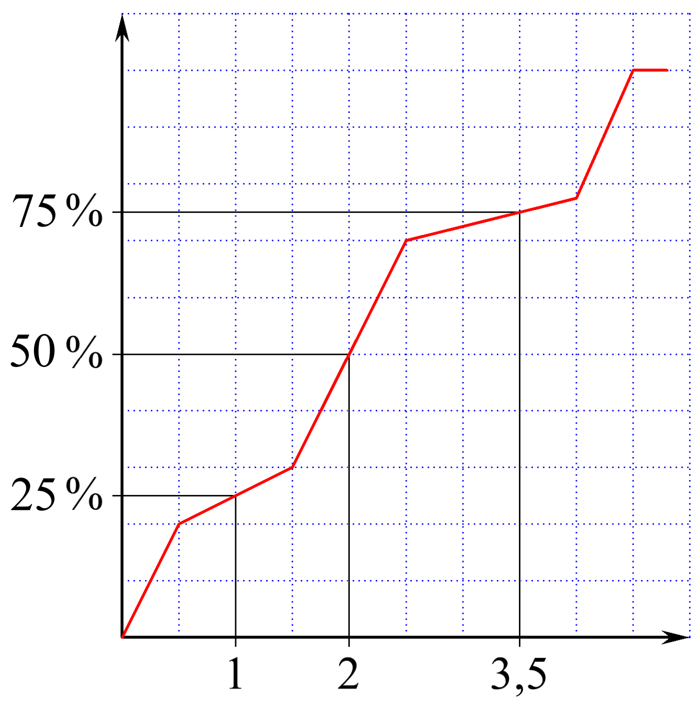
Statistik

for
gymnasiet og hf



2015 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.

Tidligere udgaver af dette hæfte har skiftet adresser til
http://mat1.dk/statistik_for_gymnasiet_og_hf_2011.pdf
http://mat1.dk/statistik_for_gymnasiet_og_hf_2013.pdf
http://mat1.dk/statistik_for_gymnasiet_og_hf_2014.pdf

18/5-2015

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

DESKRIPTIV STATISTIK

TEST

FORDELINGER

Normalfordeling

χ^2-fordeling

DESKRIPTIV STATISTIK

Hvad er deskriptiv tatistik?

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

Ugrupperede data

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

For grupperede data skal vi gøre noget andet. Se afsnit 3.7-3.8 på side 9.

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 <u>median</u> 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 <u>median</u> 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:

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

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.

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

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 <u>nedre kvartil</u>.
Dvs. nedre kvartil er 27.

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

Når vi taler om <u>kvartilsættet</u> 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 <u>nedre kvartil</u>.
Dvs. nedre kvartil er 3,5 .

Medianen for højre halvdel af tallene kalder vi <u>øvre kvartil</u>.
Dvs. øvre kvartil er 7 .

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

2.4 Hvordan tegner vi boksplot?

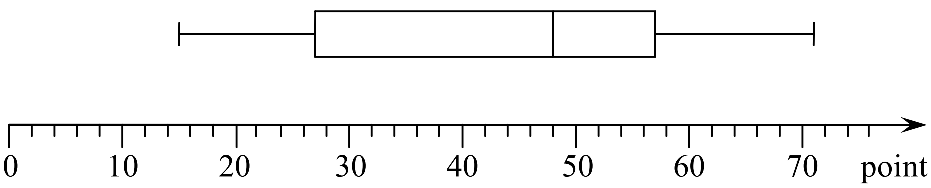
Ved at undersøge datasættet

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 <u>boksplot</u>.



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

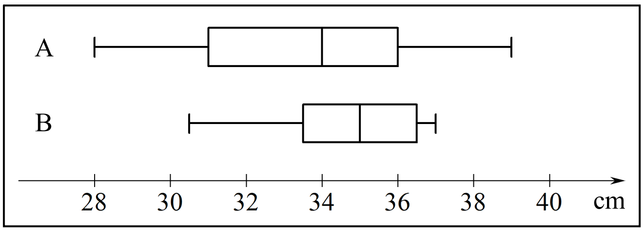
Rektanglet anskueliggør at den midterste halvdel af tallene ligger i intervallet fra 27 til 57.

