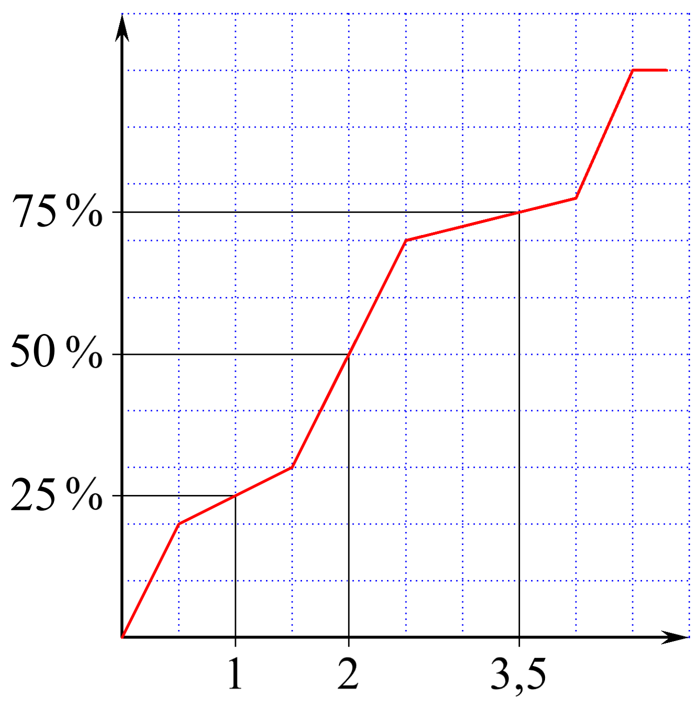
Statistik

for
gymnasiet og hf



2011 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.

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 middeltallet for ugrupperede data?

Middeltallet for nogle tal er det vi plejer at kalde gennemsnittet.
Vi kan udregne middeltallet 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 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 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?

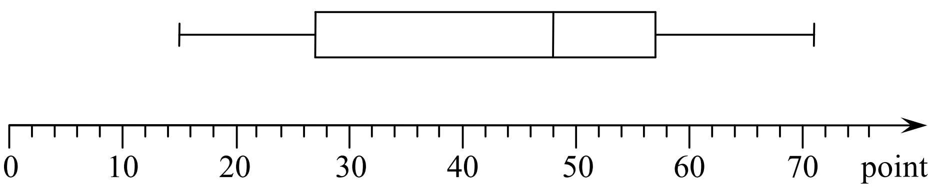
Ved at undersøge datasættet

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.

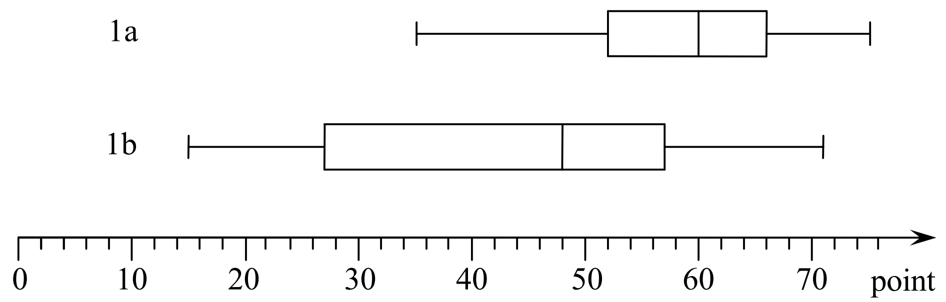
Den vandrette streg til venstre anskueliggør at den fjerdedel af tallene der er mindst, ligger i intervallet fra 15 til 27.

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.



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.

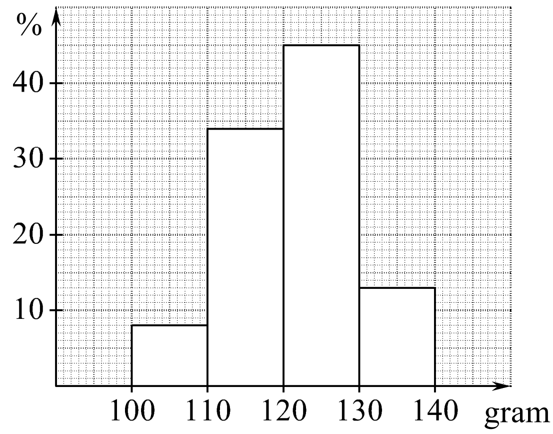
Vægt i gram	100-110	110-120	120-130	130-140
Procent	8	34	45	13

Histogrammet til højre viser oplysningerne i tabellen.

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

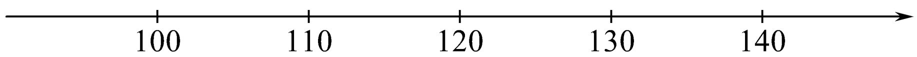
Dette viser at 8 % af frugterne vejer mellem 100 og 110 gram.

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).

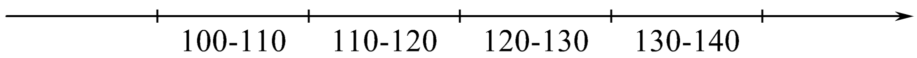


Advarsel: Den vandrette akse skal tegnes som en sædvanlig tallinje.

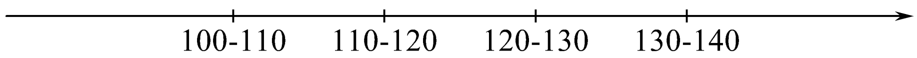
RIGTIGT:



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?

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

$$8\% + 34\% = 42\% \ , \quad 8\% + 34\% + 45\% = 87\% \ , \quad \text{osv.}$$

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

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:

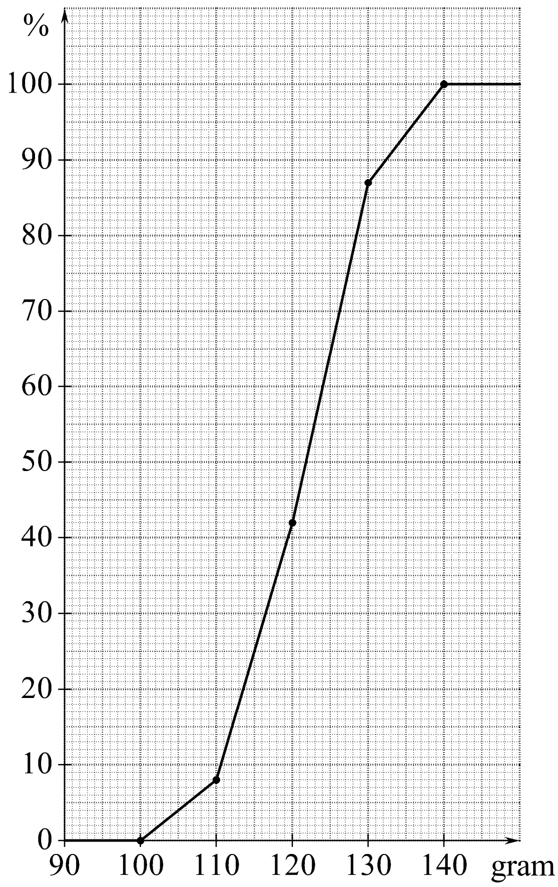
0% er mindre end 100, så ved $x = 100$ afsætter vi et punkt ud for 0% på y-aksen.

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).



Hvis vi kender antal i stedet for procent.

