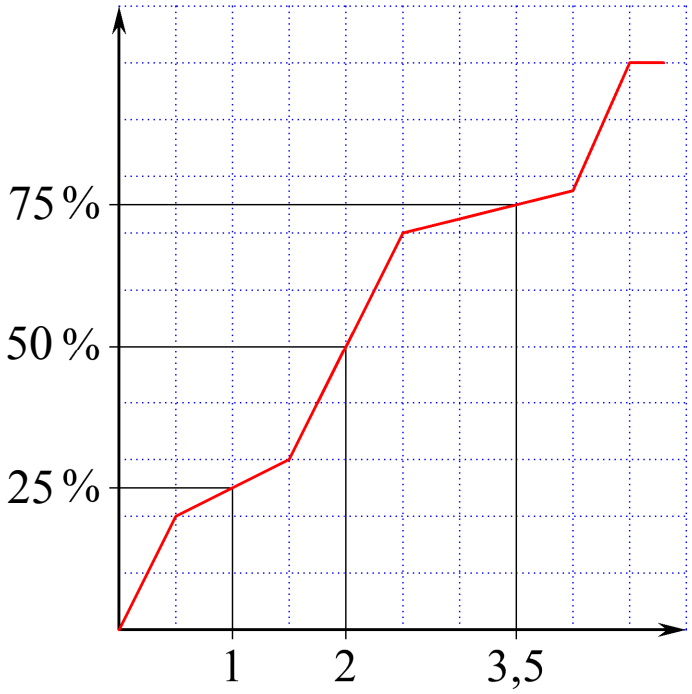
Deskriptiv statistik

for
gymnasiet og hf



2014 Karsten Juul

DESKRIPTIV STATISTIK

1.1 Hvad er deskriptiv statistik? 1
1.2 Hvad er grupperede og ugrupperede data? 1
1.21 Eksempel på ugrupperede data 1
1.22 Eksempel på grupperede data 1

2.1 Hvordan udregner vi middeltal (middelværdi) for ugrupperede data? 1
2.11 Hvordan udregner vi middeltallet når der er få data? 1
2.12 Hvordan udregner vi middeltallet når der er mange data? 1
2.2 Hvordan finder vi medianen for ugrupperede data? 2
2.21 Hvordan finder vi medianen når der er få data? 2
2.22 Hvordan finder vi medianen når der er mange data? 2
2.3 Hvordan finder vi kvartilsættet for ugrupperede data? 3
2.31 Hvis der er et midterste tal. 3
2.32 Hvis der ikke er et midterste tal 3
2.4 Hvordan tegner vi boksplot? 3
2.5 Hvordan sammenligner vi boksplot? 4
2.51 opgave. 4
2.52 opgave. 4
2.53 opgave. 4

3.1 Hvordan tegner vi et histogram? 5
3.2 Et grupperet datasæt er en model af virkeligheden der er meget forenklet. 5
3.3 Hvordan tegner vi en sumkurve? 6
3.31 Hvis der er oplyst procent for hvert interval 6
3.32 Hvis der er oplyst antal for hvert interval 6
3.4 Hvordan aflæser vi på en sumkurve? 7
3.41 Hvor mange procent af rørene er UNDER 3,7 meter? 7
3.42 Hvor mange procent af rørene er OVER 5,5 meter? 7
3.43 Hvor mange procent af rørene er MELLEM 3,7 og 5,5 meter? 7
3.44 Hvor mange procent af rørene er LIG 3,7 meter ELLER DERUNDER ? 7
3.5 Hvordan finder vi medianen for grupperede data? 8
3.6 Hvordan finder vi kvartilsættet for grupperede data? 8
3.61 Nedre kvartil. 8
3.62 Øvre kvartil. 8
3.63 Kvartilsæt. 8
3.7 Hvordan udregner vi middeltal (middelværdi) for grupperede data? 9

4.1 Hvordan grupperer vi data? 10
4.2 Hvor brede skal vi gøre intervallerne når vi grupperer data? 11
4.3 Problemer med intervallernes endepunkter når vi grupperer 12

5.1 Vi kan tegne histogrammer på to måder. 13
5.2 Hvor mange procent af dataene i et grupperet datasæt er lig et bestemt tal? 14
5.21 En vigtig egenskab ved en model af typen "grupperet datasæt". 14
5.22 Hvor mange procent af dataene er præcis lig 117? 14
5.23 Hvor mange procent af dataene er ca. 117? 14
5.24 Hvor mange procent af dataene er ca. 117,00? 14
5.3 Sumkurve og lineær sammenhæng. 15

Nyere hæfte:

http://mat1.dk/deskriptiv_statistik_for_c_niveau_i_hf.pdf

Deskriptiv statistik for gymnasiet og hf

19/3-2014

Nyeste version af dette hæfte kan downloades fra http://mat1.dk/noter.htm

DESKRIPTIV STATISTIK

1.1 Hvad er deskriptiv statistik?

Deskriptiv statistik er metoder til at få overblik over tal vi har indsamlet.
De tal vi har indsamlet, kalder vi data.

1.2 Hvad er grupperede og ugrupperede data?

Hvis der er mange forskellige data, så grupperer vi dem i intervaller.

1.21 Eksempel på ugrupperede data.

Vi har talt antallet af bær i 15 pakker.
Antal bær i en pakke: 24 24 22 24 23 22 24 23 26 26 23 28 27 22 24

1.22 Eksempel på grupperede data.

Vi har vejet 200 frugter:

Mellem 100 og 110 gram: 16 frugter
Mellem 110 og 120 gram: 68 frugter
Mellem 120 og 130 gram: 90 frugter
Mellem 130 og 140 gram: 26 frugter

2.1 Hvordan udregner vi middeltal (middelværdi) for ugrupperede data?

Middeltallet for nogle tal er det vi plejer at kalde gennemsnittet.
Vi kan udregne middeltallet (middelværdien) ved at lægge tallene sammen og dividere resultatet med antallet af tal.