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

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

2.5 Hvordan sammenligner vi boksplot?

2.51 Opgave



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

Sammenlign højderne på A og B.

Svar

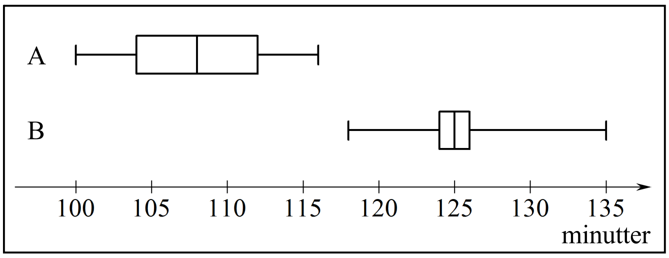
Sammenlign størrelser

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

Sammenlign spredning

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

2.52 Opgave



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

Sammenlign tiderne for A og B.

Svar

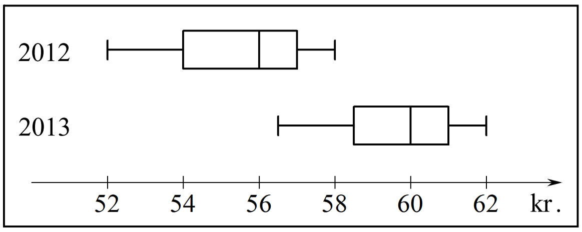
Sammenlign størrelser

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

Sammenlign spredning

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

2.53 Opgave



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

(a) Sammenlign priserne i 2012 og 2013.

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

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

Svar på (a)

Sammenlign størrelser

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

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

Sammenlign spredning

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

Svar på (b)

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

Svar på (c)

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

Grupperede data

3.1 Hvordan tegner vi et histogram?

Tabellen viser fordelingen af nogle frugters vægt.

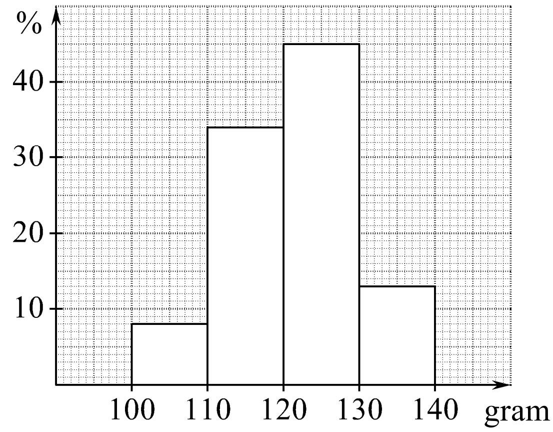
Vægt i gram	100-110	110-120	120-130	130-140
Procent	8	34	45	13

Histogrammet til højre viser oplysningerne i tabellen.

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

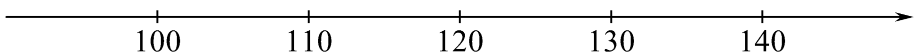
Dette viser at 8 % af frugterne vejer mellem 100 og 110 gram.

<u>Bemærk:</u> Denne måde at tegne et histogram på kan kun bruges fordi intervallerne 100-110, 110-120 osv. er lige lange. Du skal kun kende denne måde.

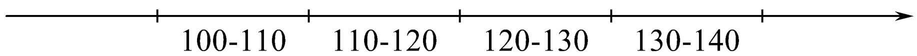


<u>Advarsel:</u> Den vandrette akse skal tegnes som en sædvanlig tallinje.