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

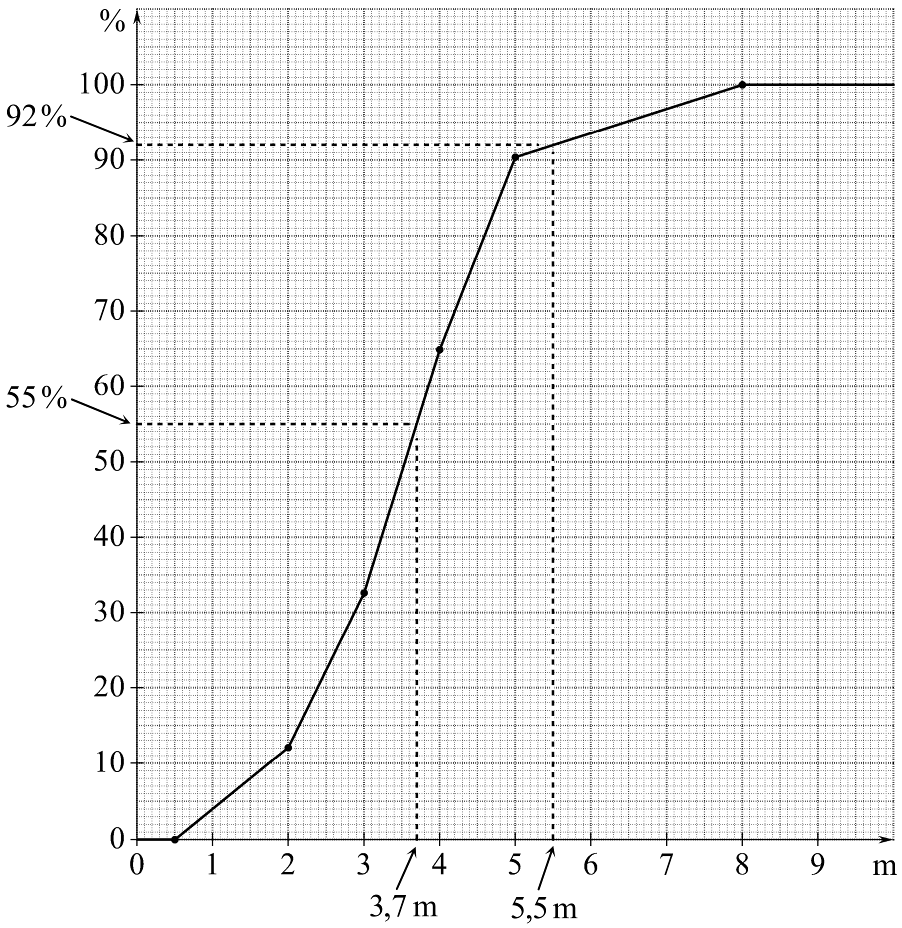
Tallene i 3. række udregner vi sådan: $34 + 58 \ = \ 92$, $34 + 58 + 91 \ = \ 183$, osv.

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 $\underline{\underline{55\%}}$ 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 $\underline{\underline{8\%}}$ 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 $\underline{\underline{37\%}}$ 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.

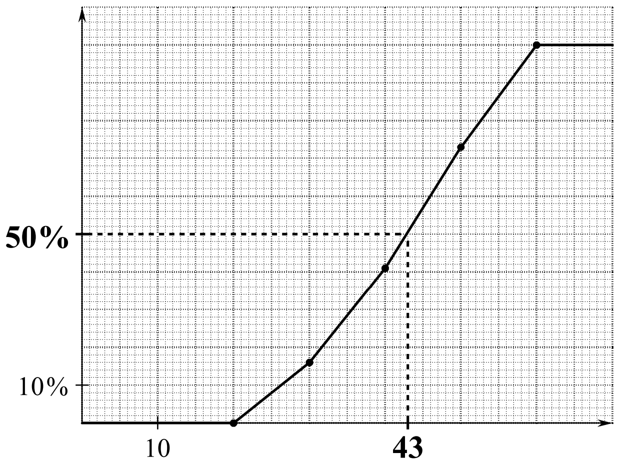
(For ugrupperede data skal vi gøre noget helt andet. Se afsnit 2.2 på side 2).

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.

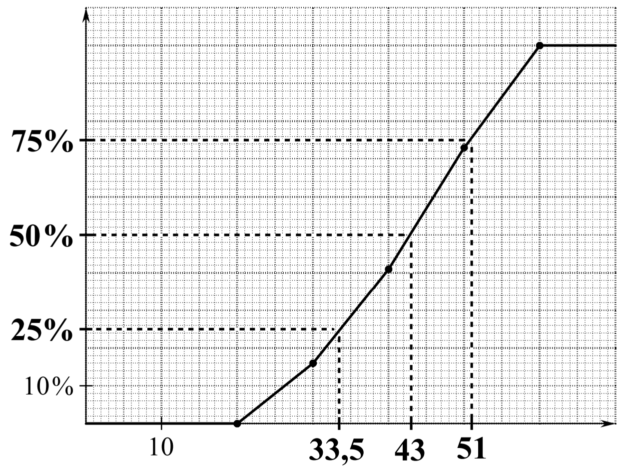
På figuren er medianen 43.



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 middeltallet for grupperede data?

Vi vil udregne middeltallet 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:

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

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

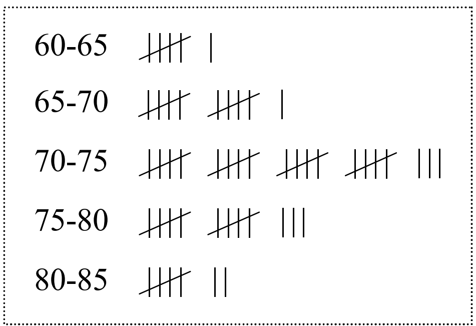
I rammen nedenfor har vi skrevet disse fem intervaller under hinanden.

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).



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

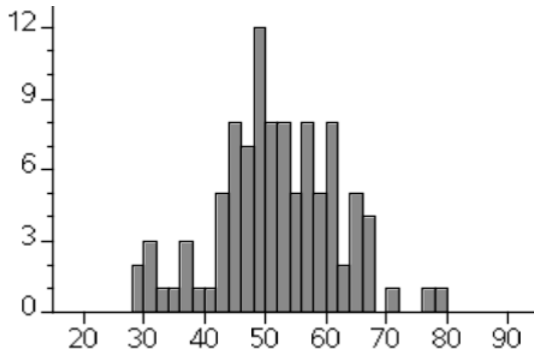
Intervallernes bredde er for lille.
Der er så få data i hvert interval at højden svinger tilfældigt op og ned.

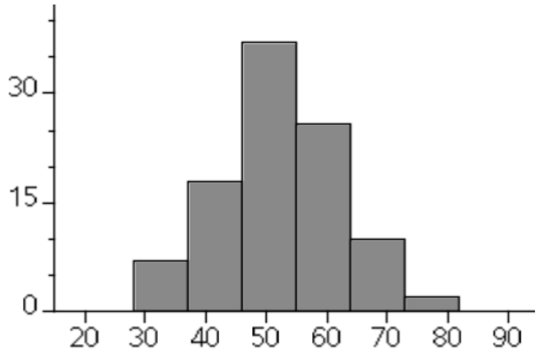
Midterste figur

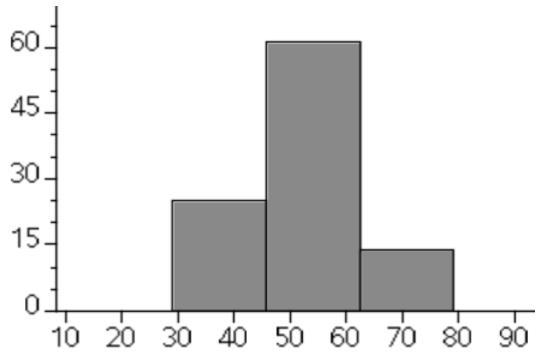
Intervallernes bredde er passende.

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 $75{,}0000\cdots$ 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.

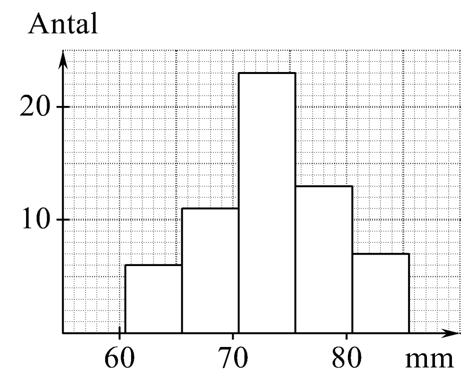
Dette problem kan vi undgå ved at bruge 75,5 som endepunkt i stedet for 75. Så bliver intervallernes endepunkter 60,5 , 65,5 , 70,5 osv.

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.

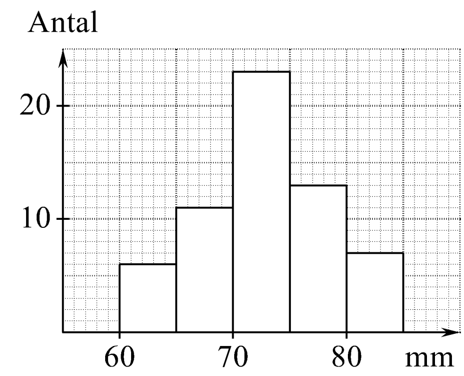
Her er der ingen data der er lig et endepunkt for et af intervallerne.