2.11 Hvordan udregner vi middeltallet når der er få data?

I 7 prøver opnåede en elev følgende pointtal: 6 9 8 8 9 7 9

Sådan udregner vi middeltallet:

$$\frac{6+9+8+8+9+7+8}{7} = 7{,}85714$$

Middeltallet for elevens pointtal er $\underline{\underline{7{,}9}}$

2.12 Hvordan udregner vi middeltallet når der er mange data?

De nye elever på en skole har været til en prøve:

Point	1	2	3	4	5	6
Antal elever	5	22	58	49	62	18

I tabellen ser vi at 5 elever har fået 1 point, 22 elever har fået 2 point, osv.
Antallet af pointtal er altså

$$5+22+58+49+62+18 = 214$$

Vi behøver ikke lægge de 58 tretaller sammen. Vi får det samme ved at udregne $3 \cdot 58$.
Middeltallet kan vi altså udregne sådan:

$$\frac{1 \cdot 5 + 2 \cdot 22 + 3 \cdot 58 + 4 \cdot 49 + 5 \cdot 62 + 6 \cdot 18}{214} = 3{,}91121$$

Middeltallet for elevernes pointtal er altså $\underline{\underline{3{,}9}}$

2.2 Hvordan finder vi medianen for ugrupperede data?

(For grupperede data skal vi gøre noget helt andet. Se afsnit 3.5 på side 8).

2.21 Hvordan finder vi medianen når der er få data?

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

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:

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

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:

$\overbrace{3\ 3\ 4\ 5}\ \overbrace{6\ 6\ 8\ 9}$

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 .

2.22 Hvordan finder vi medianen når der er mange data?

De nye elever på en skole har været til en prøve:

Point	1	2	3	4	5	6
Antal elever	5	22	58	49	62	18

I tabellen ser vi at 5 elever har fået 1 point, 22 elever har fået 2 point, osv.

Antallet af pointtal er altså

$$5+22+58+49+62+18 = 214$$

Da $214:2=107$, ser det sådan ud:

$\overbrace{1\ 1\ \ldots\ ?}^{107}\ \overbrace{?\ \ldots\ 6\ 6}^{107}$

Tal nr. 6 i denne række er første total da der ifølge tabellen er 5 ettaller.
Tal nr. 28 er første tretal da $5+22=27$.
Tal nr. 86 er første firtal da $27+58=85$.
Tal nr. 135 er første femtal da $85+49=134$.
De to midterste tal, dvs. nr. 107 og 108, er altså begge firtaller.
Da der ikke er noget midterste tal, er medianen gennemsnittet af de to midterste tal.
Medianen for elevernes pointtal er altså $\underline{\underline{4{,}0}}$

I tabellen ovenfor ændrer vi antallet 22 til 23. Så ser det sådan ud.

$\overbrace{1\ 1\ \ldots\ ?}^{107}\ ?\ \overbrace{?\ \ldots\ 6\ 6}^{107}$

Vi kan se at nu er det tal nr. 108 der er medianen, dvs. medianen er 4.

2.3 Hvordan finder vi kvartilsættet for ugrupperede data?

(For grupperede data skal vi gøre noget helt andet. Se afsnit 3.6 på side 8).

2.31 Hvis der er et midterste tal:

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

Medianen for tallene til venstre for det midterste tal kalder vi nedre kvartil.
Dvs. nedre kvartil er 27.

Medianen for tallene til højre for det midterste tal kalder vi øvre kvartil.
Dvs. øvre kvartil er 57.

Når vi taler om kvartilsættet for nogle tal, så mener vi de tre tal

nedre kvartil , median og øvre kvartil,

dvs. kvartilsættet for tallene ovenfor er de tre tal 27 , 48 , 57 .

2.32 Hvis der ikke er et midterste tal:

3 3 4 5 6 6 8 9

Medianen for den venstre halvdel af tallene kalder vi nedre kvartil.
Dvs. nedre kvartil er 3,5 .

Medianen for højre halvdel af tallene kalder vi øvre kvartil.
Dvs. øvre kvartil er 7 .

Kvartilsættet er de tre tal 3,5 , 5,5 , 7,0 .

2.4 Hvordan tegner vi boksplot?

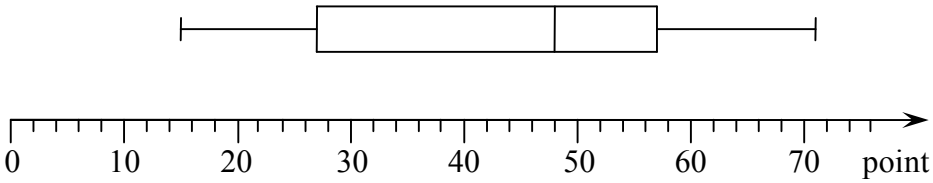
Ved at undersøge datasættet

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

kan vi se at

mindste tal	=	15
nedre kvartil	=	27
median	=	48
øvre kvartil	=	57
største tal	=	71

Disse oplysninger har vi vist på figuren. Sådan en figur kaldes et boksplot.



De to små lodrette streger i enderne viser at mindste og største tal er 15 og 71.
De to lodrette streger i hver ende af rektanglet viser at nedre og øvre kvartil er 27 og 57.
Den lodrette streg i midten af rektanglet viser at medianen er 48.

Rektanglet anskueliggør at den midterste halvdel af tallene ligger i intervallet fra 27 til 57.

Den vandrette streg til venstre anskueliggør at den fjerdedel af tallene der er mindst, ligger i intervallet fra 15 til 27.

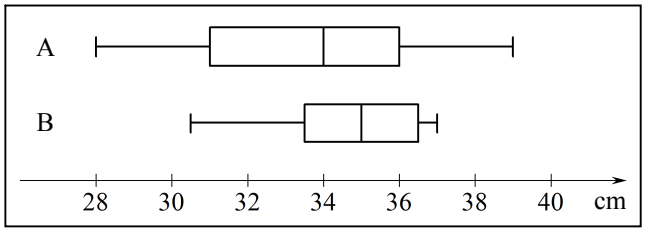
Den vandrette streg til højre anskueliggør at den fjerdedel af tallene der er størst, ligger i intervallet fra 57 til 71.

2.5 Hvordan sammenligner vi boksplot?

2.51 Opgave

Diagrammet viser højdefordelingen for en plante på to marker A og B.

Sammenlign højderne på A og B.



Svar

Sammenlign størrelser

Alle dele af diagrammet bortset fra højre endepunkt ligger længere mod højre på B, ← begrundelse
så højderne er altså overvejende større på B selv om den største højde er på A. ← resultat

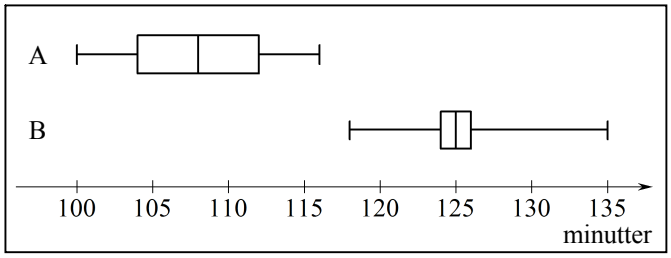
Sammenlign spredning

Både hele diagrammet og kassen er bredere på A's diagram end på B's, ← begrundelse
så højderne fra A er mere spredt end højderne fra B. ← resultat

2.52 Opgave

Diagrammet viser fordelingen af tider for to løbere A og B.