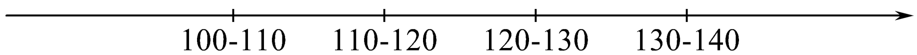
RIGTIGT:



FORKERT:



FORKERT:



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

Ovenfor har vi set på følgende <u>grupperede datasæt</u>:

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.

Dette er ikke i modstrid med virkeligheden, for når vi siger at noget vejer 110 g, mener vi ca. 110 g. Hvis vi hermed mener "mellem 109 g og 111 g", så er der ifølge tabellen 4,2 % der vejer ca. 110 g. Til eksamen plejer man ikke at spørge om sådan noget.

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 en opgave hvor du selv skal gruppere nogle data. Se afsnit 4.1 side 9.

3.3 Hvordan tegner vi en sumkurve?

3.31 Kumuleret frekvens og sumkurve

Den **kumulerede frekvens** af et tal t er den procentdel af dataene der er af størrelse t eller derunder. **Sumkurven** er grafen for den kumulerede frekvens.

3.32 Hvis der er oplyst procent for hvert interval.

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 %

For at tegne en sumkurve, udregner vi kumulerede frekvenser.
Vi har skrevet dem i tabellen, og vi har udregnet dem sådan:
$8\% + 34\% = 42\%$, $42\% + 45\% = 87\%$, osv.

Vægt i gram	100	110	120	130	140
Kumuleret frekvens	0 %	8 %	42 %	87 %	100 %

For at tegne sumkurven gør vi sådan:

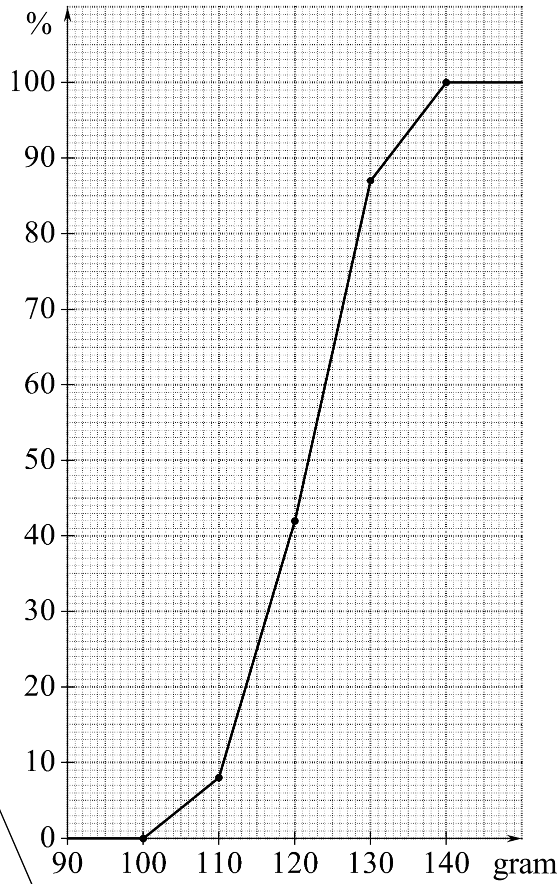
0 % er mindre end 100, så ved $x = 100$ afsætter vi et punkt ud for 0 % på y-aksen.

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. begrundelse for dette i afsnit 5 på side 10).



3.33 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 (m)	0,5-2	2-3	3-4	4-5	5-8
Antal rør	34	58	91	72	27

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

Nedenfor lægger vi sammen før vi omregner til procent. Det er for at undgå mellemfacitter med mange cifre.
Antal data er $34 + 58 + +91 + 72 + 27 \;=\; 282$.
Kumuleret hyppighed udregner vi sådan:
$34 + 58 \;=\; 92$, $92 + 91 \;=\; 183$, osv.
Kumuleret frekvens udregner vi sådan:

$\frac{34}{282} \;=\; 0{,}120567$, $\frac{92}{282} \;=\; 0{,}326241$, osv.