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

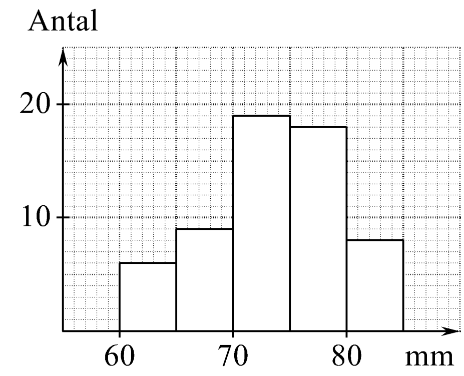
Rektanglerne har samme højder som på øverste figur, men de er anbragt en halv enhed længere mod venstre.

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.

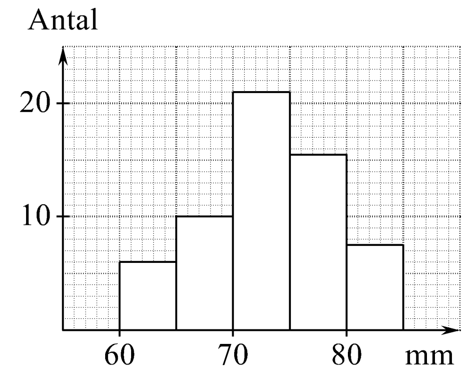


Alle data der er endepunkt for et af intervallerne, har vi talt med i intervallet til højre for endepunktet.

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



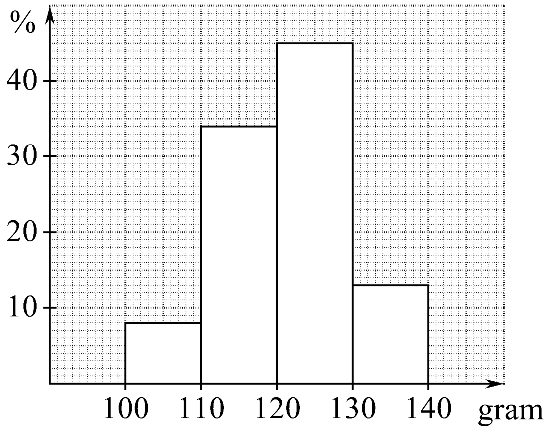
Alle data der er endepunkt for et af intervallerne, har vi talt med som en halv i hvert af de to intervaller med dette endepunkt.

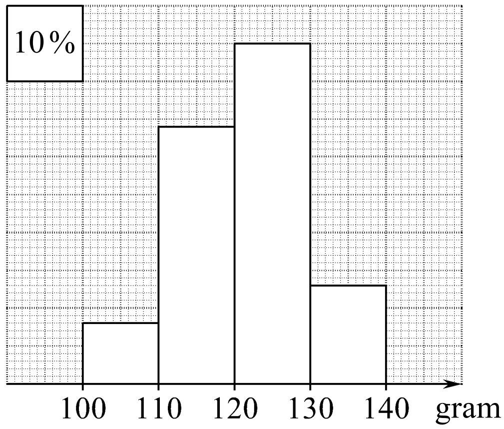


5.1 Vi kan tegne histogrammer på to måder.

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

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



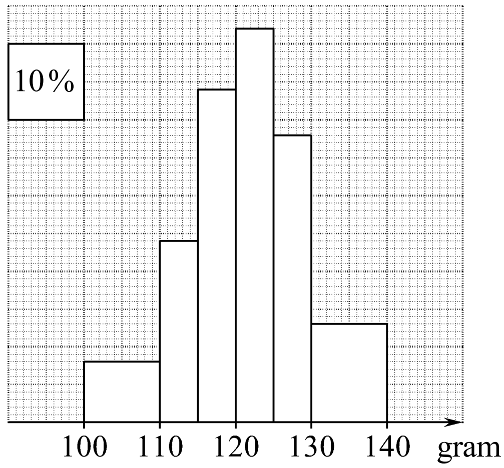


Når det er arealerne vi ser på, behøver intervallerne ikke være lige lange.

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

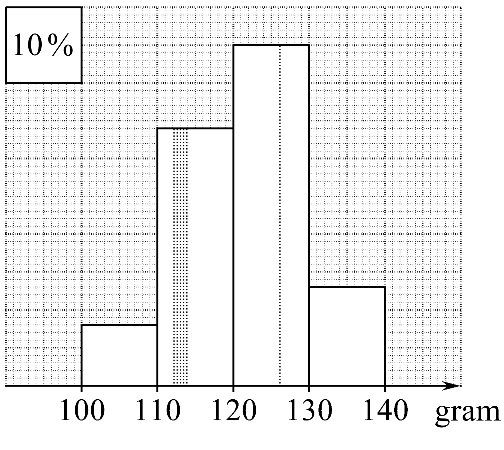
Men frekvensen er ikke tre gange så stor.

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 $117{,}00000\cdots$. 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

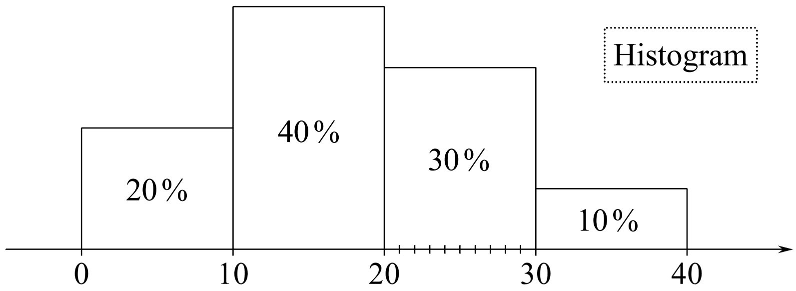
så gælder at

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

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

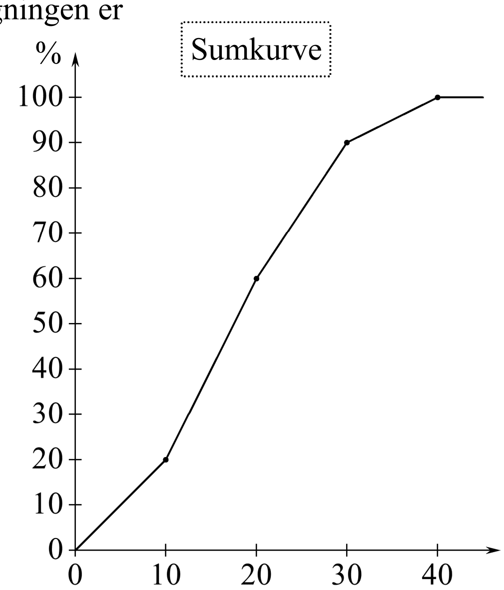
Vi udregner b :

Når $x = 20$ er $y = 0{,}60$ så

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

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

$y = 0{,}03x$.



<u>For de fire intervaller er ligningerne:</u>

0-10: $y = 0{,}02x$

10-20: $y = 0{,}04x - 0{,}2$

20-30: $y = 0{,}03x$

30-40: $y = 0{,}01x + 0{,}6$

<u>Hvor mange procent af observationerne har størrelse 27 eller derunder?</u>

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.

<u>Hvor stor er nedre kvartil?</u>

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

6.1 Stikprøver (nogle ting vi skal kunne skrive om i opgaver om stikprøver).

Skal gymnasieelever med over 10 % fravær smides ud?

Dette spørgsmål har en lokalavis stillet til personer i lokalområdet.

Vi vil undersøge om gymnasieelever har en anden holdning end indbyggerne i lokalområdet.

6.11 Hvad er populationen?

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

Er det alle landets gymnasieelever?
Er det alle elever på lokalområdets gymnasier?
Er det alle gymnasieelever som bor i lokalområdet?
Er det 1.g'erne på vores eget gymnasium?
Eller?

6.12 Hvad er stikprøven?

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

6.13 Er stikprøven repræsentativ?

Vi kan kun bruge stikprøven hvis vi med rimelighed kan regne med at den er repræsentativ,

dvs. hvis den ligner populationen med hensyn til den egenskab vi undersøger.

