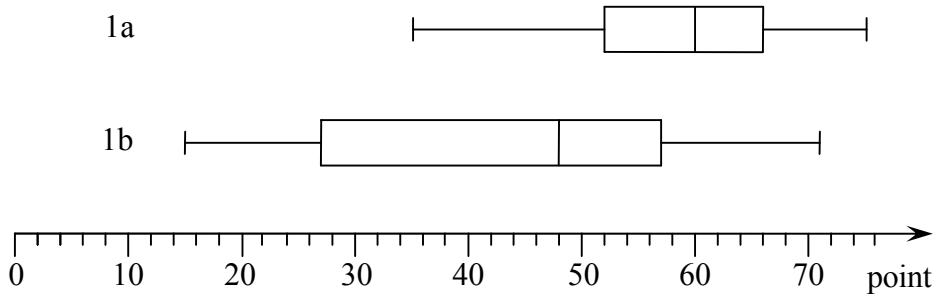
Den vandrette streg i venstre ende af boksplottet viser hvor den fjerdedel af tallene der er mindst, ligger.

Den vandrette streg i højre ende af boksplottet viser hvor den fjerdedel af tallene der er størst, ligger.

Sammenligning af boksplot

Boksplot er især nyttige når man vil sammenligne tal fra forskellige steder, f.eks. point fra to eller flere klasser.

De to klasser 1a og 1b har haft samme prøve hvor hver elev fik et antal point. Figuren viser fordelingen af point i de to klasser.



Af figuren ses:

1a har klaret sig bedre end 1b:

Både mindsteværdi, nedre kvartil, median, øvre kvartil og størsteværdi i 1a:

35, 52, 60, 66, 75

er større end den tilsvarende størrelse i 1b:

15, 27, 48, 57, 71.

Der gælder endda at mindsteværdien i 1a (35) er større end nedre kvartil i 1b (27) . Det betyder at

de mindste 25 % af pointtallene i 1b er mindre end alle pointtal i 1a.

Pointtallene ligger mindre spredt i 1a end i 1b:

Forskellen på højeste og laveste pointtal i 1a: $75-35=40$
er mindre end i 1b: $71-15=56$.

Forskellen på øvre og nedre kvartil i 1a: $66-52=14$
er mindre end i 1b: $57-27=30$.