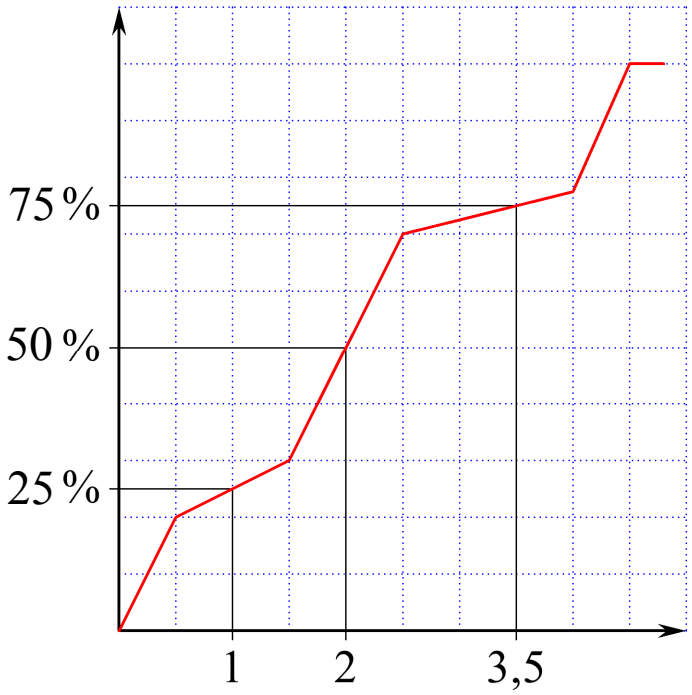
Statistik

for
gymnasiet og hf



2013 Karsten Juul

I dette hæfte er der lagt vægt på

- at det skal være egnet til at slå op i når elever løser opgaver
- at tvivlstilfælde bliver afklaret
- at det er muligt på forskellige niveauer at inddrage andre emner i et eksamensspørgsmål om statistik.

Udgaven fra 2011 har skiftet adresse til http://mat1.dk/statistik_for_gymnasiet_og_hf_2011.pdf

Gå ind på http://mat1.dk/noter.htm for at downloade nyeste version af dette hæfte.

Statistik for gymnasiet og hf

Dette hæfte kan downloades fra www.mat1.dk

DESKRIPTIV STATISTIK

TEST

FORDELINGER

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?

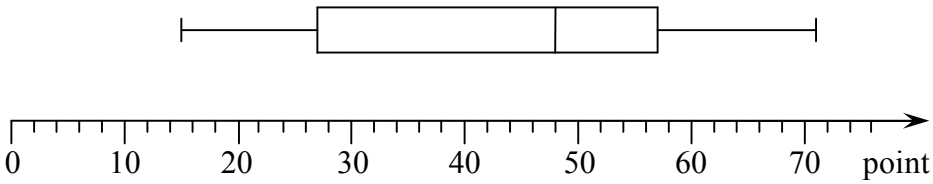
Ved at undersøge datasættet

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.

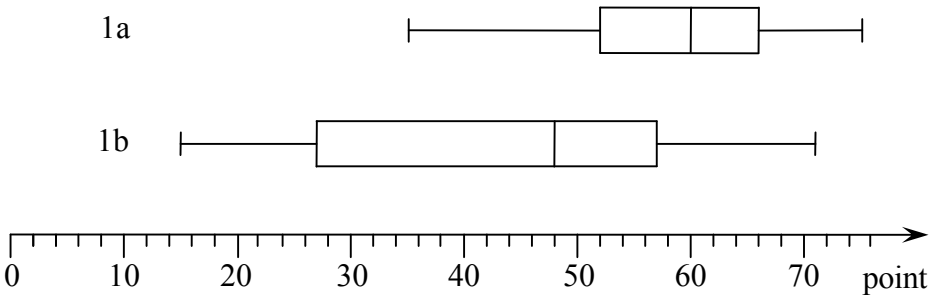
Den vandrette streg til venstre anskueliggør at den fjerdedel af tallene der er mindst, ligger i intervallet fra 15 til 27.

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?

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.



Eksempel på hvad vi bl.a. kan skrive når vi skal sammenligne to boksplot:

> Nedre kvartil for 1a's pointtal (pt) er 52, så mindst 75 % af 1a's pt er 52 eller større, og medianen for 1b's pt er 48, så mindst 50 % af 1b's pt er 48 eller mindre.
> Altså er de dårligste 50 % i 1b dårligere end de 75 % bedste i 1a.

Bemærk at vi både skriver fagudtrykket (f.eks. median) og talværdien og hvad tallet fortæller.

På figuren kan vi se:

1a har klaret sig bedre end 1b
da alle dele af diagrammet ligger tydeligt længere mod højre i 1a end i 1b:

> Mindste tal, nedre kvartil, median, øvre kvartil og største tal i 1a
> (som er 35, 52, 60, 66, 75)
> er større end de tilsvarende tal i 1b
> (som er 15, 27, 48, 57, 71).
>
> Der gælder endda at mindste tal i 1a (som er 35) er større end nedre kvartil i 1b (som er 27) . Det betyder at
>
> de mindste 25 % af pointtallene i 1b er mindre end det mindste pointtal i 1a.

Pointtallene ligger mindre spredt i 1a end i 1b
da både kassen og hele diagrammet er tydeligt bredere i 1b end i 1a:

> Forskellen på højeste og laveste pointtal i 1a (som er $75 - 35 = 40$)
> er mindre end i 1b (hvor den er $71 - 15 = 56$).
>
> Forskellen på øvre og nedre kvartil i 1a (som er $66 - 52 = 14$)
> er mindre end i 1b (hvor den er $57 - 27 = 30$).

3.1 Hvordan tegner vi et histogram?

Tabellen viser fordelingen af nogle frugters vægt.

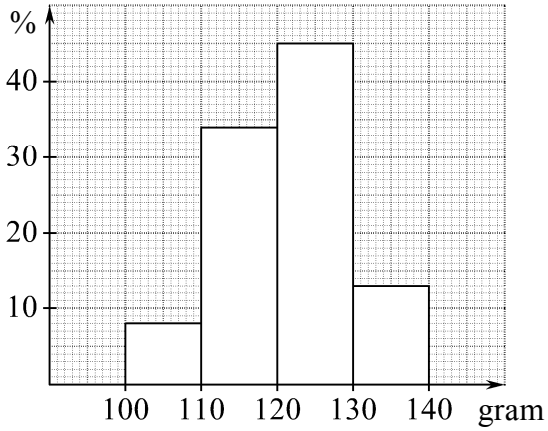
Vægt i gram	100-110	110-120	120-130	130-140
Procent	8	34	45	13

Histogrammet til højre viser oplysningerne i tabellen.

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

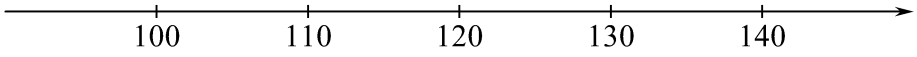
Dette viser at 8 % af frugterne vejer mellem 100 og 110 gram.

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).

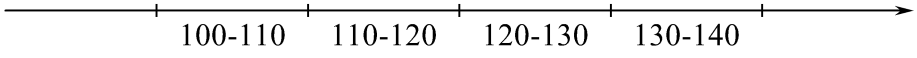


Advarsel: Den vandrette akse skal tegnes som en sædvanlig tallinje.

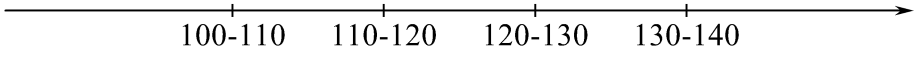
RIGTIGT:



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 $110{,}00000\cdots$

(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.

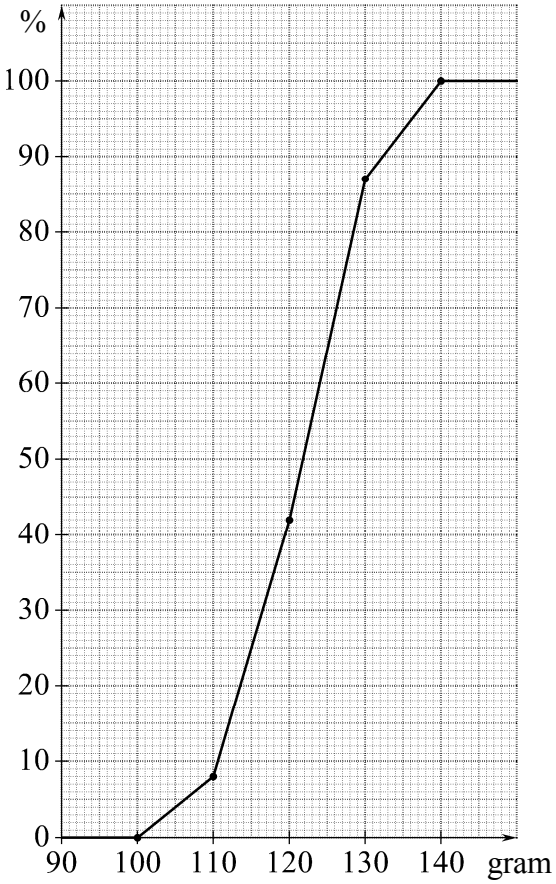
For at tegne sumkurven gør vi sådan:

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 $34 + 58 + +91 + 72 + 27 = 282$.