6.14 Systematiske fejl kan være årsag til at stikprøven ikke er repræsentativ.

Eksempel 1

Hvis populationen er alle landets gymnasieelever,
og vi kun spørger gymnasieelever fra lokalområdet,
så laver vi nok en systematisk fejl.

Fraværet er meget forskelligt i forskellige dele af landet. Dette er en af grundene til at vi ikke kan regne med at gymnasieelever fra lokalområdet har samme holdning som gymnasieelever i hele landet.

Eksempel 2

Hvis vi spørger elever pr. e-mail,
og mange ikke svarer,
så laver vi måske en systematisk fejl.

Det kan f.eks. være at nej-sigere i højere grad end ja-sigere tager sig sammen til at svare.

6.15 Tilfældige fejl kan være årsag til at stikprøven ikke er repræsentativ.

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

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

Risikoen for tilfældige fejl kan gøres mindre ved at tage en større stikprøve.

Afsnit 6.1 fortsætter på næste side.

6.16 Er der skjulte variable?

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

Eksempel 1

Måske svarer mange elever på noget andet end det de er blevet spurgt om, måske fordi de tror de bliver spurgt om en aktuel sag hvor der muligvis har været fejl i optællingen af forsømmelser. Elevens holdning til denne sag er den skjulte variabel der påvirker elevens svar.

Eksempel 2

Hvis der er flere der overlever en operation på hospital A end på hospital B, så slutter man måske at behandlingen er bedre på A end på B. Men forskellen skyldes måske at B har flere ældre patienter. Patienternes alder er så en skjult variabel der påvirker resultatet.

6.2 Hvad er sandsynlighed?

Med to eksempler viser vi hvad ordet sandsynlighed betyder.

6.21 Eksempel.

Vi trækker tre kort fra et sædvanligt spil kort (52 kort).

Man har udregnet at

sandsynligheden for at få tre røde kort er 11,8 % .

Dette betyder at

hvis vi mange gange trækker tre kort,
vil vi i 11,8 % af tilfældene få tre røde kort.

6.22 Eksempel.

Der står at 1% af varerne har fejl.
Vi udtager tilfældigt 30 varer. Tre af dem har fejl.

Man har udregnet at

hvis 1% af varerne har fejl, så er sandsynligheden 0,33 % for at vi får tre eller flere varer med fejl i en stikprøve på 30 varer.

Dette betyder at

hvis vi mange gange udtager 30 varer, så vil vi kun i 0,33 % af tilfældene få tre eller flere varer med fejl.

6.3 Test af hypotese. Signifikansniveau.

På en spillemaskine står at man får gevinst 20 % af de gange man spiller.

Hypotese: *Sandsynligheden for gevinst er 20 %* .

For at teste denne hypotese spiller vi 50 gange.

20 % af 50 er 10.

6.31 Hvis vi får gevinst 8 gange.

Man kan udregne at hvis hypotesen er rigtig, gælder at

sandsynligheden er lidt over 30 % for kun at vinde 8 eller færre gange,

dvs.

hvis vi mange gange spiller 50 gange, så vil vi i ca. en tredjedel af tilfældene vinde 8 eller færre gange.

Selv om hypotesen skulle være rigtig, er der altså ikke noget usædvanligt i at vinde så få gange som vi gjorde.

Vi kan altså ikke forkaste hypotesen.

Dette betyder ikke at vi har vist at hypotesen er rigtig.

6.32 Hvis vi får gevinst 3 gange.

Man kan udregne at hvis hypotesen er rigtig, gælder at

sandsynligheden er under 1 % for kun at vinde 3 eller færre gange,

dvs.

hvis vi mange gange spiller 50 gange, så vil vi i under en procent af tilfældene vinde 3 eller færre gange.

Hvis hypotesen er rigtig, er det altså meget usædvanligt at vinde så få gange som vi gjorde.

Derfor forkaster vi hyporesen.

Dvs. konklusionen er at der ikke er belæg for at hævde at sandsynligheden for gevinst er 20 %.

6.33 Hvad er signifikansniveau?

I første tilfælde ovenfor forkastede vi ikke hypotesen fordi sandsynligheden var over 30 %.

I andet tilfælde forkastede vi hypotesen fordi sandsynligheden var under 1%.

Hvor er grænsen for at vi forkaster? Ofte vedtager man grænsen 5 %. Så siger man at

signifikansniveauet er 5 % .

Det er også almindeligt at bruge signifikansniveauet 1 %.

7.1 Krydstabel. Hvordan udregner vi FORVENTEDE VÆRDIER?

Krydstabel

Nogle tilfældigt udvalgte elever får stillet samme spørgsmål.
De skal svare ja eller nej.

Antal svar kan skrives i en krydstabel som følgende:

	ja	nej
piger		
drenge		

Forventede værdier

Vi får oplyst at

der er 133 piger og 92 drenge, og 158 har svaret ja

Vi udregner:

Antal svar i alt: $133 + 92 = 225$

Antal nej-svar: $225 - 158 = 67$

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

Hvad er de forventede værdier hvis piger og drenge har samme holdning til det stillede spørgsmål?

158 af de 225 elever svarede ja, så vi må forvente 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 værdier 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

7.2 Krydstabel. Hvordan udregner vi χ^2-TESTSTØRRELSEN?

Symbolet χ^2 læses *ki i anden*

Nogle tilfældigt udvalgte elever får stillet samme spørgsmål.
De svarede sådan:

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

Antal piger i alt: $87+46 = 133$
Antal drenge i alt: $71+21 = 92$
Antal svar i alt: $133+92 = 225$
Antal ja-svar i alt: $87+71 = 158$
Antal nej-svar i alt: $46+21 = 67$

Faktiske tal	ja	nej	i alt
piger	**87**	**46**	133
drenge	**71**	**21**	92
i alt	158	67	225

Vi kan nu udregne de forventede tal. I afsnit 7.1 står mellemregningerne.

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

χ^2-teststørrelsen.

Vi vil udregne **ét** tal T der angiver
hvor meget de faktiske tal **afviger** fra de forventede tal .

Man har fundet ud af at det er praktisk at bruge det tal T vi får ved følgende udregning:
For hvert af de fire felter udregner vi

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

Ved at lægge disse fire tal sammen får vi størrelsen T som udtrykker hvor meget de faktiske tal afviger fra de forventede:

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

$$T = 3{,}60$$

Tallet 3,60 kaldes χ^2-teststørrelsen . (Symbolet χ^2 læses *ki i anden*).

7.3 Krydstabel. Hvordan udregner vi antallet af FRIHEDSGRADER?

7.31 Frihedsgrader for 2 gange 2 tabel.

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

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

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.

7.32 Frihedsgrader for 2 gange 3 tabel.

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

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

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.

7.33 Frihedsgrader for m gange n tabel.

Hvis der 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 $= (2-1)\cdot(2-1) = 1\cdot 1 = 1$

7.4 Krydstabel. Kan vi FORKASTE HYPOTESEN?

Vi vil teste følgende hypotese:

Hypotese: *Piger og drenge har samme holdning til det stillede spørgsmål.*

Vi bruger et signifikansniveau på 5 %.

Stikprøvens afvigelse fra hypotesen er 3,60 (det udregnede vi i 7.2).

Antallet af frihedsgrader er 1 (det udregnede vi i 7.3).

7.41 Hvordan kan vi udregne *p*-værdien?

Symbolet

$\chi^2\mathrm{Cdf}(3.60, \infty, 1)$

betyder:

sandsynligheden for at afvigelsen er 3,60 eller større (mellem 3,60 og uendelig).

1-tallet er antallet af frihedsgrader.
Sandsynligheden er udregnet under forudsætning af at hypotesen er rigtig.

Denne sandsynlighed kalder vi

p-værdien.

Vores elektroniske hjælpemiddel udregner at

$p = \chi^2\mathrm{Cdf}(3.60, \infty, 1) = 5{,}8\,\%$.

7.42 Hvordan kan vi skrive konklusionen?

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

Stikprøven giver ikke belæg for at hævde at piger og drenge har forskellig holdning til det spørgsmål der blev stillet.

7.43 Misforstå ikke procenttallene.

5,8 % 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.