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

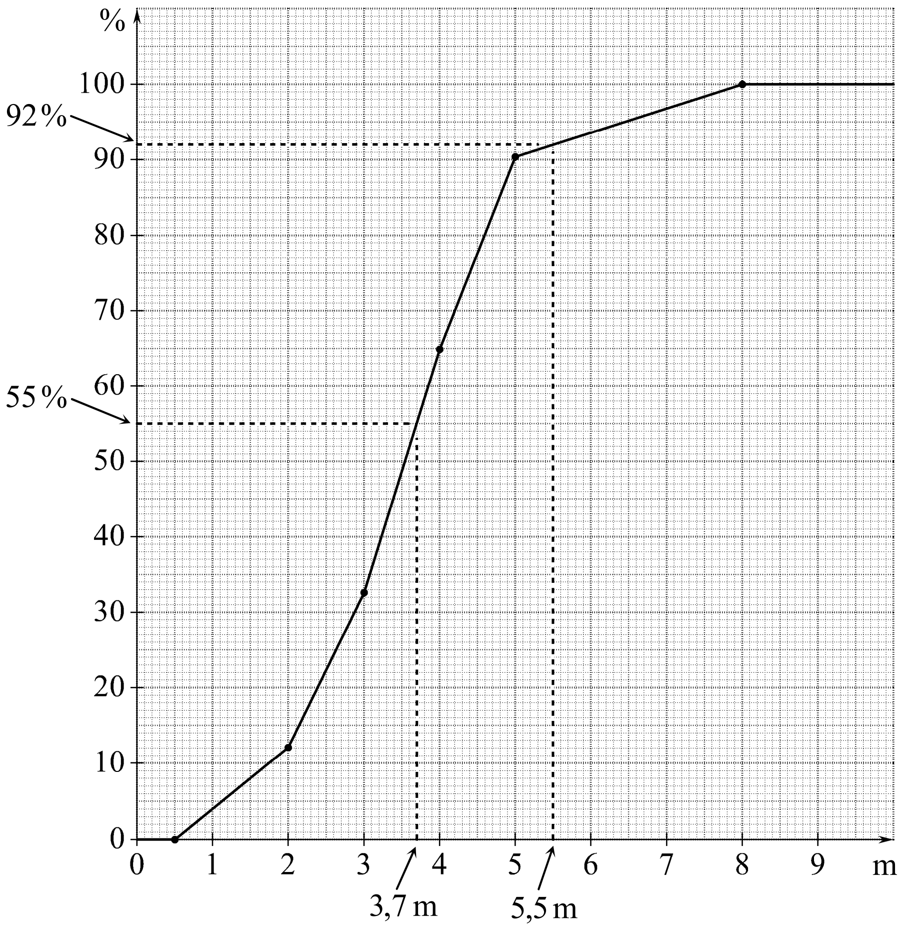
3.34 Hvis start-tal og/eller slut-tal mangler.

Længde (m)	-2	2-3	3-4	4-5	5-
Hyppighed	34	58	91	72	27

I denne tabel mangler start-tal og slut-tal. Så skal vi kun tegne kurven i de andre intervaller. Kurven består altså af tre linjestykker. Spørgsmålene i opgaven vil kunne besvares ved hjælp af den del af kurven vi har tegnet.

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 3.2 på side 5).

3.5 Hvordan finder vi medianen for grupperede data?

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

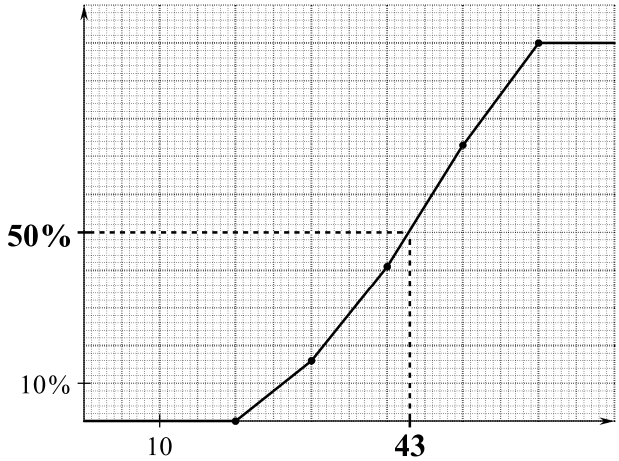
For at finde medianen skal vi bruge sumkurven når det er grupperede data.