Sammenlign tiderne for A og B.



Svar

Sammenlign størrelser

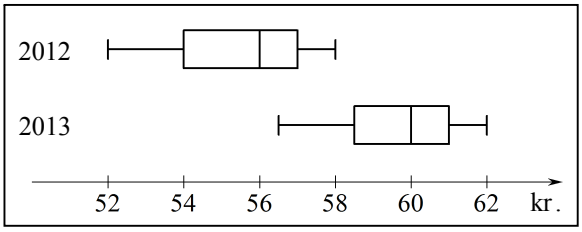
Venstre endepunkt for B-diagrammet ligger til højre for højre endepunkt for A-diagrammet, ← begrundelse
så B's mindste tid er større end A's største tid. ← resultat

Sammenlign spredning

A-kassen er meget længere end B-kassen, ← begrundelse
så midterste halvdel af tiderne er meget mere spredt for A end for B. ← resultat
Hele diagrammet har ca. samme længde for A og B, ← begrundelse
så forskellen på største og mindste tid er ca. den samme for A og B. ← resultat

2.53 Opgave

Diagrammet viser hvordan priserne på en vare er fordelt i 2012 og i 2013.



(a) Sammenlign priserne i 2012 og 2013.

(b) I 2012 betalte en person 53,50 kr. for varen.
Hvordan ligger denne pris i forhold til alle 2012-priserne for varen?

(c) En person betalte et beløb i den laveste halvdel af den højeste halvdel af 2012-priserne.
Hvad fortæller dette om størrelsen af beløbet.

Svar på (a)

Sammenlign størrelser

Hele 2013-diagrammet ligger til højre for venstre halvdel af 2012-diagrammet, ← begrundelse
så alle 2013-priserne er over laveste halvdel af 2012-priserne. ← resultat

Hele 2012-diagrammet ligger til venstre for kassen i 2013-diagrammet, ← begrundelse
så alle 2012-priserne er lavere end de 75 % højeste 2013-priser. ← resultat

Sammenlign spredning

Hverken for kassen eller hele diagrammet er længden ændret væsentligt fra 2012 til 2013, ← begrundelse
så der er ikke meget forskel på hvor spredt priserne er i 2012 og 2013. ← resultat

Svar på (b)

53,50 ligger på diagrammets venstre linjestykke, ← begrundelse
dvs. 53,50 kr. er i den nederste fjerdedel af 2012-priserne. ← resultat

Svar på (c)

Når et beløb er i den laveste halvdel af den højeste halvdel, er det i højre del af kassen, ← begrundelse
dvs. mellem 56,00 kr. og 57,00 kr. ← resultat

3.1 Hvordan tegner vi et histogram?

Tabellen viser fordelingen af nogle frugters vægt.

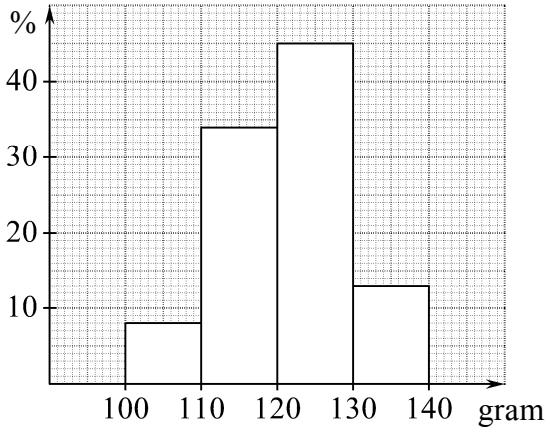
Vægt i gram	100-110	110-120	120-130	130-140
Procent	8	34	45	13

Histogrammet til højre viser oplysningerne i tabellen.

Rektanglet over intervallet 100-110 har højden højden 8 %.

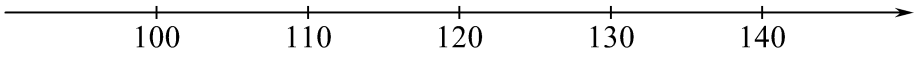
Dette viser at 8 % af frugterne vejer mellem 100 og 110 gram.

Bemærk: Denne måde at tegne et histogram på kan kun bruges fordi intervallerne 100-110, 110-120 osv. er lige lange. Til skriftlig eksamen skal du kun kende denne måde.
(Se evt. afsnit 5.1 side 13 hvor der står om en anden måde at tegne histogrammer på).

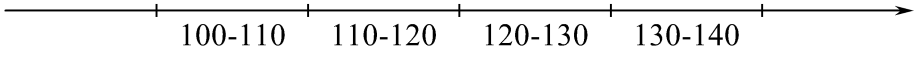


Advarsel: Den vandrette akse skal tegnes som en sædvanlig tallinje.

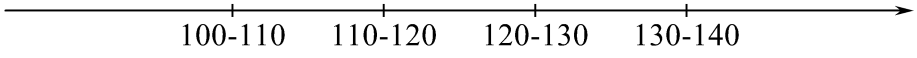
RIGTIGT:



FORKERT:



FORKERT:



3.2 Et grupperet datasæt er en model af virkeligheden der er meget forenklet.

Ovenfor har vi set på følgende grupperede datasæt:

Vægt i gram	100-110	110-120	120-130	130-140
Procent	8	34	45	13

Da dette datasæt er grupperet, skal vi regne som om

de 8 % i første interval er helt jævnt fordelt i dette interval

de 34 % er helt jævnt fordelt i andet interval

osv.

Dette betyder bl.a. et vi f.eks. skal regne som om

0 % af dataene er præcis lig 110, dvs. lig 110,00000…

(Se evt. afsnit 5.2 side 14 for at få en forklaring).

Der gælder altså:

Den procentdel af dataene der er 110 eller mindre,
er lig den procentdel der er mindre end 110.

Det giver ingen mening at spørge om 110 er talt med i intervallet 100-110 eller i intervallet 110-120.
Dette spørgsmål giver mening i andre opgaver (se afsnit 4.1 side 10 og evt. afsnit 4.3 side 12).

3.3 Hvordan tegner vi en sumkurve?

3.31 Hvis der er oplyst procent for hvert interval

Et intervals frekvens, er den procentdel af dataene som intervallet indeholder. Ordet "kumuleret" betyder ophobet.

For at tegne en sumkurve, udregner vi kumulerede frekvenser. Vi har skrevet dem i tabellen, og vi har udregnet dem sådan:

$8\% + 34\% = 42\%$, $8\% + 34\% + 45\% = 87\%$, osv.