7.5 Krydstabel. Forskellige ord der beskriver samme hypotese.

Bemærk at 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.

7.6 Krydstabel. Gennemregnet eksempel med to frihedsgrader.

På en skole er der 356 elever elever i 1.g som er på en studieretning med matematik på A-niveau eller B-niveau. Vi vil teste følgende hypotese på 5 %-signifikansniveau:

Hypotese: *Der er ikke forskel på hvordan eleverne på A- og B-niveau klarer start-matematikprøven.*

Alle elever fik samme prøve. Hver elev fik

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

Resultatet står i tabellen.

Faktiske værdier:

	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 værdier, 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 værdier:

	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-teststørrelsen:

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

Vi får $T = 6{,}31$.

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

Vores elektroniske hjælpemiddel udregner at

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

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

Der er forskel på hvordan eleverne på A- og B-niveau klarer start-matematikprøven.

7.7 Krydstabel. At acceptere hypotesen $\neq$ at tro at hypotesen er rigtig.

Vi stiller samme spørgsmål til 100 tilfældigt valgte piger og til 100 tilfældigt valgte drenge. De skal svare ja eller nej. Resultatet står i tabellen.

	ja	nej	i alt
piger	**91**	**9**	100
drenge	**82**	**18**	100
i alt	173	27	200

Vi har stillet spørgsmålet for at teste følgende:

Hypotese: *Holdningen til spørgsmålet er uafhængig af køn.*

Vi har valgt signifikansniveauet 5%.
Vi bruger en χ^2-test og får $p = 6{,}26\%$.

Da p ikke er mindre end 5%, kan vi ikke forkaste hypotesen om uafhængighed.

Da vi ikke kan forkaste hypotesen, må vi acceptere hypotesen.

At vi accepterer hypotesen betyder IKKE at vi betragter hypotesen som mere troværdig end flere andre hypoteser som er i modstrid med den.

Dette skyldes følgende overvejelser:

I stikprøven er der 9% af drengene og 18% af pigerne der siger nej.
I stikprøven er andelen af piger der siger nej, altså dobbelt så stor som andelen af drenge der siger nej.

Hvis populationen ligner stikprøven meget, så er holdningen altså ikke uafhængig af køn.

Det kan tænkes at det er en skæv stikprøve, og at der i populationen er uafhængighed.

Men stikprøven kan også være skæv i modsat retning så forskellen på pigers og drenges holdning er større i populationen end i stikprøven.

Det er kun når vi forkaster en hypotese at vi har vist noget.

8.1 Test for fordeling. Hvordan udregner vi FORVENTEDE VÆRDIER?

Et stort antal gymnasieelever er medlem af en bestemt klub på internettet. De er delt i holdningen til et bestemt spørgsmål hvor nogle siger nej, og de andre siger ja.

Hypotese: *Fordelingen er som vist i tabellen.*

Hypotese:

Piger, ja	Piger, nej	Drenge, ja	Drenge, nej
30%	10%	40%	20%

For at teste denne hypotese spørger vi 150 elever der er tilfældigt valgt blandt klubbens medlemmer. Vi får følgende svar:

Faktiske værdier:

Piger, ja	Piger, nej	Drenge, ja	Drenge, nej
44	26	55	25

Hvis hypotesen er rigtig, må vi forvente følgende værdier:

Piger, ja: $150 \cdot 0{,}30 = 45$
Piger, nej: $150 \cdot 0{,}10 = 15$
Drenge, ja: $150 \cdot 0{,}40 = 60$
Drenge, nej: $150 \cdot 0{,}20 = 30$

Forventede værdier:

Piger, ja	Piger, nej	Drenge, ja	Drenge, nej
45	15	60	30

8.2 Test for fordeling. Hvordan udregner vi χ^2-TESTSTØRRELSEN?

I 8.1 ovenfor er der en tabel med faktiske tal og en tabel med forventede tal .

Symbolet χ^2 læses *ki i anden*

Vi vil udregne **ét** tal T der angiver
hvor meget de faktiske tal **afviger** fra de forventede tal .

Man har fundet ud af at det er praktisk at bruge det tal T man får ved følgende udregning:

For hvert af de fire felter udregner vi

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

Ved at lægge disse fire tal sammen får vi størrelsen T som udtrykker hvor meget de faktiske tal afviger fra de forventede:

$$T = \frac{(44-45)^2}{45} + \frac{(26-15)^2}{15} + \frac{(55-60)^2}{60} + \frac{(25-30)^2}{30}$$

$$T = 9{,}34$$

Tallet 9,34 kaldes χ^2-teststørrelsen . (Symbolet χ^2 læses *ki i anden*).

8.3 Test for fordeling. Hvordan udregner vi antallet af FRIHEDSGRADER?

I 8.1 og 8.2 skrev vi tal i følgende skema:

Piger, ja	Piger, nej	Drenge, ja	Drenge, nej	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.

Der gælder:

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

	ja	nej
piger		
drenge		

Nogle forfattere skriver skemaet som vist til højre.
Antal frihedsgrader er 3 selv om skemaet er skrevet sådan.
Det er en fælde at skrive skemaet sådan, da vi kunne tro at vi skulle bruge en anden regel til at udregne antal frihedsgrader (se evt. afsnit 7.3 på side 21).

8.4 Test for fordeling. Kan vi FORKASTE HYPOTESEN?

Vi vil teste hypotesen

Hypotese: *Fordelingen er som skrevet i øverste tabel i 8.1* .

Vi bruger et signifikansniveau på 5 %.

Stikprøvens afvigelse fra hypotesen er 9,34 (det udregnede vi i 8.2).
Antallet af frihedsgrader er 3 (det udregnede vi i 8.3).

8.41 Hvordan kan vi udregne *p*-værdien?

Symbolet

$\chi^2\text{Cdf}(9.34, \infty, 3)$

betyder:

sandsynligheden for at afvigelsen er 9,34 eller større (mellem 9,34 og uendelig).

3-tallet er antallet af frihedsgrader.
Sandsynligheden er udregnet under forudsætning af at hypotesen er rigtig.

Denne sandsynlighed kalder vi

p-værdien.

Vores elektroniske hjælpemiddel udregner at

$$p \;=\; \chi^2\text{Cdf}(9.34, \infty, 3) \;=\; 2{,}5\,\%\ .$$

8.42 Hvordan kan vi skrive konklusionen?

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

Fordelingen er ikke som vist i øverste tabel i 8.1 .

8.43 Misforstå ikke procenttallene.

2,5 % 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.

8.5 Test for fordeling. At acceptere hypotesen ≠ at tro at hypotesen er rigtig.

Der er fire typer varer som vi kalder A, B, C og D.

En person påstår at fordelingen er sådan:

Hypotese 1:

A	B	C	D
40%	20%	20%	20%

For at teste denne hypotese, udtager vi tilfældigt 40 varer. Fordelingen i stikprøven er sådan:

Stikprøve:

A	B	C	D
12	12	7	9

Vi udregner χ^2-teststørrelsen og får 3,25. Med 3 frihedsgrader giver dette $p = 35{,}5\%$, så vi accepterer hypotese 1.

En anden person påstår at fordelingen er sådan:

Hypotese 2:

A	B	C	D
20%	40%	20%	20%

Vi tester hypotese 2 med stikprøven ovenfor og får $p = 35{,}5\%$, så vi accepterer hypotese 2.

Ovenstående tydeliggør følgende:

At vi accepterer hypotesen betyder IKKE at vi betragter hypotesen som mere troværdig end flere andre hypoteser som er i modstrid med den.

Det er kun når vi forkaster en hypotese at vi har vist noget.

9.1 Hvordan opstiller vi en nulhypotese?

Når vi udfører en test, så undersøger vi om en bestemt hypotese kan forkastes. Denne hypotese kaldes nulhypotesen.

(Ordet *nulhypotese* er en fordanskning af det engelske udtryk *null hypothesis*.
Ordet *null* betyder *ugyldig*).

Når vi skal skrive hvad nulhypotesen påstår, skal vi (normalt) skrive

at noget er ens .

Dette kan vi også udtrykke ved at skrive

at der ikke er forskel ,
eller at noget er uafhængigt af noget andet,
eller at noget ikke har betydning for noget andet,
eller at der ikke er nogen sammenhæng .