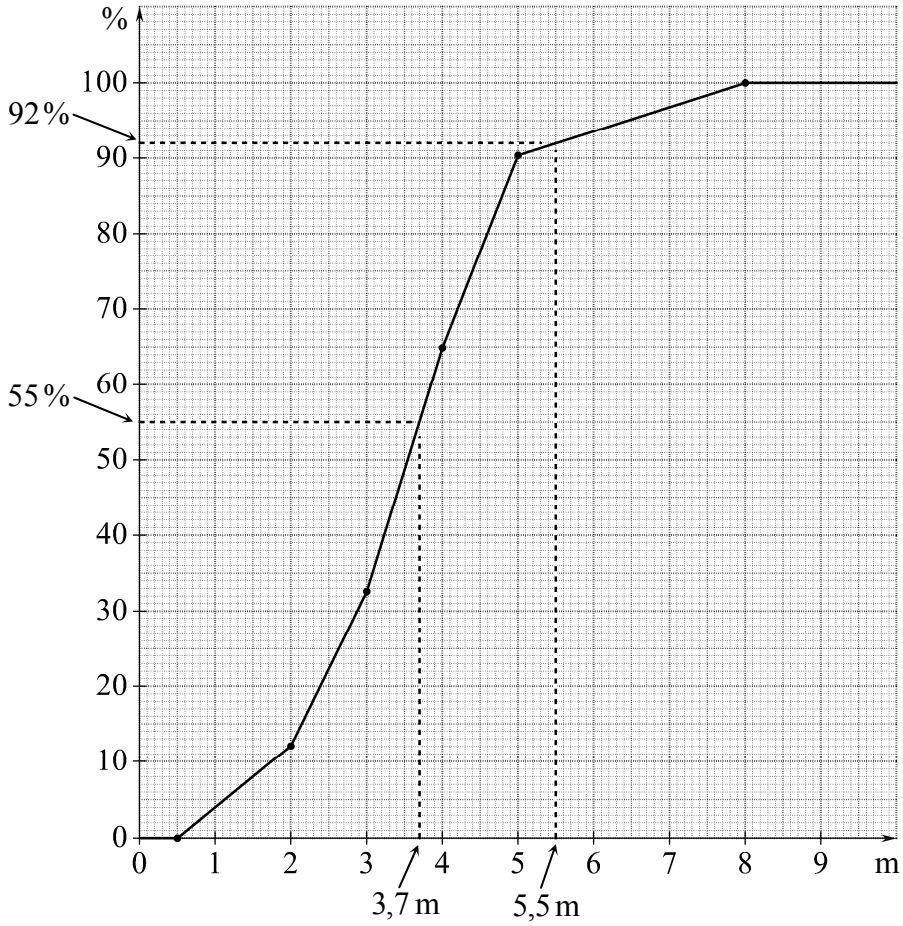
Tallene i 3. række udregner vi sådan: $34 + 58 = 92$, $34 + 58 + 91 = 183$, osv.

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.

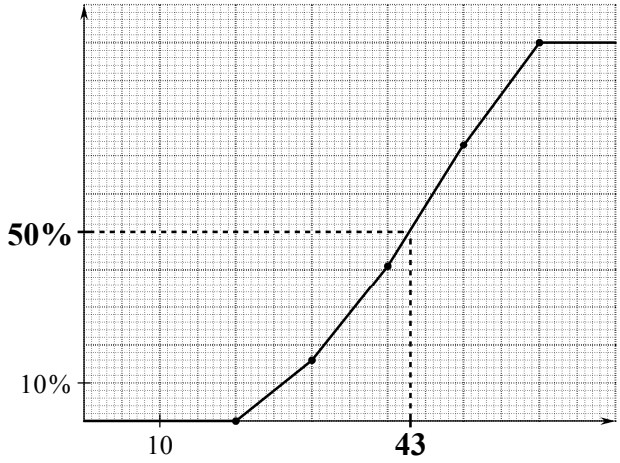
(For ugrupperede data skal vi gøre noget helt andet. Se afsnit 2.2 på side 2).

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.

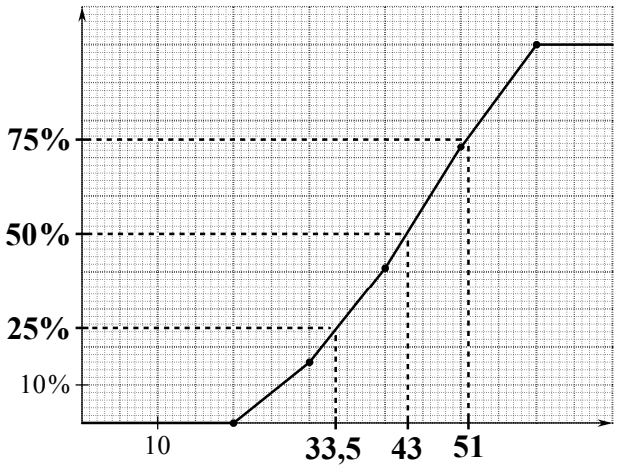
På figuren er medianen 43.



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.

<u>Bemærk:</u> 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 ||

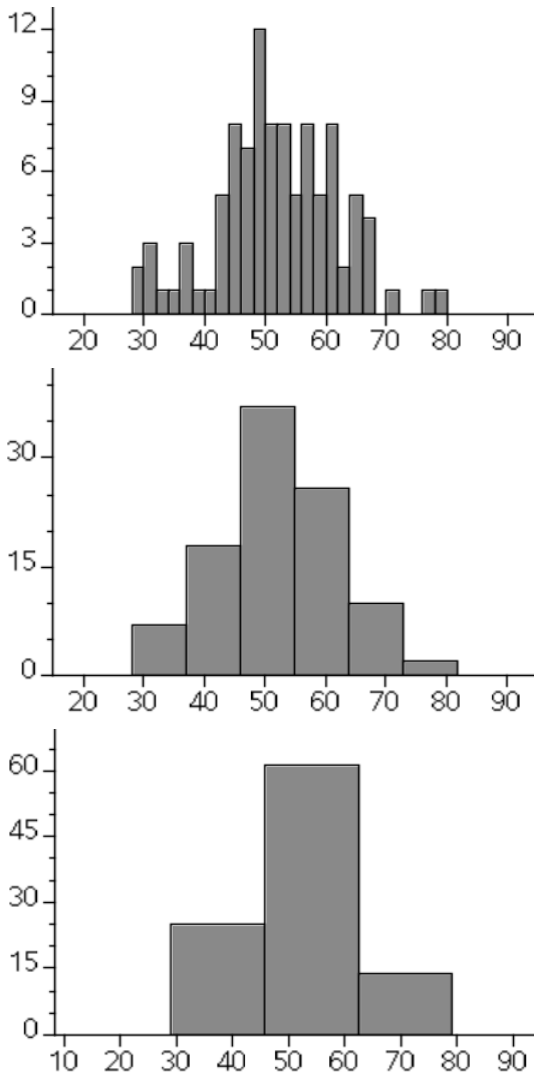
Efter at vi har foretaget denne optælling, kan vi opskrive det grupperede datasæt:

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.

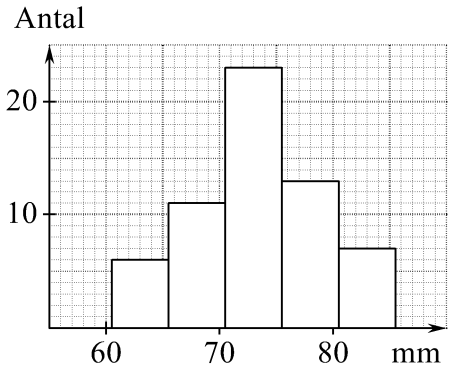
Dette problem kan vi undgå ved at bruge 75,5 som endepunkt i stedet for 75. Så bliver intervallernes endepunkter 60,5 , 65,5 , 70,5 osv.