Vægt i gram	100-110	110-120	120-130	130-140
Frekvens	8 %	34 %	45 %	13 %
Kumuleret frekvens	8 %	42 %	87 %	100 %

I andet interval står 42 %. Det betyder at i de to første intervaller er der 42 % af dataene, dvs. 42 % af dataene er under 120. Sumkurven skal bruges til at aflæse hvor mange procent af dataene der er mindre end et tal.

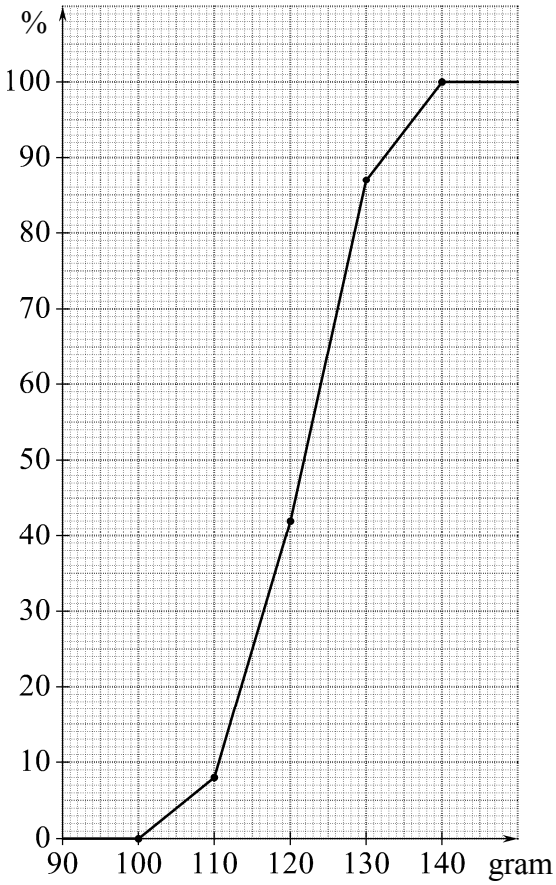
For at tegne sumkurven gør vi sådan:

0 % er mindre end 100, så ved $x = 100$ afsætter vi et punkt ud for 0 % på y-aksen.

8 % er mindre end 110, så ved $x = 110$ afsætter vi et punkt ud for 8 % på y-aksen.

42 % er mindre end 120, så ved $x = 120$ afsætter vi et punkt ud for 42 % på y-aksen.

Osv.



Da dataene er jævnt fordelt i hvert interval, skal vi forbinde punkterne med rette linjestykker.
(Se evt. begrundelsen for dette i afsnit 5.3 på side 15).

3.32 Hvis der er oplyst antal for hvert interval.

I tabellen står antal i stedet for procent.
Så må vi omregne til procent for at kunne tegne sumkurven.

Længde i meter	0,5-2	2-3	3-4	4-5	5-8
Antal rør	34	58	91	72	27

I tabellen nedenfor lægger vi sammen før vi omregner til procent. Det er for at undgå mellemfacitter med mange cifre.

I tabellen ovenfor kan vi skrive "hyppighed" i stedet for "antal rør". Det har vi gjort i tabellen nedenfor.

Antal data er $34 + 58 + +91 + 72 + 27 = 282$.

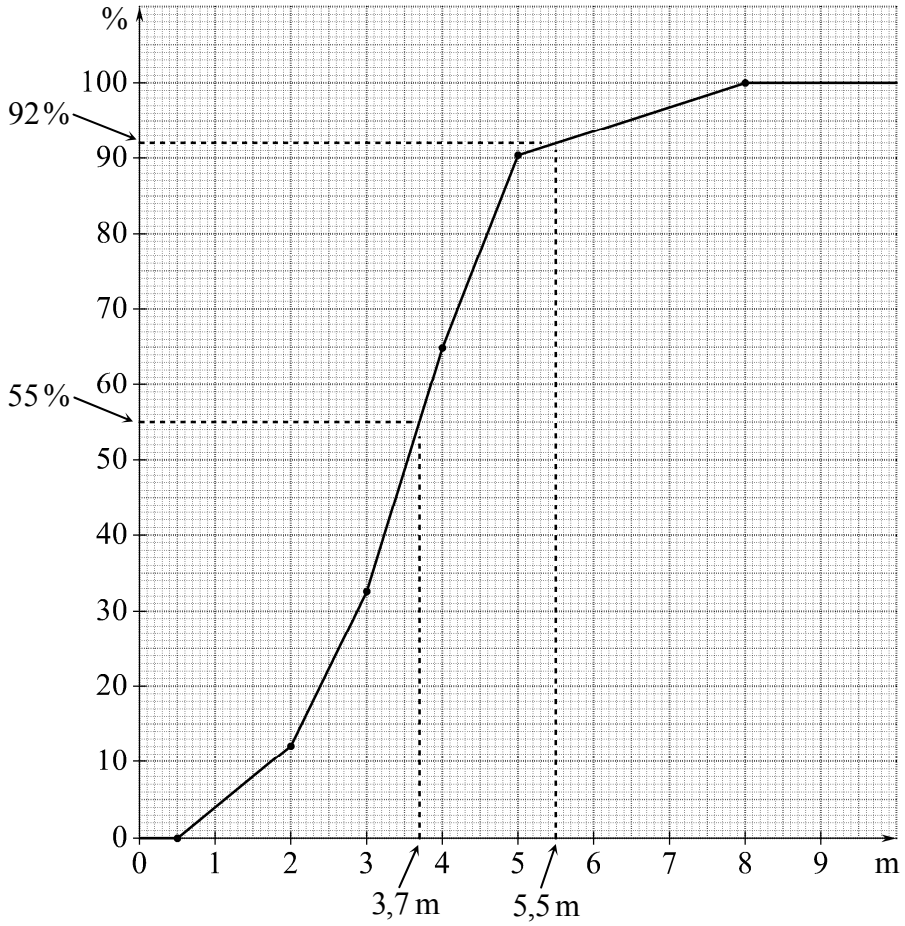
Tallene i 3. række udregner vi sådan: $34 + 58 = 92$, $34 + 58 + 91 = 183$, osv.

Tallene i 4. række udregner vi sådan: $\frac{34}{282} = 0{,}120567$, $\frac{92}{282} = 0{,}326241$, osv.