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.

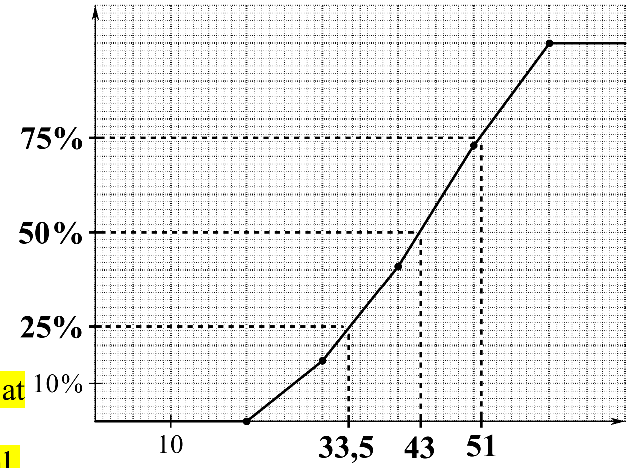


Dette har du brug for at vide når du har fundet medianen og skal svare på hvad dette tal fortæller. I dit svar skal du i stedet for "data" skrive det ord der står i opgaven, f.eks. "længde", og i stedet for "dette tal" skal du skrive det tal du har fundet, f.eks. 43.

3.6 Hvordan finder vi kvartilsættet for grupperede data?

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

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



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 .

Dette har du brug for at vide når du har fundet nedre kvartil og skal svare på hvad dette tal fortæller. I dit svar skal du i stedet for "data" skrive det ord der står i opgaven, f.eks. "længde", og i stedet for "dette tal" skal du skrive det tal du har fundet, f.eks. 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.

Dette har du brug for at vide når du har fundet øvre kvartil og skal svare på hvad dette tal fortæller. I dit svar skal du i stedet for "data" skrive det ord der står i opgaven, f.eks. "længde", og i stedet for "dette tal" skal du skrive det tal du har fundet, f.eks. 51.

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 Middeltal for grupperede data når antal (hyppighed) er oplyst

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: $\frac{0{,}5+2}{2}=1{,}25$, $\frac{2+3}{2}=2{,}5$, 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 $34+58+91+72+27 = 282$. Middeltallet udregnes sådan:

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

Middeltal for rørs længde er $\underline{\underline{3{,}57\,\mathrm{cm}}}$.

Det er nemmere at gange med 27 end 27 gange at skrive 6,5.

3.8 Middeltal for grupperede data når procent (frekvens) er oplyst

Vi vil udregne middeltallet for følgende grupperede datasæt:

Længde i meter	0,5-2	2-3	3-4	4-5	5-8
Frekvens	12 %	18 %	35 %	25 %	10 %

For at udregne middeltallet forestiller vi os at

de 12 % i første interval alle er lig tallet i midten af dette interval,
de 18 % 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
Frekvens	12 %	18 %	35 %	25 %	10 %

Middeltallet udregnes sådan:

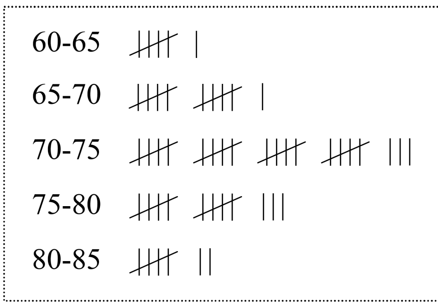
$$\frac{1{,}25\cdot 12+2{,}5\cdot 18+3{,}5\cdot 35+4{,}5\cdot 25+6{,}5\cdot 10}{100} = 3{,}6$$

Middeltal for rørs længde er $\underline{\underline{3{,}6\,\mathrm{cm}}}$.