Den alternative hypotese er den hypotese som vi vil anse som troværdig hvis vi kan forkaste 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.*

9.2 Kritisk værdi.

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

> En dag er teststørrelsen $T = 6{,}88$. Så er $p = 3{,}2\,\%$, dvs. vi forkaster.
>
> En dag er teststørrelsen $T = 4{,}68$. Så er $p = 9{,}6\,\%$, dvs. vi forkaster ikke.

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

> Vores regnetekniske hjælpemiddel
>
> løser ligningen $\chi^2\mathrm{Cdf}(T,\infty,2) = 0{,}05$ mht. T
>
> og får $T = 5{,}99$.

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

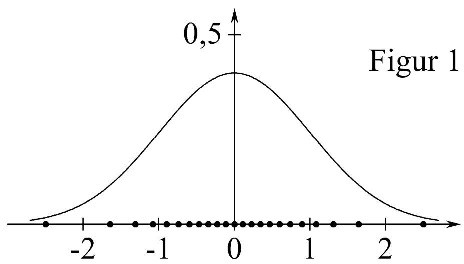
(I andre test vil de kritiske værdier normalt være andre tal).

FORDELINGER

10.1 Normalfordeling. Tæthedsfunktionen 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.



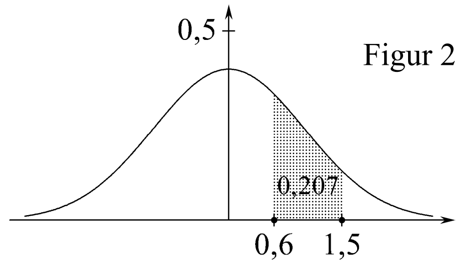
Figur 1

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

Denne funktion viser vi hvordan vi har fordelt tallene:

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



Figur 2

Man siger at f er <u>tæthedsfunktionen</u> for de ti mio. tal.
Grafen for en tæthedsfunktion er altså et slags afrundet histogram.

$$\text{Gråt areal} = \int_{0,6}^{1,5} f(x)\,dx = 0{,}207 \ .$$

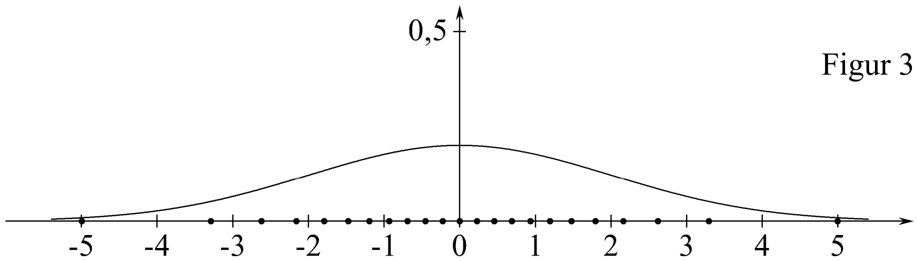
Der er altså $20{,}7\%$ af tallene der ligger i intervallet $0{,}6 \le x \le 1{,}5$. Hele arealet under grafen er $1 = 100\%$. (Vi brugte vores regnetekniske hjælpemiddel til at udregne integralet).

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. <u>sandsynligheden</u> for at få et tal i dette interval er $20{,}7\%$.

10.2 Normalfordeling. Nogle tal der er mere spredt.

Hvert af de ti mio. tal fra 10.1 ganger vi med 2:

$0{,}6 \cdot 2 = 1{,}2$
$1{,}5 \cdot 2 = 3{,}0$
osv.



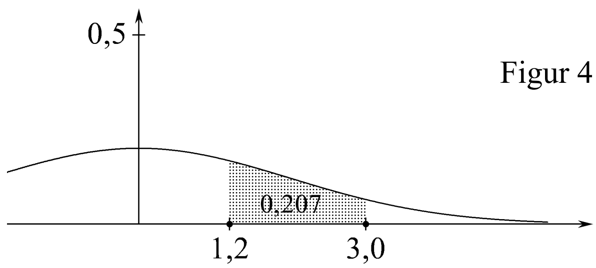
Figur 3

De tal vi får som resultater, ligger mere spredt.

På figur 1 øverst har vi tegnet nogle få af tallene.
Når vi ganger disse tal med 2, så får vi de tal vi har tegnet på figur 3.

Grafen på figur 3 er dobbelt så bred som grafen på figur 1. Derfor må højden være den halve da arealet under grafen stadig skal være $1 = 100\%$.

Når vi ganger de af de ti mio. tal som ligger i intervallet $0{,}6 \le x \le 1{,}5$ med 2, så bliver disse tal ført over i intervallet $1{,}2 \le x \le 3{,}0$. Da $20{,}7\%$ af de gamle tal er i intervallet $0{,}6 \le x \le 1{,}5$, er $20{,}7\%$ af de nye tal i intervallet $1{,}2 \le x \le 3{,}0$. Se figur 4.



Figur 4

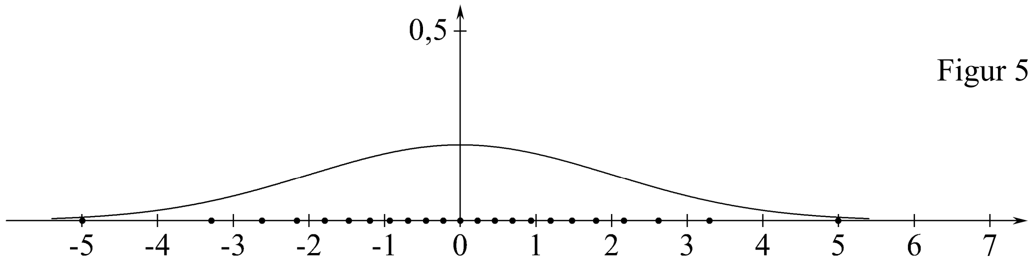
10.3 Normalfordeling. Forskydning af tallene.

Til hvert af de ti mio. tal fra 10.2 lægger vi $1{,}5$:

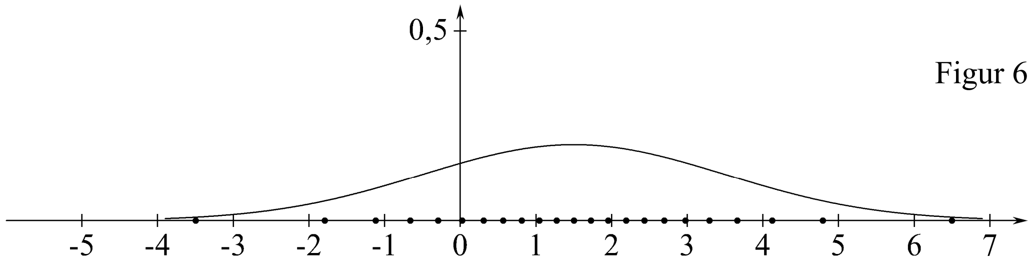
$$1{,}2 + 1{,}5 \;=\; 2{,}7$$
$$3{,}0 + 1{,}5 = \;4{,}5$$

osv.

Når vi lægger $1{,}5$ til tallene på figur 5, så får vi tallene på figur 6.
Grafen bliver forskudt 1,5 enheder mod højre.