Længde i meter	0,5-2	2-3	3-4	4-5	5-8
Hyppighed	34	58	91	72	27
Kumuleret hyppighed	34	92	183	255	282
Kumuleret frekvens	12,1 %	32,6 %	64,9 %	90,4 %	100,0 %

3.4 Hvordan aflæser vi på en sumkurve?

Figuren viser sumkurven for rørene fra tabellen på foregående side.



3.41 Hvor mange procent af rørene er UNDER 3,7 meter?

Svar: Som vist på figuren aflæser vi at 55% af rørene er under 3,7 meter.

3.42 Hvor mange procent af rørene er OVER 5,5 meter?

Svar: Som vist på figuren aflæser vi at 92% af rørene er under 5,5 meter.
Da $100\% - 92\% = 8\%$, er 8% af rørene over 5,5 meter.

3.43 Hvor mange procent af rørene er MELLEM 3,7 og 5,5 meter?

Svar: Fra de 92% der er under 5,5 meter, skal fraregnes de 55% der er under 3,7 meter.
Da $92\% - 55\% = 37\%$, er 37% af rørene mellem 3,7 og 5,5 meter.

3.44 Hvor mange procent af rørene er LIG 3,7 meter ELLER DERUNDER?

Svar: Det er samme spørgsmål som spørgsmålet 3.41 ovenfor da
0% af rørene er præcis lig $3{,}70000\cdots$ meter.
Det at der på sumkurven er 0% der er lig 3,7 meter, er ikke i modstrid med at nogle af rørene er målt til 3,7 meter. (Læs evt. forklaringen på dette i afsnit 5.2 på side 14).

3.5 Hvordan finder vi medianen for grupperede data?

For at finde medianen skal vi bruge sumkurven når det er grupperede data.

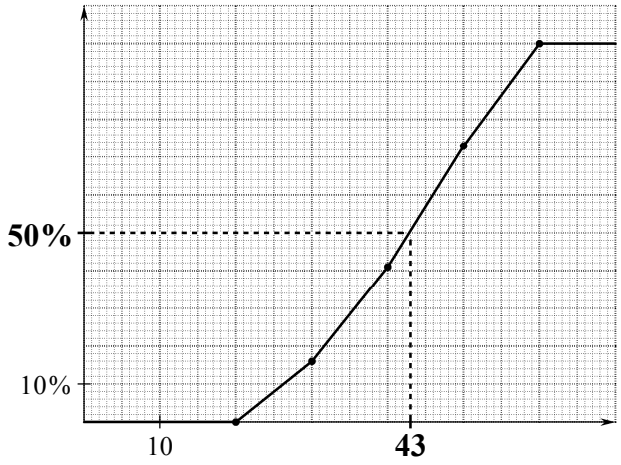
(For ugrupperede data skal vi gøre noget helt andet. Se afsnit 2.2 på side 2).

Vi starter i 50 % på y-aksen,
går vandret hen til sumkurven,
går lodret ned på x-aksen,
og aflæser x-værdien.

Denne x-værdi er medianen.

At et tal er median, betyder altså at
50 % af dataene er mindre end dette tal
og 50 % af dataene er større end dette tal.

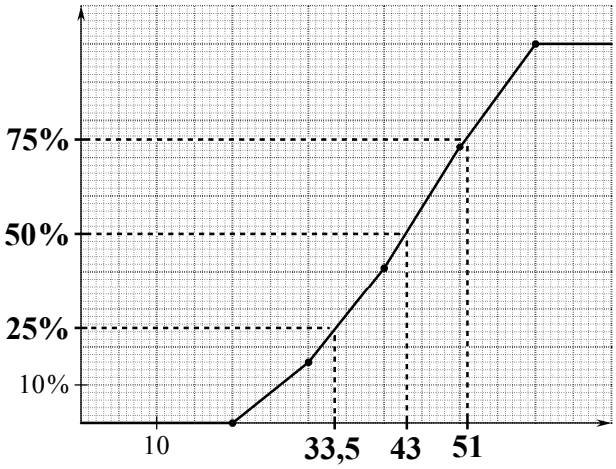
På figuren er medianen 43.



3.6 Hvordan finder vi kvartilsættet for grupperede data?

For at finde kvartilsættet skal vi bruge sumkurven når det er grupperede data.

(For ugrupperede data skal vi gøre noget helt andet. Se afsnit 2.3 på side 3).



3.61 Nedre kvartil.

Vi starter i 25 % på y-aksen,
går vandret hen til sumkurven,
går lodret ned på x-aksen,
og aflæser x-værdien.

Denne x-værdi er nedre kvartil.

At et tal er nedre kvartil, betyder altså at
25 % af dataene er mindre end dette tal
og 75 % af dataene er større end dette tal.

På figuren er nedre kvartil 33,5 .

3.62 Øvre kvartil.

Vi starter i 75 % på y-aksen,
går vandret hen til sumkurven,
går lodret ned på x-aksen,
og aflæser x-værdien.

Denne x-værdi er øvre kvartil.

At et tal er øvre kvartil, betyder altså at
75 % af dataene er mindre end dette tal
og 25 % af dataene er større end dette tal.

På figuren er øvre kvartil 51 .

3.63 Kvartilsæt.

Når vi taler om kvartilsættet for nogle tal, så mener vi de tre tal

nedre kvartil , median , øvre kvartil,

dvs. kvartilsættet er de tre tal 33,5 , 43 , 51 .

3.7 Hvordan udregner vi middeltal (middelværdi) for grupperede data?

Vi vil udregne middeltallet (middelværdien) for følgende grupperede datasæt:

Længde i meter	0,5-2	2-3	3-4	4-5	5-8
Antal rør	34	58	91	72	27

For at udregne middeltallet forestiller vi os at

de 34 tal i første interval alle er lig tallet i midten af dette interval,
de 58 tal i andet interval alle er lig tallet i midten af dette interval,
osv.

Dette ændrer ikke middeltallet da tallene er jævnt fordelt i hvert interval.

Tallet i midten af intervallet udregner vi sådan:

$\frac{0,5+2}{2}=1,25$, $\frac{2+3}{2}=2,5$, osv.

Tal i midten af intervallet	1,25	2,5	3,5	4,5	6,5
Hyppighed	34	58	91	72	27

Antal data er $34+58++91+72+27 = 282$.

Nu kan vi udregne middeltallet sådan (se afsnit 2.12 på side 1):

$$\frac{1,25\cdot 34+2,5\cdot 58+3,5\cdot 91+4,5\cdot 72+6,5\cdot 27}{282} = 3,56560$$

Middeltallet for rørenes længde er $\underline{\underline{3,57\,\text{cm}}}$.