4.1 Hvordan grupperer vi data?

Vi har modtaget et datasæt på 60 tal (enhed er mm):

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



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 har vi skrevet disse 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.

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

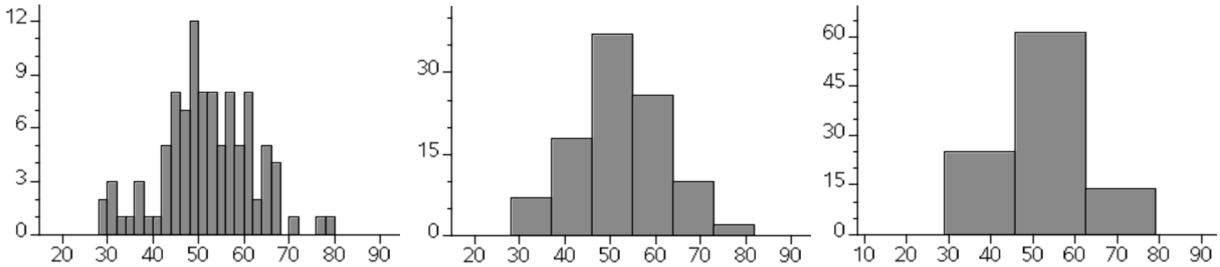
Dette er ikke eneste måde at gøre det på, men der er tradition for at bruge denne måde i det danske gymnasium og hf.

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 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.

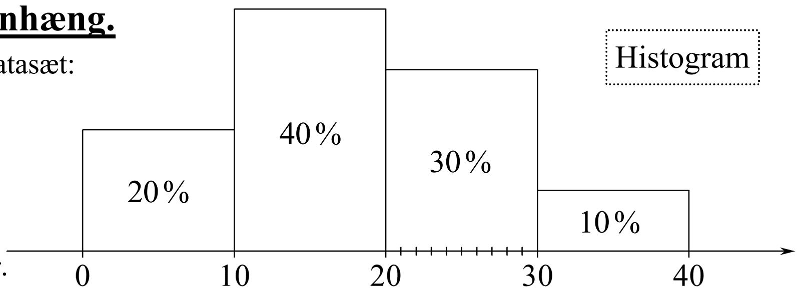


Venstre figur: For lille intervalbredde. Få data i hvert interval får højde til at svinge tilfældigt.

Midterste figur: Intervallernes bredde er passende.

Nederste figur: For stor intervalbredde. Unødig forenklet beskrivelse af hvordan data er fordelt.

5 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

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

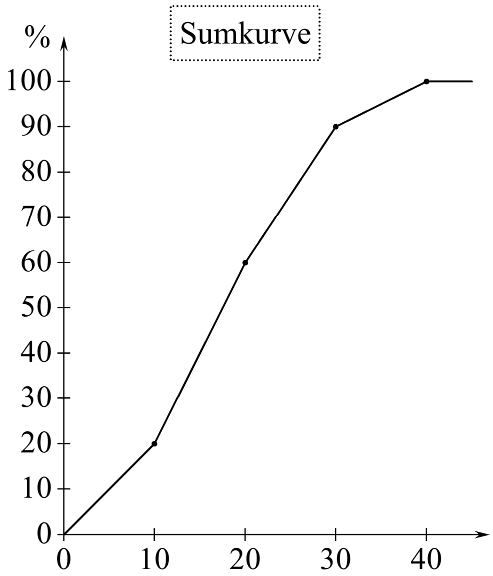
Vi udregner b :

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 .

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: 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 brikkerne er gule*.