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.

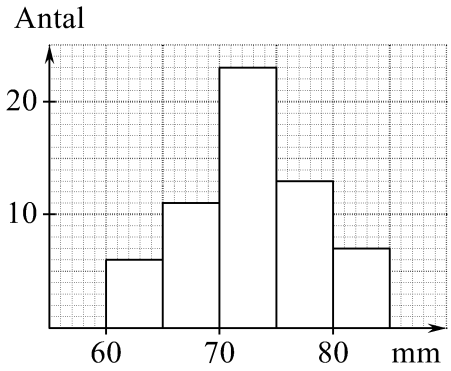
Her er der ingen data der er lig et endepunkt for et af intervallerne.



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

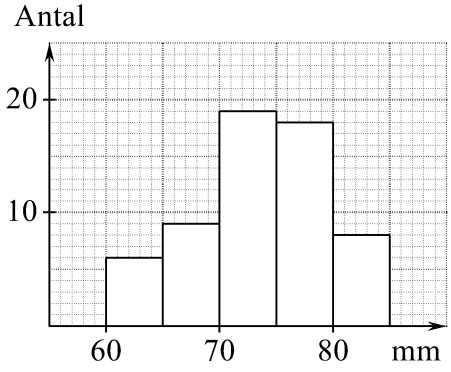
Rektanglerne har samme højder som på øverste figur, men de er anbragt en halv enhed længere mod venstre.

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.

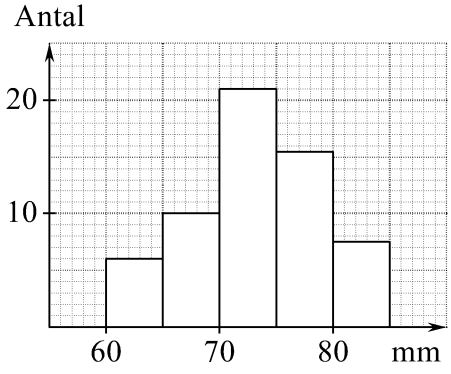


Alle data der er endepunkt for et af intervallerne, har vi talt med i intervallet til højre for endepunktet.

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



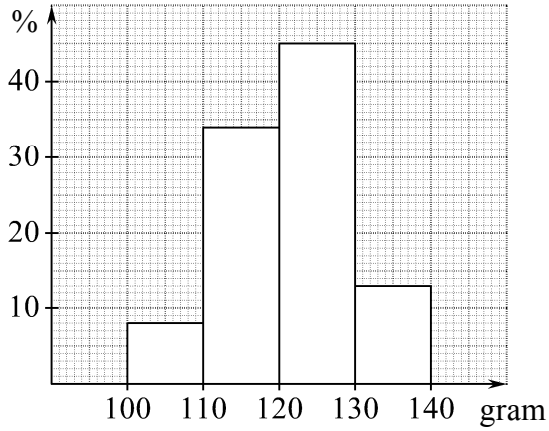
Alle data der er endepunkt for et af intervallerne, har vi talt med som en halv i hvert af de to intervaller med dette endepunkt.

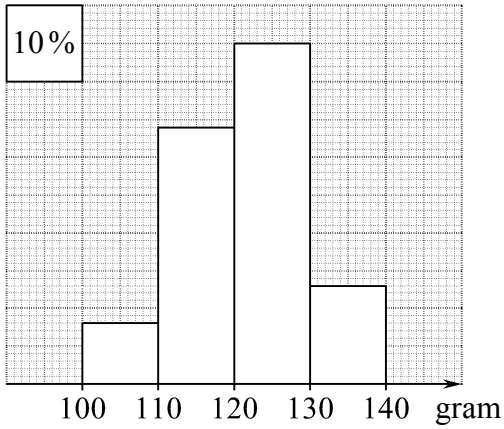


5.1 Vi kan tegne histogrammer på to måder.

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

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



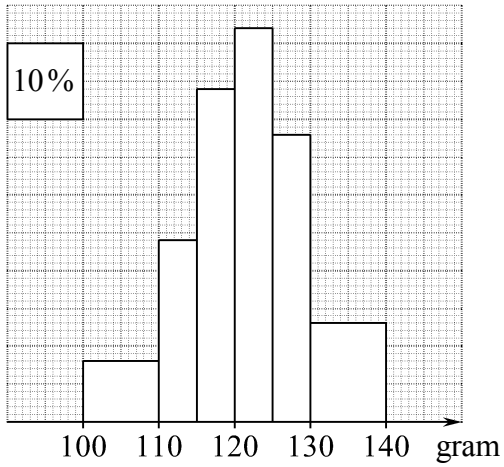


Når det er arealerne vi ser på, behøver intervallerne ikke være lige lange.

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

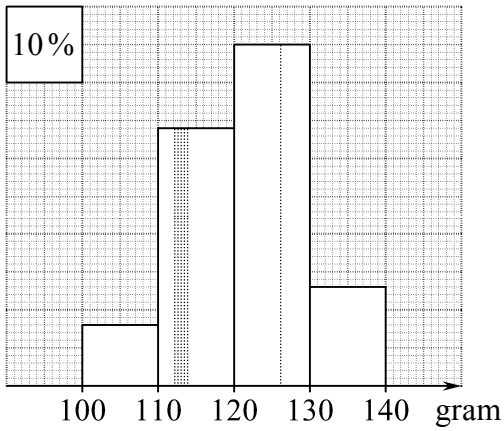
Men frekvensen er ikke tre gange så stor.

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

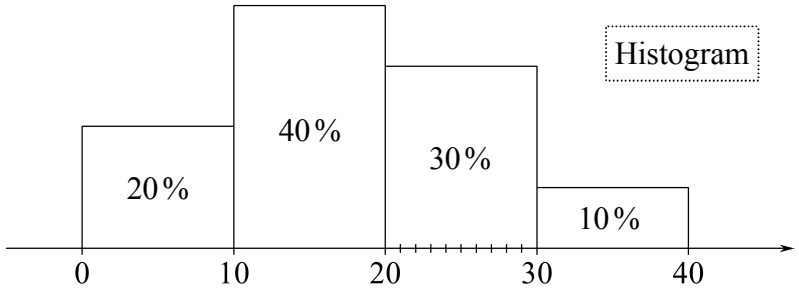
så gælder at

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

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

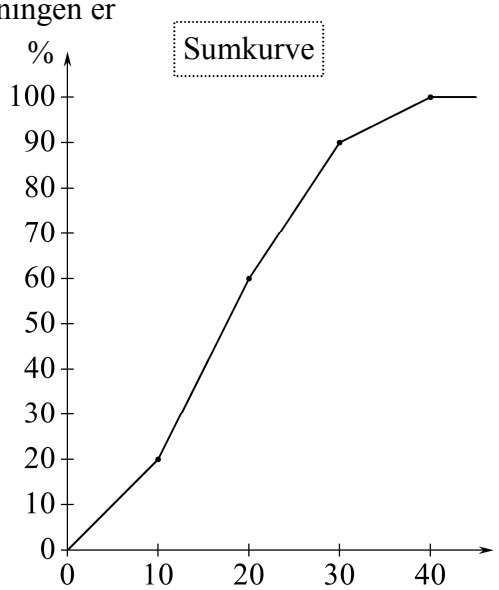
Vi udregner b :

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 .

TEST

6 Stikprøver.

Nogen på et gymnasium mener at der er forskel på hvad piger og drenge mener om et bestemt spørgsmål. For at undesøge denne hypotese, spørger vi nogle piger og drenge.

6.1 Hvad er populationen?

De ting eller personer som vi vil påstå noget om, kaldes populationen.

Er det alle personer i europa som nu er mellem 10 og 20 år?
Er det alle elever på vores gymnasium?
Eller?

Når vi laver en statistisk undersøgelse, skal vi **skrive** en præcisering af

hvad det er for en population vi vil påstå noget om.

6.2 Hvad er stikprøven?

Vi undersøger kun en lille del af hele populationen.
De personer vi får et svar fra (eller de ting vi undersøger), kaldes stikprøven.

Når vi laver en statistisk undersøgelse, skal vi **skrive** en præcisering af

hvordan vi har valgt stikprøven.

Det er **IKKE nok at skrive**:

"Vi har spurgt 47 elever på vores gymnasium."

Det er **nok at skrive**

"Den 20. februar mellem kl. 8:50 og 9:10 spurgte vi de 47 elever der sad på gangen, og vi fik svar fra dem alle. 10 af drengene og 8 af pigerne var fra 3g FY, 13 af drengene og 16 af pigerne var fra 3g Fy."

eller

"Den 20. februar kl. 8:50 sendte vi en besked til alle elever på skolen. Stikprøven er de 47 elever der svarede inden kl. 10:00 den 22. februar."