4.1 Hvordan grupperer vi data?

Vi har modtaget et datasæt som består af 60 tal:

63	71	72	78	67	78	84	74	73	66
66	70	72	75	71	72	76	75	82	77
71	62	73	66	75	74	79	68	64	71
72	76	76	82	71	63	62	69	70	69
73	72	78	79	82	75	72	76	77	63
80	83	68	83	66	75	75	82	73	77

Disse tal er længder målt i mm.

For at få overblik over disse tal vil vi gruppere dem i følgende intervaller:

60-65 65-70 70-75 75-80 80-85

I rammen nedenfor har vi skrevet disse fem intervaller under hinanden.

Første tal i datasættet er 63. Derfor sætter vi en streg ud for 60-65.
Andet tal i datasættet er 71. Derfor sætter vi en streg ud for 70-75.
Osv.

Når vi i datasættet kommer til 70, sætter vi en streg ud for 65-70.
Når vi i datasættet kommer til 75, sætter vi en streg ud for 70-75.
Vi bruger altså følgende regel:

Et tal i datasættet der er lig et af intervalendepunkterne, tæller vi med i intervallet til venstre for tallet.

Bemærk: Dette er ikke den eneste måde at gøre det på, og det er ikke den mest nøjagtige måde, men der er tradition for at bruge denne måde i det danske gymnasium og hf. (Se evt. om andre måder i afsnit 4.3 på side 12).

60-65 𝍸 |
65-70 𝍸 𝍸 |
70-75 𝍸 𝍸 𝍸 𝍸 |||
75-80 𝍸 𝍸 |||
80-85 𝍸 ||

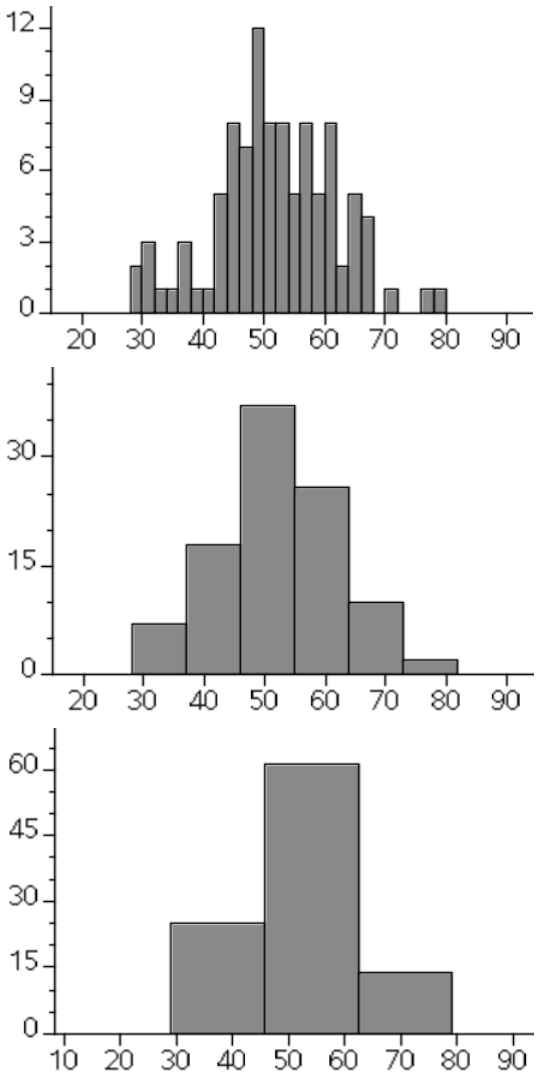
Efter at vi har foretaget denne optælling, kan vi opskrive det grupperede datasæt:

Længde i mm	60-65	65-70	70-75	75-80	80-85
Antal	6	11	23	13	7

4.2 Hvor brede skal vi gøre intervallerne når vi grupperer data?

På lommeregner eller computer kan vi nemt ændre intervallernes bredde og se hvordan histogrammet ændres.

Histogrammerne viser tre forskellige grupperinger af samme data.
På y-aksen står antal.



Øverste figur

Intervallernes bredde er for lille.
Der er så få data i hvert interval at højden svinger tilfældigt op og ned.

Midterste figur

Intervallernes bredde er passende.

Nederste figur

Intervallernes bredde er større end nødvendig, så vi får en unødig forenklet beskrivelse af hvordan dataene er fordelt.

4.3 Problemer med intervallernes endepunkter når vi grupperer.

I datasættet i 4.1 står tallet 75 seks steder. Det betyder ikke at seks af længderne er præcis 75,0000⋯ mm. Hvis længden f.eks. er ca. 75,4 mm vil måleresultatet være 75. Alle længder mellem ca. 74,5 mm og ca. 75,5 mm giver måleresultatet 75 mm.
De seks længder der er målt til 75 mm, har måske følgende længder:

ca. 75,3 ca. 74,9 ca. 74,9 ca. 74,5 ca. 75,1 ca. 75,4

Vi talte 75 med i intervallet 70-75, så alle seks længder ovenfor tæller altså med i intervallet 70-75 selv om tre af dem ikke ligger i dette interval.

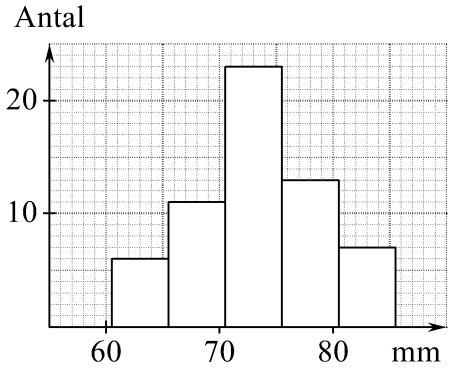
Dette problem kan vi undgå ved at bruge 75,5 som endepunkt i stedet for 75. Så bliver intervallernes endepunkter 60,5 , 65,5 , 70,5 osv.

Nedenfor er vist fire forskellige grupperinger af dataene fra 4.1 .

Intervallernes endepunkter ligger midt mellem to mulige "nabo-data".

TI-Nspire laver denne gruppering hvis vi taster bredde 5 og søjlestart 60,5.

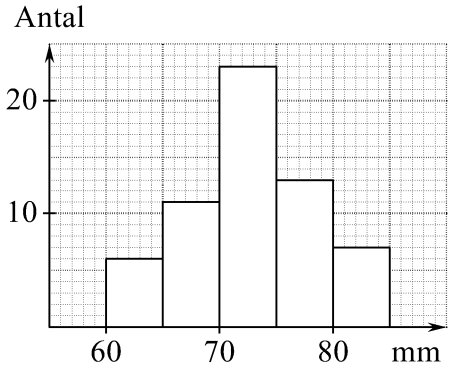
Her er der ingen data der er lig et endepunkt for et af intervallerne.