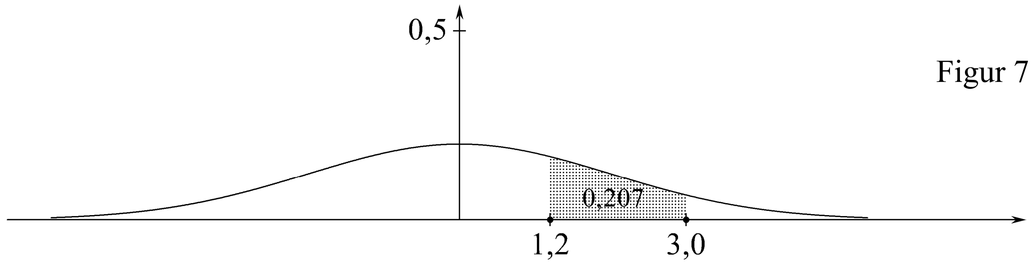


Figur 5

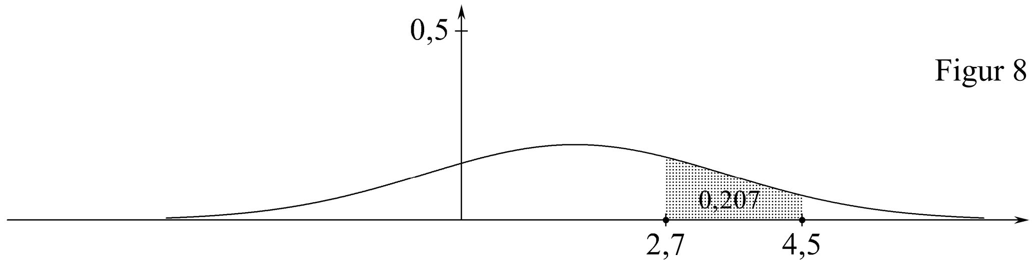


Figur 6

Når vi lægger $1{,}5$ til tallene i intervallet $1{,}2 \leq x \leq 3{,}0$, så bliver disse tal ført over i intervallet $2{,}7 \leq x \leq 4{,}5$. Da $20{,}7\,\%$ af de gamle tal er i intervallet $1{,}2 \leq x \leq 3{,}0$, må der være $20{,}7\,\%$ af de nye tal i intervallet $2{,}7 \leq x \leq 4{,}5$. Se figur 7 og 8.



Figur 7



Figur 8

10.4 Normalfordeling. Middelværdi.

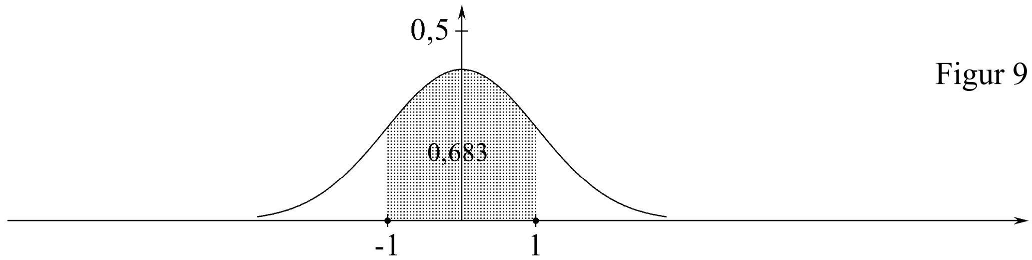
Når nogle tal er normalfordelt, så er deres middelværdi tallet i midten der hvor grafen er højest.
Både de ti mio. tal på figur 1 og de ti mio. tal på figur 5 har middelværdien 0.
De ti mio. tal på figur 6 har middelværdien $1{,}5$.

10.5 Normalfordeling. Spredning.

Vi vil udregne hvor mange procent af de ti mio. tal fra 10.1 der ligger i intervallet $-1 \leq x \leq 1$.
Vi skal altså udregne det grå areal på figur 9. Når f er funktionen fra 10.1 gælder at

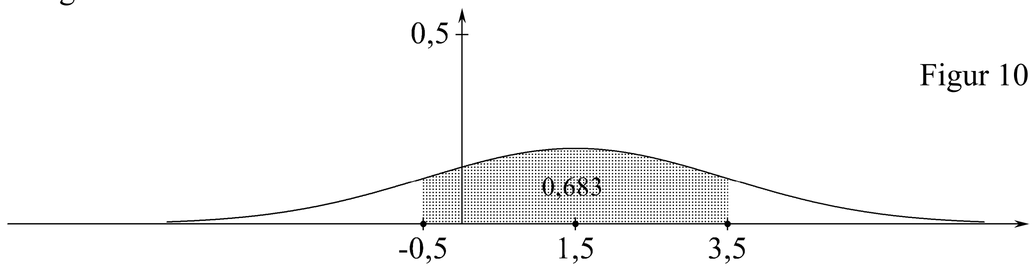
$$\text{gråt areal på figur 9} = \int_{-1}^{1} f(x)\,dx = 0{,}683$$

Så må 68,3 % af de ti mio. tal fra 10.1 ligge i intervallet $-1 \leq x \leq 1$



Figur 9

For hvert af de ti mio. tal på figur 1 gjorde vi følgende: Vi gangede tallet med 2 og lagde 1,5 til resultatet. Derved føres tallene på figur 1 over i tallene på figur 6. Tallene i intervallet $-1 \leq x \leq 1$ føres over i intervallet $-0{,}5 \leq x \leq 2{,}5$, så i dette interval ligger 68,3 % af de nye tal. Se figur 10.



Figur 10

De ti mio. tal på figur 6 ligger mere spredt end de ti mio. tal på figur 1. Vi angiver tallenes spredning med ét tal. Følgende to eksempler viser hvordan dette tal er bestemt for normalfordelinger:

> <u>På figur 10 er tallenes spredning 2</u> fordi vi skal gå 2 enheder ud på begge sider af middelværdien for at få 68,3 % af tallene med.
>
> <u>På figur 9 er tallenes spredning 1</u> fordi vi skal gå 1 enhed ud på begge sider af middelværdien for at få 68,3 % af tallene med.

10.6 Normalfordeling. Hvilke fordelinger er normalfordelinger?

Ovenfor startede vi med nogle tal der var fordelt som angivet med funktionen $f(x)$ fra 10.1 . Derefter gangede vi disse tal med 2 og lagde 1,5 til resultatet. Hvis vi i stedet for 2 og 1,5 bruger andre tal, så får vi en anden fordeling af samme type. De fordelinger vi kan få på denne måde kaldes normalfordelinger.

10.7 Normalfordeling. Mange steder ude i virkeligheden er tal normalfordelt.

Tal fra virkeligheden er ofte normalfordelt:

- En maskine fylder sukker i poser. Når vi vejer poserne, får vi en række tal der er normalfordelt.
- Når vi måler højderne af drengene i 3.g, får vi en række tal der er normalfordelt.
- Når vi mange gange måler højden af samme flagstang, så får vi en række tal der er normalfordelt.
- Tiderne i et løb med mange deltagere er normalfordelt.

10.8 Normalfordeling. Udregning.

Symbolet

normCdf(9,12,10,3)

betyder

den procentdel af tallene der ligger mellem 9 og 12
når tallene er normalfordelt med middelværdi 10 og spredning 3.

Vores regnetekniske hjælpemiddel udregner at

normCdf(9,12,10,3) = 0,378

så 37,8 % af tallene ligger mellem 9 og 12.

10.9 Normalfordeling. En anvendelse.

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

Vores regnetekniske hjælpemiddel udregner at

normCdf(-∞,75,76,0.4) = 0,00621

så 0,6 % af flaskerne indeholder under 75 centiliter.

11.1 χ^2-fordeling. Tæthedsfunktion.

Nogle tal er normalfordelt med middelværdi 0 og spredning 1.
På figur 11 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:

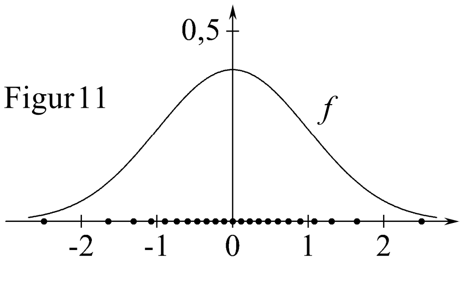
$(-0{,}6)^2 = 0{,}36$

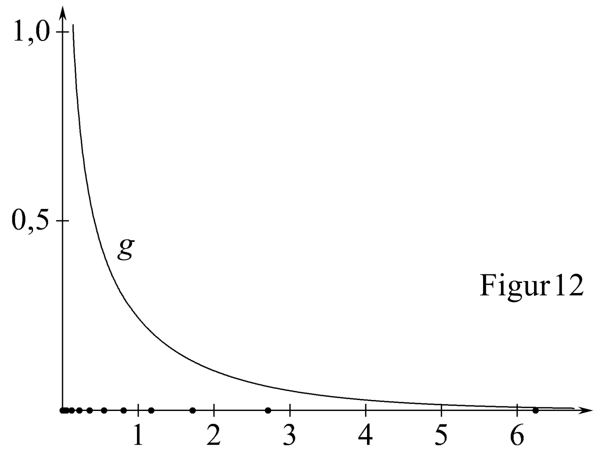
$2{,}5^2 = 6{,}25$

osv.