Disse to beskrivelser af en indsamling af stikprøve er så grundige at læseren kan se om der er grund til tro at der kan være systematiske fejl.

6.3 Systematiske fejl ved valg af stikprøven.

Eksempel 1

Population: Eleverne på vores gymnasium.

Stikprøve: Eleverne i en sproglig klase.

Her kan vi have lavet en systematisk fejl ved valg af stikprøven, for det kan være at en bestemt holdning oftere er blandt sproglige end blandt andre.

Eksempel 2

Hvis vi spørger elever pr. e-mail, og mange ikke svarer, så kan vi have lavet en systematisk fejl, for det er måske især elever med en bestemt holdning der svarer.

6.4 Tilfældige fejl ved valg af stikprøven.

Selv om vi vælger stikprøven tilfældigt blandt hele populationen,
er det ikke helt sikkert at den ligner populationen.

Det kan f.eks. være at vi tilfældigt har fået for mange ja-sigere med i stikprøven.

Det er muligheden for tilfældige fejl vi beskæftiger os med når vi udregner tallet *p*. (se afsnit 9.3 og 13.4).

Afsnit 6 fortsætter på næste side.

6.5 Er der skjulte variable?

En skjult variabel er noget der kan ødelægge resultatet selv om stikprøven er udvalgt tilfældigt blandt hele populationen.

Eksempel 1

Mange elever svarer på noget andet end det de er blevet spurgt om, fordi de tror de bliver spurgt om en aktuel sag. Elevens holdning til denne sag er en skjult variabel der påvirker elevens svar.

Eksempel 2

Der er flere der overlever på hospital A end på hospital B. Man slutter at behandlingen er bedre på A end på B. Men forskellen skyldes at B har flere ældre patienter. Patienternes alder er en skjult variabel der påvirker resultatet.

7 Hvad er sandsynlighed?

7.1 Eksempel.

At **sandsynligheden for at vinde = 25 %**
betyder at **vi vinder 25 % af gangene**.

7.2 Eksempel.

At **sandsynligheden for at en pose har 3 eller flere defekte = 4 %**
betyder at **4 % af poserne har 3 eller flere defekte**.

8 Test af hypotese.

8.1 Signifikansniveau.

Hypotese: *Halvdelen af dåserne er bulede*.

Vi undersøger 10 tilfældige dåser: kun 1 er bulet.
Hvis halvdelen af alle dåser var bulede, ville der nok have været mere end 1 bulet i en stikprøve på 10 dåser.
Derfor forkaster vi hypotesen.

Sandsynligheden p for at få 1 eller færre bulede er ca. 1 % hvis hypotesen er rigtig.
Hvor lille skal p være for at vi forkaster hypotesen?
Dette fastsætter vi ved at angive en sandsynlighed som vi kalder signifikansniveauet.
Ofte er signifikansniveauet 5 % eller 1 %.

Vi forkaster hypotesen hvis p er mindre end signifikansniveauet.
p er sandsynligheden for at få en stikprøve der ligger i hvert fald så langt fra hypotesen som vores stikprøve gør.

8.2 Hvornår har vi vist noget med en test?

Kun når vi forkaster en hypotese, har vi vist noget.
Når vi ikke forkaster hypotesen, har vi IKKE vist at hypotesen sandsynligvis er rigtig.

Mange (også lærere) tror at når vi ikke kan forkaste hypotesen, så er det sandsynligt at hypotesen er rigtig. Dette er en katastrofal misforståelse. Det har ingen forbindelse med virkeligheden.
I nogle eksamensopgaver og vejledende opgaver spørges om der er belæg for uafhængighed. At spørge sådan er en grov fejl da testen aldrig kan give belæg for uafhængighed.

9. Test for uafhængighed i 2×2 tabel.

Vi vil teste om piger og drenge har samme holdning til et spørgsmål.
Populationen er alle danskere hvis alder er 16 til 18 år.

Hypotese: *Andelen der siger ja, er ens for piger og drenge.*

Hypotesen er altså at svaret er uafhængigt af om det er en pige eller en dreng.
I den slags test vi laver her, gælder altid: Hypotesen er at noget er ens.

Vi vælger: **signifikansniveau = 5%**

Nogle tilfældigt udvalgte piger og drenge får stillet samme spørgsmål.
Stikprøven er de piger og drenge som svarede.
De svarede sådan:

Faktiske tal	ja	nej
piger	**87**	**46**
drenge	**71**	**21**

9.1 Sådan udregner vi FORVENTEDE TAL.

For at teste hypotesen, udregner vi først noget vi kalder de forventede tal, dvs. hvordan svarene skulle være fordelt mellem de fire felter hvis ja-andelen i tabellen skulle være ens for piger og drenge.

For at kunne udregne de forventede tal udregner vi først følgende tal:

Antal piger: $87+46 = 133$
Antal drenge: $71+21 = 92$
Antal ja-sigere: $87+71 = 158$
Antal nej-sigere: $46+21 = 67$
Antal piger og drenge: $133+92 = 225$

Disse tal skriver vi i et skema:

	ja	nej	i alt
piger			133
drenge			92
i alt	158	67	225

I tabellen med forventede tal skal andelen af ja-sigere være den samme for piger og drenge, så da 158 af de 225 elever svarede ja, skal vi i denne tabel skrive at $\frac{158}{225}$ af de 133 piger svarede ja:

forventet antal ja-svar fra piger: $133 \cdot \frac{158}{225} = 93{,}40$
forventet antal nej-svar fra piger: $133-93{,}40 = 39{,}60$
forventet antal ja-svar fra drenge: $158-93{,}40 = 64{,}60$
forventet antal nej-svar fra drenge: $92-64{,}60 = 27{,}40$

Her er de fire forventede tal skrevet ind i skemaet:

Forventet	ja	nej	i alt
piger	**93,40**	**39,60**	133
drenge	**64,60**	**27,40**	92
i alt	158	67	225

9.2 Sådan udregner vi χ^2.

Symbolet χ^2 læses sådan: *ki i anden*

For hvert af de fire felter udregner vi

$$\frac{(faktisk - forventet)^2}{forventet}$$

Ved at lægge disse fire tal sammen får vi tallet χ^2 som er afstanden mellem de faktiske tal og de forventede tal:

$$\chi^2 = \frac{(87-93{,}40)^2}{93{,}40} + \frac{(46-39{,}60)^2}{39{,}60} + \frac{(71-64{,}60)^2}{64{,}60} + \frac{(21-27{,}40)^2}{27{,}40}$$

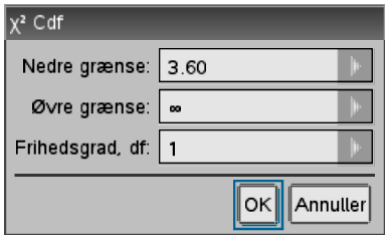
$$\chi^2 = 3{,}60$$

9.3 Sådan udregner vi *p*.

Ovenfor udregnede vi at afstanden mellem faktiske og forventede tal er $\chi^2 = 3{,}60$.

Tallet p er sandsynligheden hvis hypotesen er rigtig,
for at afstanden χ^2 er 3,60 eller større (dvs. mellem 3,60 og ∞).

Vi kan få Nspire til at udregne p ved i beregningsmenuen at vælge
Statistik / Fordelinger / χ^2 Cdf... og udfylde sådan:



Tallet 1 er antal frihedsgrader.
Se afsnit 10.

På Nspire-lommeregneren kan vi skrive ∞ ved hjælp af π-tasten.
På computeren skriver vi infinity i stedet for ∞.

$p = \chi^2\text{Cdf}(3.60, \infty, 1) = 5{,}8\%$ *udregnet på Nspire*

9.4 Sådan skriver vi konklusionen.

Da sigifikansniveauet er 5 %, og p ikke er mindre end 5 %, kan vi ikke forkaste hypotesen.

Stikprøven giver ikke belæg for at hævde at andelen der siger ja,

er forskellig for piger og drenge.

9.5 Misforstå ikke procenterne.

5,8 % er IKKE sandsynligheden for at hypotesen er rigtig.
5,8 % er IKKE sandsynligheden for at hypotesen er forkert.
94,2 % er IKKE sandsynligheden for at hypotesen er rigtig.
94,2 % er IKKE sandsynligheden for at hypotesen er forkert.

De 5,8 % er udregnet under den forudsætning at hypotesen er rigtig
og er sandsynligheden for at få en stikprøve hvis afvigelse fra hypotesen
er så stor som eller større end afvigelsen i den stikprøve vi fik.