Alle data der er endepunkt for et af intervallerne, har vi talt med i intervallet til venstre for endepunktet.

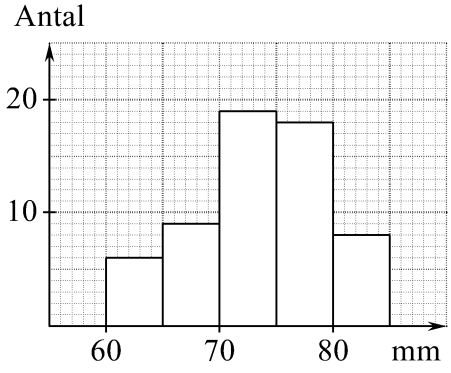
Rektanglerne har samme højder som på øverste figur, men de er anbragt en halv enhed længere mod venstre.

Dette er metoden som vi brugte i 4.1, og som der er en vis tradition for at bruge i det danske gymnasium og hf.

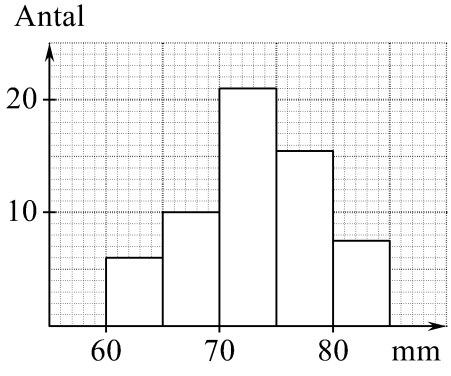


Alle data der er endepunkt for et af intervallerne, har vi talt med i intervallet til højre for endepunktet.

TI-Nspire laver denne gruppering hvis vi taster bredde 5 og søjlestart 60.



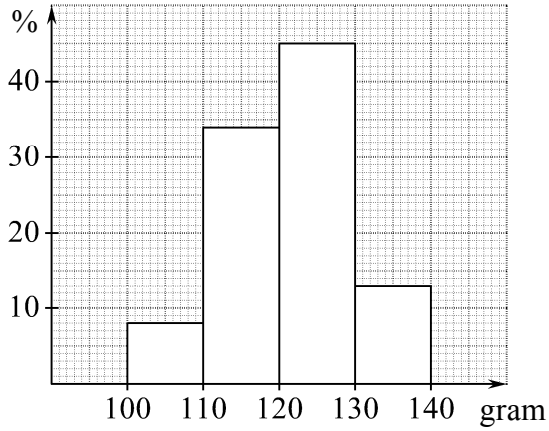
Alle data der er endepunkt for et af intervallerne, har vi talt med som en halv i hvert af de to intervaller med dette endepunkt.

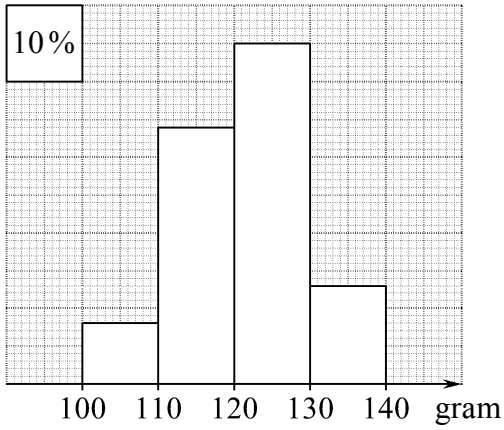


5.1 Vi kan tegne histogrammer på to måder.

På histogrammet til venstre kan vi aflæse frekvenserne på y-aksen.

På histogrammet til højre er det søjlernes areal der er frekvenserne.



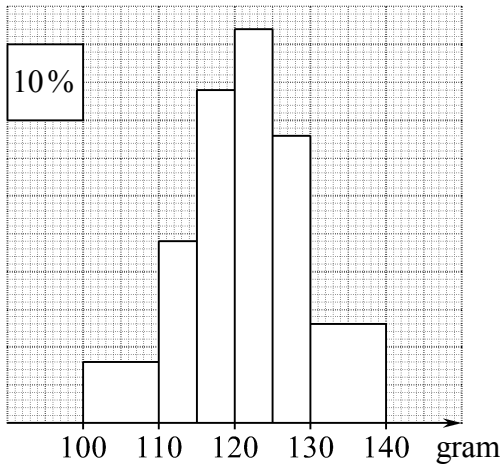


Når det er arealerne vi ser på, behøver intervallerne ikke være lige lange.

På figuren til højre er rektanglet over intervallet 110-115 tre gange så højt som rektanglet over intervallet 100-110.

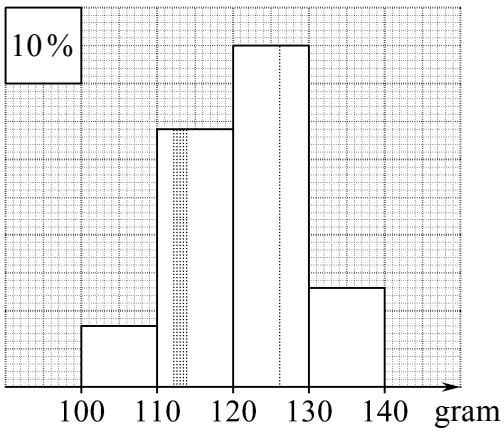
Men frekvensen er ikke tre gange så stor.

Frekvensen er kun 1,5 gange så stor
da arealet kun er 1,5 gange så stort.



På figuren har vi markeret arealet over intervallet 112-114. Arealet er $34\,\%:5=6{,}8\,\%$, så 6,8 % af dataene er mellem 112 og 114 gram.

På figuren har vi også markeret arealet over tallet 126. Arealet er 0 %, så 0 % af dataene er præcis $126{,}0000\cdots$ gram.



5.2 Hvor mange procent af dataene i et grupperet datasæt er lig et bestemt tal?

I 3.1-3.3 så vi på følgende grupperede datasæt:

Vægt i gram	100-110	110-120	120-130	130-140
Procent	8	34	45	13

5.21 En vigtig egenskab ved en model af typen "grupperet datasæt".

Vi ved ikke hvordan de oprindelige data var fordelt i det enkelte interval.
F.eks. ved vi ikke hvordan de 34 % var fordelt i intervallet 110-120.

Derfor har man vedtaget at man skal regne som om dataene i det enkelte interval er helt jævnt fordelt.