De tal vi får på denne måde, er fordelt som antydet på figur 12. Vi vil finde forskriften for tæthedsfunktionen g for disse tal. (Kun få af tallene er vist som prikker).

Figur 11





Figur 12

11.2 χ^2-fordeling. Arealfunktion.

Arealfunktionerne for f og g kalder vi F og G. (f og g er funktionerne fra 11.1).

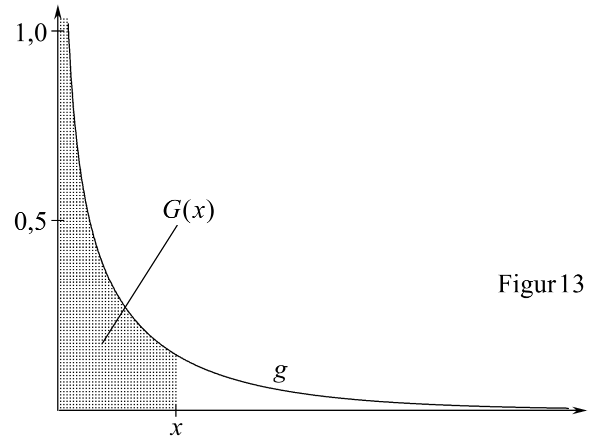
Når x er et positivt tal, gælder altså at

(1) $G(x)$ = gråt areal på figur 13.

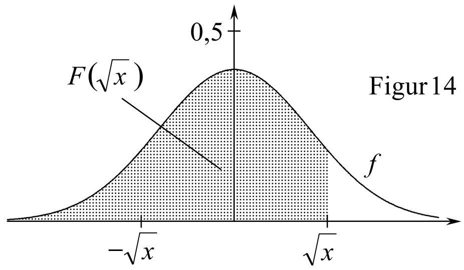
Husk at gråt areal på figur 13 er den procentdel af alle tallene som ligger i intervallet $0 \le t \le x$.

De tal der føres over i intervallet $0 \le t \le x$ når vi opløfter til anden, er tallene i intervallet $-\sqrt{x} \le t \le \sqrt{x}$, så den procentdel af de nye tal som ligger i intervallet $0 \le t \le x$, er lig den procentdel af de gamle tal som ligger i intervallet $-\sqrt{x} \le t \le \sqrt{x}$. Dette betyder at

gråt areal på figur 13 = gråt areal på figur 16 .



Figur 13



Figur 14

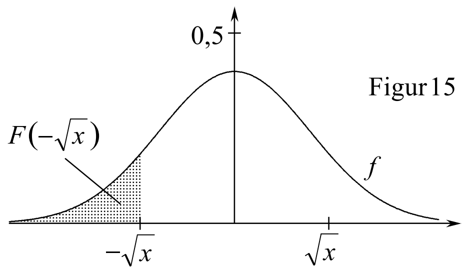
Af dette og (1) får vi

(2) $G(x)$ = gråt areal på figur 16.

Da F er arealfunktionen for f, er

$$F(\sqrt{x}) = \text{gråt areal på figur 14 .}$$

$$F(-\sqrt{x}) = \text{gråt areal på figur 15 .}$$



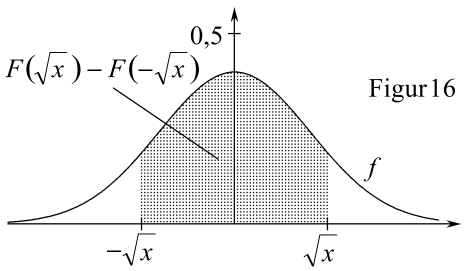
Figur 15

Når vi fra gråt areal på figur 14 fjerner gråt areal på figur 15, så får vi gråt areal på figur 16, så

$$F(\sqrt{x}) - F(-\sqrt{x}) = \text{gråt areal på figur 16 .}$$

Af dette og (2) får vi

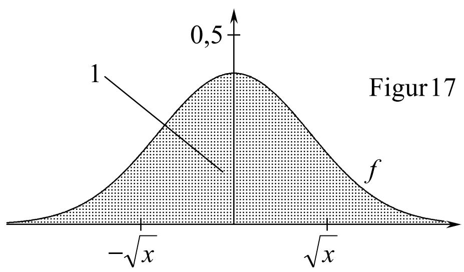
(3) $G(x) = F(\sqrt{x}) - F(-\sqrt{x})$.



Figur 16

Det grå areal på figur 17 er 1 (=100 %).
Når vi fra dette fjerner gråt areal på figur 14, får vi gråt areal på figur 18, så

$$1 - F(\sqrt{x}) = \text{gråt areal på figur 18.}$$



Figur 17

Da grafen er symmetrisk, er de to grå områder på figur 19 lige store, så

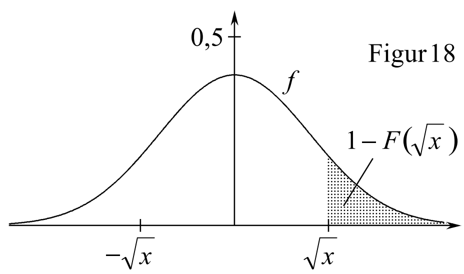
(4) $F(-\sqrt{x}) = 1 - F(\sqrt{x})$

Af (3) og (4) får vi

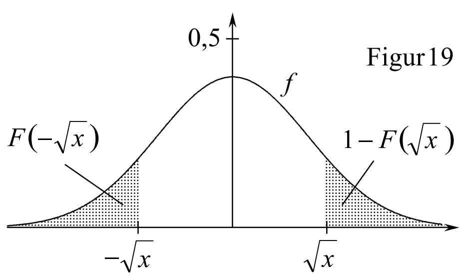
$$G(x) = F(\sqrt{x}) - \left(1 - F(\sqrt{x})\right)$$

Vi omskriver højresiden og får

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



Figur 18



Figur 19

11.3 χ^2-fordeling. Forskrift for tæthedsfunktion.

Formlen (5) fra 11.2 kan vi også skrive sådan:

$$G(x) = 2F(x^{\frac{1}{2}}) - 1$$

Da G er arealfunktionen for g, er

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

dvs.

$$g(x) = \left(2F(x^{\frac{1}{2}}) - 1\right)'$$

Af reglen for at differentiere sammensat funktion får vi $\left(F(x^{\frac{1}{2}})\right)' = F'(x^{\frac{1}{2}}) \cdot (x^{\frac{1}{2}})'$, så

$$g(x) = 2F'(x^{\frac{1}{2}}) \cdot \tfrac{1}{2} x^{-\frac{1}{2}}$$

Dette kan også vi også skrive sådan:

$$g(x) = x^{-\frac{1}{2}} \cdot f(x^{\frac{1}{2}})$$

Dette omskriver vi ved at bruge forskriften for f som står i i 10.1 . Vi får:

$$(6) \quad g(x) = \frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} \cdot \mathrm{e}^{-\frac{x}{2}} \quad , \quad x > 0 .$$

Når nogle tal er fordelt sådan at dette er tæthedsfunktionen, så siger vi at tallene er χ^2-fordelt med 1 frihedsgrad.

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

I (6) 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

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

Vi får vores regnetekniske hjælpemiddel til at løse ligningen

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

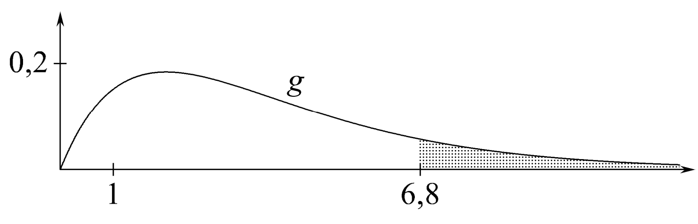
mht. k og får $k = \frac{1}{4}$, så

$$(7) \quad g(x) = \tfrac{1}{4} \cdot x \cdot \mathrm{e}^{-\frac{x}{2}} \quad , \quad x > 0$$

er tæthedsfunktionen for tal der er χ^2-fordelt med 4 frihedsgrader.

11.5 χ^2-fordeling. Test og tæthedsfunktion.

Figuren viser grafen for funktionen (7) fra 11.4 .
Hvis vi i en χ^2-test med 4 frihedsgrader får at χ^2-teststørrelsen er 6,8 , så er p-værdien det grå areal:



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

Normalt regner vi dette tal ud sådan:

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

Stikordsregister