10 Hvordan udregner vi antal FRIHEDSGRADER i test for uafhængighed?

10.1 Frihedsgrader for 2 gange 2 tabel.

	ja	nej	i alt
piger	**87**		133
drenge			92
i alt	158	67	225

Tabellen stammer fra en undersøgelse af om der er forskel på drenges og pigers holdning til det stillede spørgsmål.

I tabellen er der

2 rækker (piger og drenge)

og

2 søjler (ja og nej).

Hvis vi skriver et tal i ét af de fire felter, så er det fastlagt hvad der skal stå i de tre andre felter. Det er fordi både antal piger, antal drenge, antal ja og antal nej er kendt.

Fordi man kun kan vælge 1 tal, er antallet af frihedsgrader 1.

10.2 Frihedsgrader for 2 gange 3 tabel.

	ja	nej	?	i alt
piger	**35**	**12**		61
drenge				45
i alt	58	23	25	106

Tabellen stammer fra en undersøgelse af om der er forskel på drenges og pigers holdning til det stillede spørgsmål.

I tabellen er der

2 rækker (piger og drenge)

og

3 søjler (ja, nej og ?).

Hvis vi skriver tal i to af de seks felter, så er det fastlagt hvad der skal stå i de fire andre felter. Det er fordi både antal piger, antal drenge, antal ?, antal ja og antal nej er kendt.

Fordi man kun kan vælge 2 tal, er antallet af frihedsgrader 2.

10.3 Frihedsgrader for *m* gange *n* tabel.

Hvis der i test for uafhængighed er

m rækker og n søjler,

så er

antal frihedsgrader $= (m-1)\cdot(n-1)$

Hvis $m=2$ og $n=2$, så får vi af denne formel at

antal frihedsgrader $= (2-1)\cdot(2-1) = 1\cdot 1 = 1$

ADVARSEL: Nogle gange skal vi udregne antal frihedsgrader på en anden måde.
Se afsnit 13.3 .

11 Eksempel med to frihedsgrader i test for uafhængighed.

Populationen er alle danske elever i 1.g som har matematik på A- eller B-niveau. Vi vil teste følgende hypotese på 5 %-signifikansniveau:

Hypotese: *Eleverne på A-niveau og B-niveau er lige gode til brøkregning.*

Stikprøven er nogle elever som fik en prøve i brøkregning. Hver elev fik

under middel, *middel* eller *over middel*.

Resultatet står i tabellen.

Faktiske tal:

	under	middel	over	i alt
mat A	**51**	**65**	**78**	194
mat B	**44**	**72**	**46**	162
i alt	95	137	124	356

Vi udregner de forventede værdier:

95 af de 356 elever fik *under middel*, så hvis hypotesen er rigtig, må vi forvente at også $\frac{95}{356}$ af mat A-eleverne fik *under middel*.

Det forventede antal er altså $\frac{95}{356} \cdot 194 = 51{,}77$.

Det forventede antal mat A-elever med karakteren *middel* er $\frac{137}{356} \cdot 194 = 74{,}66$.

Vi kan fortsætte sådan for at udregne resten af de forventede tal, men vi kan også udregne resten af de forventede værdier ved at bruge at vi kender summen af hver række og søjle.

Det forventede antal mat B-elever med karakteren *under middel* kan vi altså udregne sådan: $95 - 51{,}77 = 43{,}23$.

Forventede tal:

	under	middel	over	i alt
mat A	**51,77**	**74,66**	**67,57**	194
mat B	**43,23**	**62,34**	**56,43**	162
i alt	95	137	124	356

Vi udregner χ^2 som er afstanden mellem de faktiske tal og de forventede tal:

$$\chi^2 = \frac{(51-51{,}77)^2}{51{,}77} + \frac{(65-74{,}66)^2}{74{,}66} + \frac{(78-67{,}57)^2}{67{,}57} + \frac{(44-43{,}23)^2}{43{,}23} + \frac{(72-62{,}34)^2}{62{,}34} + \frac{(46-56{,}43)^2}{56{,}43}$$

Vi får $\chi^2 = 6{,}31$.

$$\text{Antal frihedsgrader} = (\text{antal rækker} - 1) \cdot (\text{antal søjler} - 1) = (2-1) \cdot (3-1) = 2$$

Vi udregner p som er sandsynligheden hvis hypotesen er rigtig for at χ^2 er i hvert fald 6,31:

$p = \chi^2\text{Cdf}(6.31, \infty, 2) = 4{,}3\,\%$.

Da p er mindre end 5 %, forkaster vi hypotesen, så på 5 %-signifikansniveau har vi vist:

Eleverne på A-niveau og B-niveau er ikke lige dygtige til brøkregning.

12 Nulhypotese.

Når vi udfører en test, så undersøger vi om en bestemt hypotese kan forkastes. Denne hypotese kaldes nulhypotesen. Nulhypotesen skal påstå at noget er ens.
Dette kan udtrykkes på flere måder. Følgende fire sætninger er samme hypotese:

Der er ikke forskel på pigers og drenges holdning til spørgsmålet.
Pigers og drenges holdning til spørgsmålet er ens.
Elevers holdning til spørgsmålet er uafhængig af køn.
Elevers køn har ikke betydning for deres holdning til spørgsmålet.
Der er ikke sammenhæng mellem køn og holdning til spørgsmålet.

Den alternative hypotese er den hypotese som vi har belæg for at hævde hvis testen forkaster nulhypotesen.

Hvis vi vil undersøge om der er forskel på pigers og drenges holdning til et spørgsmål, så skal vi skrive følgende nulhypotese:

Nulhypotese: *Pigers og drenges holdning er ens*.

Alternativ hypotese: *Pigers og drenges holdning er forskellig*.

13 Test for fordeling når stikprøven er angivet som antal.

I et land er det muligt at stemme på på fire partier A, B, C og D.

Hypotese: *Tilslutningen til partierne er som vist i tabellen.*

Hypotese:

A	B	C	D
33 %	8%	39 %	20 %

Vi spørger folk der er tilfældigt valgt blandt dem der må stemme, og får følgende 150 svar:

Faktiske tal:

A	B	C	D
46	22	55	27

Populationen er dem der må stemme. Stikprøven er dem der svarede.

Vi vil teste hypotesen på signifikansniveu 5 %.

13.1 Sådan udregner vi forventede tal.

Hvis hypotesen er rigtig, må vi forvente følgende tal:

A: $150 \cdot 0{,}33 = 49{,}5$ B: $150 \cdot 0{,}08 = 12$ C: $150 \cdot 0{,}39 = 58{,}5$ D: $150 \cdot 0{,}20 = 30$

Forventede tal:

A	B	C	D
49,5	12	58,5	30

13.2 Sådan udregner vi χ^2.

Symbolet χ^2 læses *ki i anden*

For hvert af de fire felter udregner vi

$$\frac{(faktisk - forventet)^2}{forventet}$$

Ved at lægge disse fire tal sammen får vi tallet χ^2 som er afstanden mellem de faktiske tal og de forventede tal:

$$\chi^2 = \frac{(46-49{,}5)^2}{49{,}5} + \frac{(22-12)^2}{12} + \frac{(55-58{,}5)^2}{58{,}5} + \frac{(27-30)^2}{30}$$

$$\chi^2 = 9{,}09$$

13.3 Sådan udregner vi antal frihedsgrader.

Ovenfor skrev vi tal i følgende skema:

A	B	C	D	I alt
				150

Når vi har skrevet tal i tre af felterne, så er det fastlagt hvad der skal stå i det sidste felt da summen skal være 150.

Antal frihedsgrader er 3 da vi kan vælge tre af tallene.

I **test for fordeling** gælder:

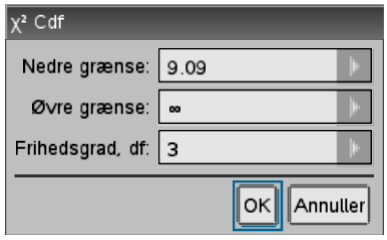
$$\text{antal frihedsgrader} = \text{antal felter} - 1 .$$

ADVARSEL: Nogle gange skal vi udregne antal frihedsgrader på en anden måde.
Se afsnit 10.3 .

13.4 Sådan udregner vi p.

Tallet p er sandsynligheden hvis hypotesen er rigtig,
for at afstanden χ^2 er 9,09 eller større (dvs. mellem 9,09 og ∞).

Vi kan få Nspire til at udregne p ved i beregningsmenuen at vælge
Statistik / Fordelinger / χ^2 Cdf... og udfylde sådan:



På Nspire-lommeregneren kan vi skrive ∞ ved hjælp af π-tasten.
På computeren skriver vi infinity i stedet for ∞.