Vi tager en stikprøve for at teste hypotesen.
Vi får en stikprøve der ligger langt fra hypotesen.
Vi udregner at hvis hypotesen er rigtig,

så er sandsynligheden kun 3%
for at få en stikprøve der ligger så langt eller længere fra hypotesen,

Hvis 3% er mindre end det valgte signifikansniveau, så forkaster vi hypotesen.
Hvis vi har valgt signifikansniveau = **5%** , så forkaster vi da $3\% < \mathbf{5\%}$.
Hvis vi har valgt signifikansniveau = **1%** , så forkaster vi ikke da $3\% \geq \mathbf{1\%}$.

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 og lærebøger) 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.

8.3 Tankegangen bag det at forkaste eller ikke forkaste en hypotese.

Antag at vi har fundet ud af følgende:
Hvis det hypotese A siger, er rigtigt, så er det usandsynligt at få en stikprøve der ser ud som den vi har fået.
Så tror vi ikke at det hypotese A siger, er rigtigt. Vi forkaster hypotese A.

Antag at vi har fundet ud af følgende:
Hvis det hypotese A siger, er rigtigt, så er det IKKE usandsynligt at få en stikprøve der ser ud som den vi har fået.
Så forkaster vi ikke A.

Hvis vi ikke forkaster A, er det så sandsynligt at A er rigtig? NEJ, for der er altid flere hypoteser B, C, D ... hvorom der gælder:
Hvis det hypotesen siger, er rigtigt, så er det IKKE usandsynligt at få en stikprøve der ser ud som den vi har fået.

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

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

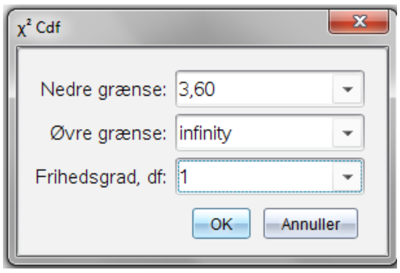
Ved at lægge disse fire tal sammen får vi tallet χ^2 som er <u>afstanden mellem de faktiske tal og de forventede tal</u>:

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

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

9.3 Sådan udregner vi *p*.

Ovenfor udregnede vi at afstanden mellem faktiske og forventede tal er $\chi^2 = 3{,}60$.
I beregningsmenuen vælger vi Statistik / Fordelinger / χ^2 Cdf... og udfylder sådan:



Tallet 1 er antal frihedsgrader. Se afsnit 10.

Vi kan skrive:

Nspire: $\chi^2\text{Cdf}(3.6,\infty,1) \blacktriangleright 0.05778$

Dvs. $p = 5{,}8\,\%$

hvor p er sandsynligheden hvis hypotesen er rigtig,
for at afstanden χ^2 er 3,60 eller større
når antal frihedsgrader er 1.

9.4 Sådan skriver vi konklusionen.

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

<u>Stikprøven giver ikke belæg for at hævde at andelen der siger ja,</u>

<u>er forskellig for piger og drenge.</u>

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 <u>udregnet under den forudsætning at hypotesen er rigtig</u>
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.

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.

10.2 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.

10.3 Frihedsgrader for m gange n tabel.

Hvis der i test for uafhængighed er

m rækker og n søjler,

så er

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

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

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 $\chi^2 = 6{,}31$.

Antal frihedsgrader = (antal rækker − 1) · (antal søjler − 1) = (2–1) · (3–1) = 2

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

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

Nspire: χ²Cdf(6.31,∞,2) ▸ 0.042638

Dvs. $p = 4{,}3\,\%$

hvor p er sandsynligheden hvis hypotesen er rigtig,
for at afstanden χ^2 er 6,31 eller større
når antal frihedsgrader er 2.

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.