Dette bevirker at det grupperede datasæt på nogle punkter adskiller sig meget fra virkeligheden. Det grupperede datasæt er en model af virkeligheden

- der giver overblik over nogle hovedtræk,
- men ikke i detaljer svarer til virkeligheden.

5.22 Hvor mange procent af dataene er præcis lig 117?

I det interval som har længde 1 og hvis midtpunkt er 117, er 3,4 % af dataene.
Dette interval er nemlig en tiendedel af intervallet 110-120, som indeholder 34 % af dataene.

Ved at bruge samme metode kan vi udregne at

I intervallet med længde 0,01 og midtpunkt 117 er 0,034 % af dataene.

I intervallet med længde 0,0001 og midtpunkt 117 er 0,00034 % af dataene.

Osv.

Heraf slutter vi at 0 % af dataene er præcis lig 117,00000… . Dette fortæller ikke noget om virkeligheden, men vi skal bruge det når vi regner inden for modellen.

5.23 Hvor mange procent af dataene er ca. 117?

Hvis

tallet er ca. 117

betyder

tallet ligger mellem 116,5 og 117,5

så gælder at

3,4 % af dataene er ca. 117.

Hvis vi skriver målte længder som hele tal, så vil alle længder mellem 116,5 og 117,5 blive skrevet som 117.

5.24 Hvor mange procent af dataene er ca. 117,00?

Hvis

tallet er ca. 117,00

betyder

tallet ligger mellem 116,995 og 117,005

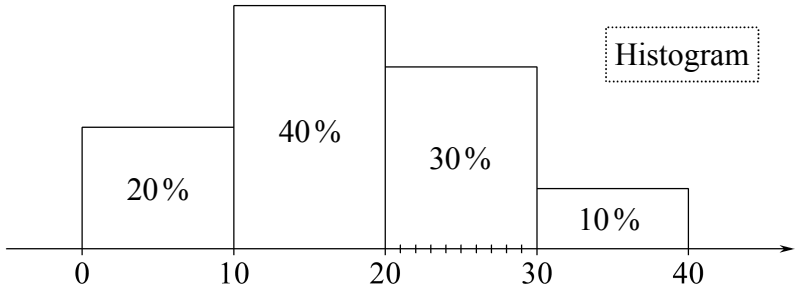
så gælder at

0,034 % af dataene er ca. 117,00.

Hvis vi skriver målte længder med to decimaler, så vil alle længder mellem 116,995 og 117,005 blive skrevet som 117,00.

5.3 Sumkurve og lineær sammenhæng.

Histogrammet viser et grupperet datasæt:



Intervallet 20-30 deler vi op i 10 lige store dele (se figur).
Hver af disse små intervaller må indeholde en tiendedel af hele intervallets observationer, dvs. de indeholder hver 3 % af samtlige observationer.

(x, y) er et punkt på sumkurven, dvs.

y er den procentdel af observationerne der har størrelse x eller derunder.

Af histogrammet ovenfor ser vi:

Når $x = 20$ er $y = 0{,}20 + 0{,}40 = 0{,}60$

Når $x = 21$ er $y = 0{,}60 + 0{,}03 = 0{,}63$

Når $x = 22$ er $y = 0{,}63 + 0{,}03 = 0{,}66$

Hver gang x bliver 1 større, vil y blive 0,03 enheder større,
så y vokser lineært i intervallet fra $x = 20$ til $x = 30$.
Derfor er grafen en ret linje i dette interval, og ligningen er

$y = 0{,}03x + b$.

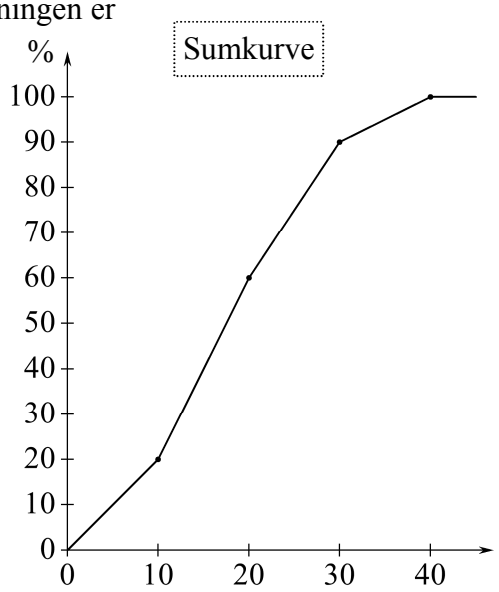
Vi udregner b :

Når $x = 20$ er $y = 0{,}60$ så

$0{,}60 = 0{,}03 \cdot 20 + b$.

Heraf ser vi at $b = 0$, så ligningen er

$y = 0{,}03x$.



For de fire intervaller er ligningerne:

0-10: $y = 0{,}02x$
10-20: $y = 0{,}04x - 0{,}2$
20-30: $y = 0{,}03x$
30-40: $y = 0{,}01x + 0{,}6$

Hvor mange procent af observationerne har størrelse 27 eller derunder?

Vi ser at vi skal bruge ligningen fra tredje interval:

$y = 0{,}03 \cdot 27 = 0{,}81$

dvs. 81 % af observationerne er 27 eller derunder.

Hvor stor er nedre kvartil?

Vi skal gå ud fra 25 % på y-aksen. Vi ser at vi skal bruge ligningen fra andet interval:

$0{,}25 = 0{,}04x - 0{,}2$.

Vi løser denne ligning mht. x og får 11,25 ,
dvs. nedre kvartil er 11,25 .

Stikordsregister

B

boksplot, sammenligne....4
boksplot, tegne....3

D

data....1
deskriptiv statistik....1

F

frekvens....6

G

grupperede data....1, 14
grupperede data afviger fra virkeligheden....5
gruppering af data....10, 11

H

histogram....5, 11, 12, 13, 15
hyppighed....6

I

intervallers bredde....11
intervallers endepunkter....10, 12
intervals frekvens....6

K

kumuleret frekvens....6
kumuleret hyppighed....6
kvartilsæt for grupperede data....8
kvartilsæt for ugrupperede data....3

M

median for grupperede data....8
median for ugrupperede data....2
middeltal for grupperede data....9
middeltal for ugrupperede data....1
middelværdi for grupperede data....9
middelværdi for ugrupperede data....1

N

nedre kvartil for grupperede data....8
nedre kvartil for ugrupperede data....3

S

sumkurve og lineær sammenhæng....15
sumkurve, aflæse....7, 8
sumkurve, antal oplyst....6
sumkurve, procent oplyst....6
sumkurve, tegne....6

U

ugrupperede data....1

Ø

øvre kvartil for grupperede data....8
øvre kvartil for ugrupperede data....3