$p = \chi^2\text{Cdf}(9.09, \infty, 3) = 2{,}8\,\%$ *udregnet på Nspire*

13.5 Sådan skriver vi konklusionen.

Da p er mindre end 5 %, forkaster vi hypotesen. På 5 %-signifikansniveau har vi belæg for at hævde at:

Fordelingen er ikke som vist i første tabel.

13.6 Misforstå ikke procenttallene.

2,5 % er IKKE sandsynligheden for at hypotesen er rigtig.
2,5 % er IKKE sandsynligheden for at hypotesen er forkert.
97,5 % er IKKE sandsynligheden for at hypotesen er rigtig.
97,5 % er IKKE sandsynligheden for at hypotesen er forkert.

De 2,5 % er udregnet under den forudsætning at hypotesen er rigtig
og er sandsynligheden for at få en stikprøve hvis afvigelse fra hypotesen
er så stor som eller større end afvigelsen i den stikprøve vi fik.

14 Test for fordeling når stikprøven er angivet med procenter.

I et land er det muligt at stemme på på fire partier A, B, C og D.

Hypotese: *Tilslutningen til partierne er som vist i tabellen.*

Hypotese:

A	B	C	D
33 %	8%	39 %	20 %

Nogen har spurgt folk der er tilfældigt valgt blandt dem der må stemme, og har fået svar fra 150. De har angivet resultatet i procenter sådan:

Vores data:

A	B	C	D
30,7 %	14,7 %	36,7 %	18,0 %

For at kunne teste hypotesen må vi udregne de faktiske tal som disse procenter er udregnet ud fra.

I tabellen med faktiske tal SKAL alle tal være hele tal.

(I tabellen med forventede tal må der gerne stå kommatal).

A: $150 \cdot 0{,}307 = 46{,}05$ B: $150 \cdot 0{,}147 = 22{,}05$ C: $150 \cdot 0{,}367 = 55{,}05$ D: $150 \cdot 0{,}18 = 27$

Faktiske tal:

A	B	C	D
46	22	55	27

Da vi nu har de faktiske tal, kan vi udføre testen på samme måde som i afsnit 13.

15 Kritisk værdi.

Hver dag undersøger vi nogle varer med en χ^2-test med 2 frihedsgrader og signifikansniveau 5 %.

En dag er $\chi^2 = 6{,}88$. Så er $p = 3{,}2\%$, dvs. vi forkaster.

En dag er $\chi^2 = 4{,}68$. Så er $p = 9{,}6\%$, dvs. vi forkaster ikke.

Når $\chi^2 = 6{,}88$ forkaster vi, og når $\chi^2 = 4{,}68$ forkaster vi ikke.
Der må være en χ^2-værdi α mellem 6,88 og 4,68 som skiller mellem at forkaste og ikke forkaste.
Vi vil finde den værdi der skiller, dvs. den χ^2-værdi hvor $p = 5\%$:

Nspire

løser ligningen $\chi^2\text{Cdf}(\alpha,\infty,2) = 0{,}05$ mht. α

og får $\alpha = 5{,}99$.

Tallet 5,99 kaldes den kritiske værdi .
De dage hvor χ^2 er større end 5,99 , forkaster vi hypotesen.

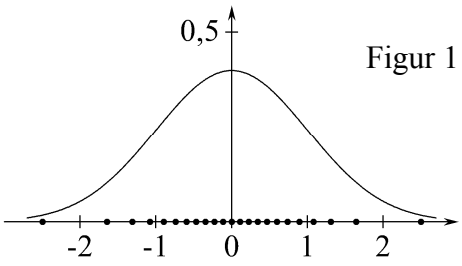
(I andre test vil de kritiske værdier normalt være andre tal).

FORDELINGER

16 Normalfordeling. Grafen viser tallenes fordeling.

Vi har anbragt ti millioner tal på en tallinje.

På figuren har vi tegnet nogle få af disse tal som prikker. Disse få af tallene har vi valgt sådan at de giver et indtryk af hvordan alle de ti mio. tal er fordelt.

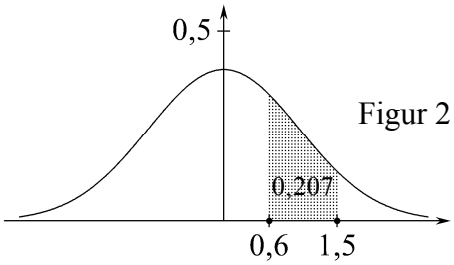


Figur 1

Vi ser at de ti mio. tal er fordelt sådan:
- Der er mange i midten.
- De ligger symmetrisk om midten.
- Der bliver færre jo længere vi kommer væk fra midten.

Vi har tegnet grafen for funktionen

$$f(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}$$



Figur 2

Denne funktion viser hvordan vi har fordelt tallene:

> Den procentdel af tallene der ligger i et interval,
> er lig arealet under grafen i dette interval.

Grafen for f er altså et slags afrundet histogram.

$$\text{Gråt areal} = \int_{0,6}^{1,5} f(x)\,dx = 0{,}207 \qquad \textit{udregnet på Nspire}$$

Der er altså $20{,}7\,\%$ af tallene der ligger i intervallet $0{,}6 \le x \le 1{,}5$. Hele arealet under grafen er $1 = 100\,\%$.

Hvis vi mange gange udpeger et tilfældigt af tallene, så vil vi $20{,}7\,\%$ af gangene få et tal i intervallet $0{,}6 \le x \le 1{,}5$, dvs. sandsynligheden for at få et tal i dette interval er $20{,}7\,\%$.

17 Nogle regler om grafer.

Vi ser på en funktion *g* der er positiv (dvs. grafen ligger over *x*-aksen) og et positivt tal *k*.

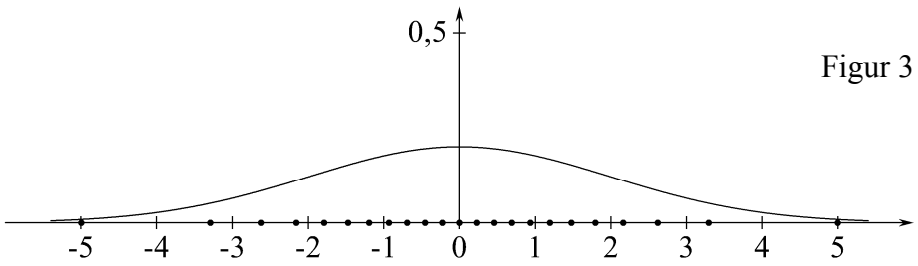
Når vi i forskriften for *g* erstatter x med $\frac{x}{k}$, så vil grafen blive strakt i vandret retning så alle grafpunkters afstand til *y*-aksen bliver ganget med k . Arealet mellem *x*-akse og graf for $g(\frac{x}{k})$ er altså k gange arealet mellem *x*-akse og graf for $g(x)$.
Hvis *k* er under 1, bliver grafen trykket sammen i vandret retning.

Når vi dividerer forskriften for g med k , så bliver grafen trykket sammen i lodret retning så alle grafpunkters afstand til *x*-aksen bliver divideret med k . Arealet mellem *x*-aksen og grafen for $\frac{1}{k}g(\frac{x}{k})$ er altså lig arealet mellem *x*-aksen og grafen for $g(x)$.

Når vi i forskriften for en funktion erstatter x med $x-m$, så forskydes grafen m enheder mod højre. Hvis m f.eks. er -3 , bliver grafen altså forskudt 3 enheder mod venstre. Grafen for $\frac{1}{k}g(\frac{x-m}{k})$ fremkommer altså ved at grafen for $\frac{1}{k}g(\frac{x}{k})$ forskydes m enheder mod højre.

18 Normalfordeling. Tal der er mere spredt.

$f(x)$ er stadig funktionen fra afsnit 5. Grafen for $f(x)$ strækker vi i vandret retning med faktoren 2, og i lodret retning med faktoren $\frac{1}{2}$. Så fremkommer grafen på figur 3. Ifølge afsnit 6 er dette grafen for $\frac{1}{2}f(\frac{x}{2})$, og arealet mellem *x*-aksen og grafen vil være 1. Hvis vi anbringer ti millioner tal på *x*-aksen så de er fordelt som denne graf viser, så vil disse tal ligge dobbelt så spredt som tallene fra afsnit 5. På figuren er nogle få af tallene angivet som prikker for at give et indtryk af hvordan de er fordelt.

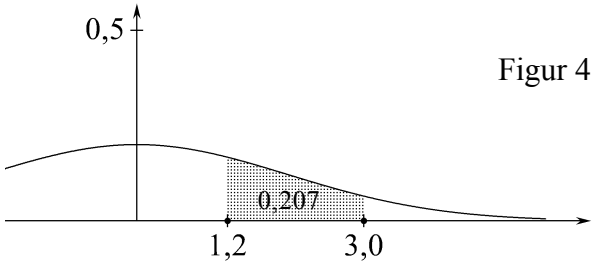


Figur 3

F.eks. gælder (se figur 4):

$$\text{Gråt areal} = \int_{1,2}^{3,0} \frac{1}{2}f(\frac{x}{2})\,dx = 0{,}207$$

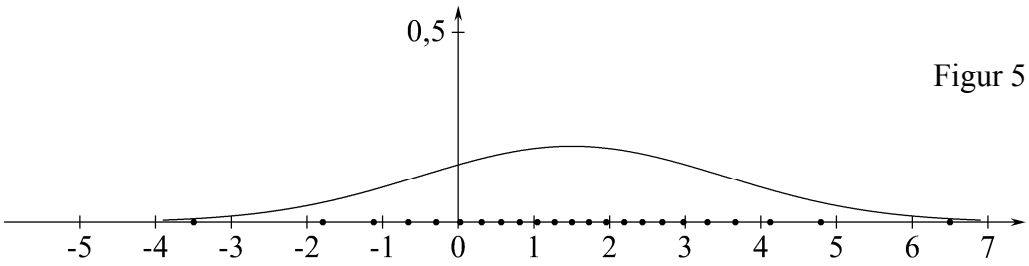
dvs. 20,7 % af tallene ligger mellem 1,2 og 3,0.



Figur 4

19 Normalfordeling. Forskydning af tallene.

Figur 3 viser grafen for $\frac{1}{2}f(\frac{x}{2})$. Denne graf forskyder vi 1,5 mod højre. Så får vi grafen på figur 5. Ifølge afsnit 6 er dette grafen for $\frac{1}{2}f(\frac{x-1,5}{2})$.

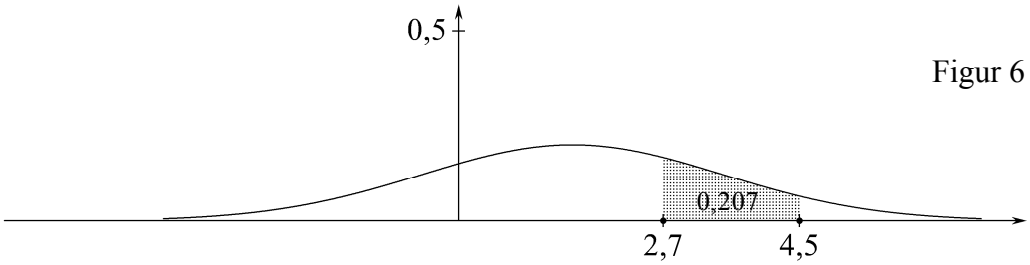


Figur 5

Hvis tallene fra afsnit 7 forskydes 1,5 enheder mod højre, så vil de være fordelt på *x*-aksen som grafen på figur 5 viser. F.eks. gælder (se figur 6):

$$\text{Gråt areal} = \int_{2,7}^{4,5} \frac{1}{2}f(\frac{x-1,5}{2})\,dx = 0{,}207$$

dvs. 20,7 % af tallene ligger mellem 2,7 og 4,5.



Figur 6

20 Normalfordeling. Middelværdi og spredning

20.1 Hvad er middelværdi og spredning for normalfordelte tal?

Når tal er normalfordelt, så er middelværdien tallet i midten der hvor grafen er højest.
Både de ti mio. tal på figur 1 og de ti mio. tal på figur 3 har middelværdien 0.
De ti mio. tal på figur 5 har middelværdien 1,5 .

De ti mio. tal på figur 1 har spredning 1.
De ti mio. tal på figur 3 og 5 ligger dobbelt så spredt, så deres spredning er 2.

Når tal er fordelt som angivet med funktionen

$$\frac{1}{s} f\left(\frac{x-m}{s}\right)$$

er de

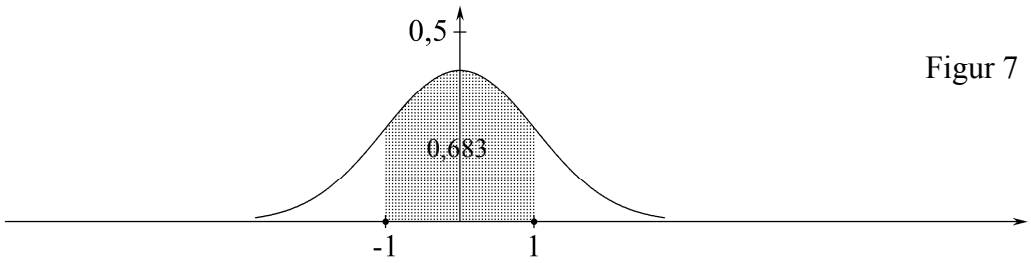
normalfordelt med middelværdi m og spredning s .

20.2 68,3 % af tallene fra figur 1.

Vi ser på tallene fra figur 1. Fra middelværdien 0 går vi spredningen 1 ud til begge sider. Så får vi intervallet $-1 \le x \le 1$. Den procentdel af tallene der ligger i dette interval er

$$\text{gråt areal på figur 7} = \int_{-1}^{1} f(x)\,dx = 0{,}683$$

Dvs. 68,3 % af tallene ligger i intervallet $-1 \le x \le 1$.



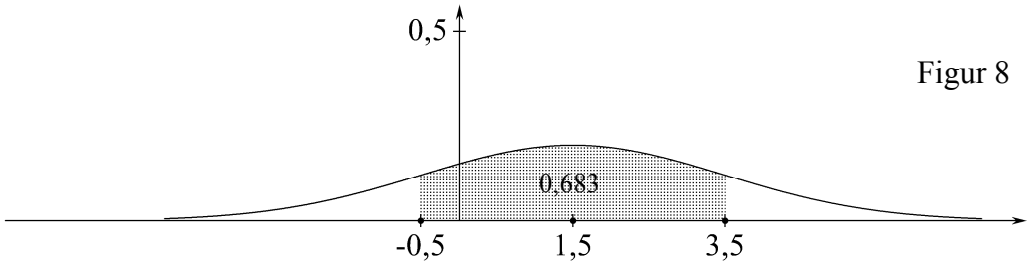
Figur 7

20.3 68,3 % af tallene fra figur 5.

Vi ser på tallene fra figur 5. Fra middelværdien 1,5 går vi spredningen 2 ud til begge sider . Så får vi intervallet $-0{,}5 \le x \le 3{,}5$. Den procentdel af tallene der ligger i dette interval er

$$\text{gråt areal på figur 8} = \int_{-0{,}5}^{3{,}5} \frac{1}{2} f\left(\frac{x-1{,}5}{2}\right) dx = 0{,}683$$

Dvs. 68,3 % af tallene ligger i intervallet $-0{,}5 \le x \le 3{,}5$.



Figur 8

20.4 68,3 % af normalfordelte tal.

Når vi fra middelværdien går spredningen ud på begge sider, så får vi et interval der indeholder 68,3 % af tallene.

21 Normalfordeling. En anvendelse.

Mange tal fra virkeligheden er normalfordelt. Her er et eksempel:

Opgave

En maskine fylder vin på flasker. Maskinens nøjagtighed er sådan at hvis vi måler mængden af vin i mange flasker, så vil vi få en række tal der er normalfordelt med spredningen 0,4 centiliter.

Vi indstiller maskinen så middelværdien af flaskernes indhold er 76 centiliter.

Hvor mange procent af flaskerne indeholder under 75 centiliter?

Svar

Nspire udregner at

$$\int_{-\infty}^{75} \frac{1}{0{,}4} f\left(\frac{x-76}{0{,}4}\right) dx \;=\; 0{,}00621$$

Dvs. $0{,}6\,\%$ af flaskerne indeholder under 75 centiliter.

22 χ^2-fordeling.

Nogle tal er normalfordelt med middelværdi 0 og spredning 1.
På figur 9 har vi tegnet nogle få af disse tal som prikker.
Disse få af tallene har vi valgt sådan at de giver et indtryk af hvordan alle tallene er fordelt.

Hvert af tallene opløfter vi til anden:

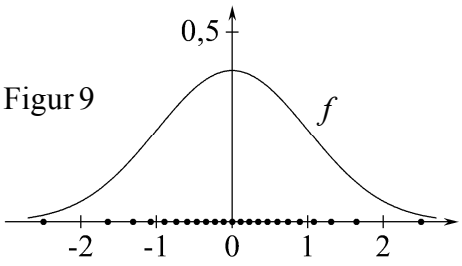
$(-0{,}6)^2 = 0{,}36$

$2{,}5^2 = 6{,}25$

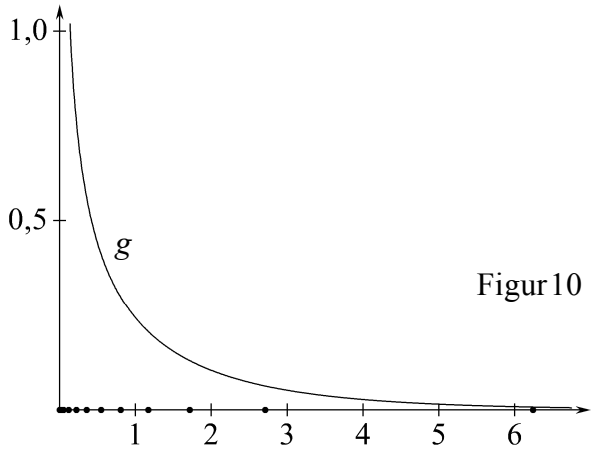
osv.

De tal vi får på denne måde, er fordelt som antydet på figur 10. Vi vil finde forskriften for funktionen g der viser hvordan disse tal er fordelt. (Kun få af tallene er vist som prikker).

Figur 9



Figur 10



g-grafen viser hvordan tallene χ^2 fra 9.3 er fordelt, dvs.

hvis vi mange gange tager en stikprøve og hver gang udregner χ^2 ,
så vil vi få nogle tal der er fordelt som *g*-grafen viser.

Da vi i afsnit 9 tog en stikprøve hvor $\chi^2 = 3{,}60$, kunne vi altså have udregnet p sådan:

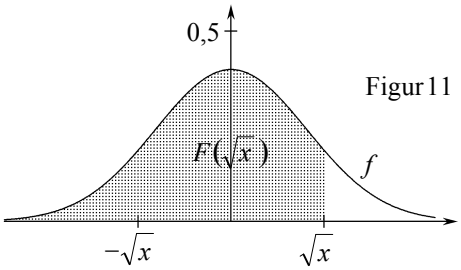
$$p \;=\; \int_{3{,}60}^{\infty} g(x)\,dx \;=\; 0{,}058 \;.$$

Her har vi brugt forskriften for g som står i linje (8) nederst i afsnit 12 .

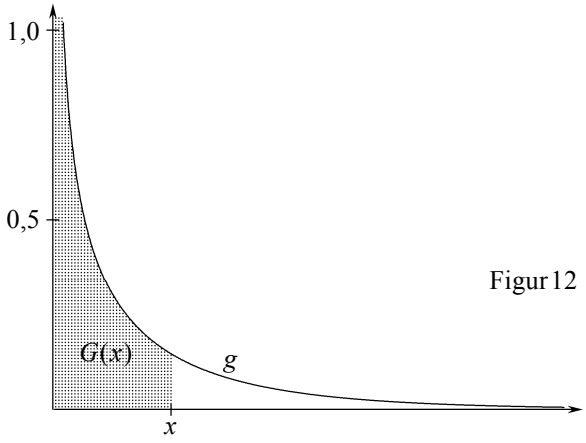
23 Forskrift for *g*.

f og g er funktionerne fra afsnit 11. Arealfunktionerne for f og g kalder vi F og G, dvs.

$F(x)$ = gråt areal på figur 11 og $G(x)$ = gråt areal på figur 12 .



Figur 11



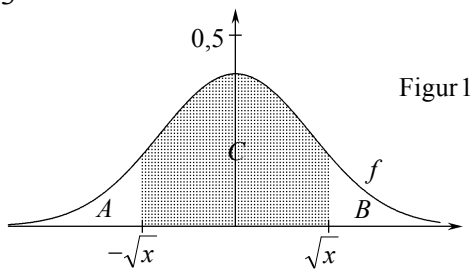
Figur 12

Tallene i intervallet $0 \le t \le x$ stammer fra tallene i intervallet $-\sqrt{x} \le t \le \sqrt{x}$, så

gråt areal på figur 12 = gråt areal på figur 13

dvs.

(1) $G(x) = C$



Figur 13

Af figur 11 og 13 ser vi at $C = F(\sqrt{x}) - A$.
Vi indsætter dette i (1):

(2) $G(x) = F(\sqrt{x}) - A$

Da *f*-grafen er symmetrisk, er $A = B$, så i (2) kan vi erstatte *A* med *B*:

(3) $G(x) = F(\sqrt{x}) - B$

Hele arealet under *f*-grafen er 1, så af figur 13 og 11 ser vi at $B = 1 - F(\sqrt{x})$.
Dette indsætter vi i (3):

(4) $G(x) = F(\sqrt{x}) - (1 - F(\sqrt{x}))$

Vi reducerer (4) og får:

(5) $G(x) = 2F(\sqrt{x}) - 1$

Da G er arealfunktion for g, er

(6) $g(x) = G'(x)$

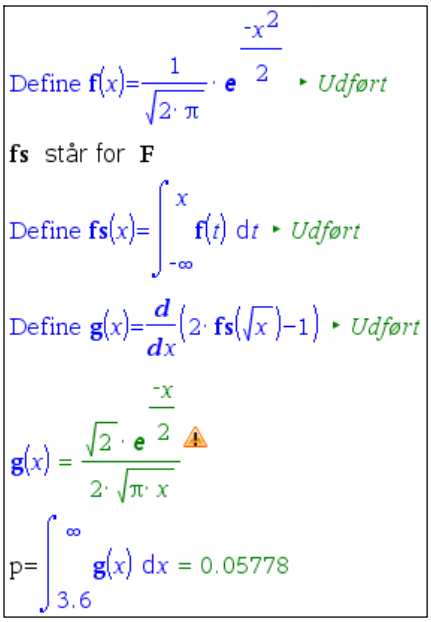
Af (5) og (6) får vi

(7) $g(x) = (2F(\sqrt{x}) - 1)'$

Nspire udregner højresiden og får

(8) $g(x) = \dfrac{\sqrt{2} \cdot e^{\frac{-x}{2}}}{2 \cdot \sqrt{\pi \cdot x}}$ *Denne formel skal du IKKE huske!*

I rammen ses hvordan vi taster for at få (8).



24 χ^2-fordeling når antal frihedsgrader ikke er 1.

Forskriften (8) omskriver vi til

(9) $g(x) = \frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} \cdot e^{-\frac{x}{2}}$, $x > 0$.

funktion viser fordelingen af tal der er χ^2-fordelt med 1 frihedsgrad. I (9) har x eksponenten $-\frac{1}{2}$. Hver gang vi lægger $\frac{1}{2}$ til eksponenten, bliver antallet af frihedsgrader 1 større. Samtidig må vi erstatte konstanten $\frac{1}{\sqrt{2\pi}}$ med en anden konstant så arealet under grafen stadig er 1. Hvis antallet af frihedsgrader er 4, så er tæthedsfunktionen altså af typen

$$g(x) = k \cdot x \cdot e^{-\frac{x}{2}}$$

Nspire løser ligningen

$$\int_0^{\infty} k \cdot x \cdot e^{-\frac{x}{2}} dx = 1$$

mht. k og får $k = \frac{1}{4}$, så

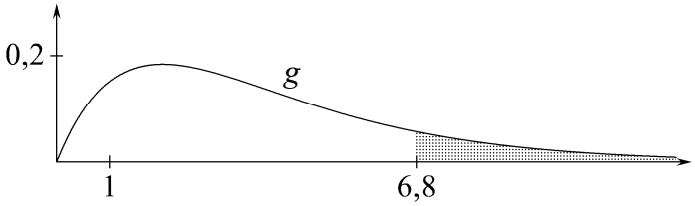
(10) $g(x) = \frac{1}{4} \cdot x \cdot e^{-\frac{x}{2}}$, $x > 0$

er funktionen der viser fordelingen af tal der er χ^2-fordelt med 4 frihedsgrader.

25 χ^2-fordeling og test.

Figuren viser grafen for funktionen (10) fra afsnit 24.
Hvis vi i en χ^2-test med 4 frihedsgrader får at χ^2 er 6,8, så er p lig det grå areal:

$$p = \int_{6,8}^{\infty} \frac{1}{4} \cdot x \cdot e^{-\frac{x}{2}} dx = 0{,}146842$$



Normalt regner vi dette tal ud sådan:

$p = \chi^2\text{Cdf}(6.8, \infty, 4) = 0{,}146842$

Stikordsregister