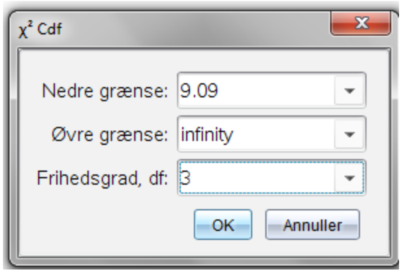
I **test for fordeling** gælder:

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

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

13.4 Sådan udregner vi *p*.

Ovenfor udregnede vi at afstanden mellem faktiske og forventede tal er $\chi^2 = 9{,}09$.
I beregningsmenuen vælger vi Statistik / Fordelinger / χ^2 Cdf... og udfylder sådan:



Vi kan skrive:

Nspire: $\chi^2\text{Cdf}(9.09,\infty,3)$ ▸ 0.028118

Dvs. $p = 2{,}8\,\%$

hvor p er sandsynligheden hvis hypotesen er rigtig,
for at afstanden χ^2 er 9,09 eller større
når antal frihedsgrader er 3.

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 χ²Cdf(α,∞,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

Normalfordeling

16 Tal der er normalfordelt.

Vi måler højden af alle drenge i 3.g .

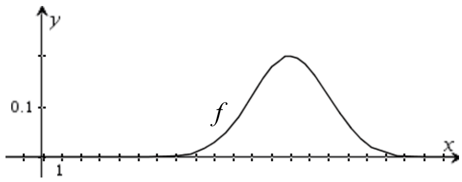
De målte tal afsætter vi som prikker på en tallinje.

Tallene ligger ikke jævnt fordelt. De er fordelt sådan:

Der er et sted på tallinjen hvor tallene ligger tæt.
Jo længere man kommer fra dette sted, jo mindre tæt ligger tallene.

Når vi måler ting af samme slags, vil tallene ofte være fordelt på en bestemt måde som vi kalder **normalfordelt.** F.eks. er højderne af drengene i 3.g normalfordelt.

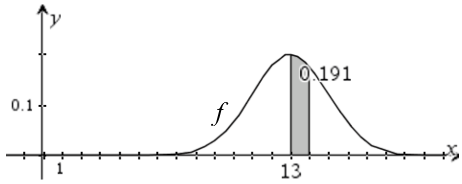
17 Graf for tal der er normalfordelt.



Grafen til venstre viser hvordan nogle målte tal er fordelt.

Hvis vi afsætter tallene som prikker på x-aksen, vil prikkerne ligge tættest der hvor y er størst, dvs. nær 13 på x-aksen ligger prikkerne tættest.

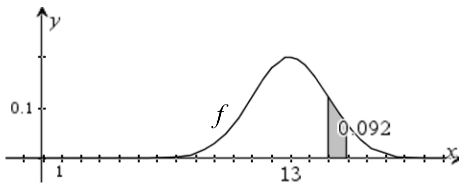
Nær 11 på x-aksen er y mindre, så nær 11 vil prikkerne ligge mindre tæt.



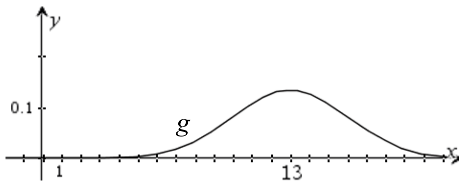
Nær 6 vil prikkerne ligge meget mindre tæt da y er tæt på 0.

Arealet under grafen er $1 = 100\,\%$.

I intervallet $13 \leq x \leq 14$ er arealet under grafen $0{,}191$, dvs. i dette interval ligger $19{,}1\,\%$ af de målte tal.



I intervallet $15 \leq x \leq 16$ er arealet under grafen $0{,}092$, dvs. i dette interval ligger $9{,}2\,\%$ af de målte tal.



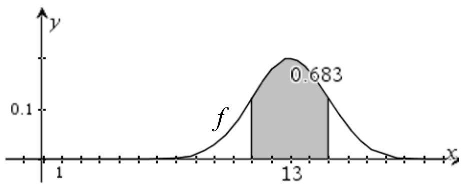
Grafen for g er bredere end grafen for f. Til gengæld er den lavere da arealet under grafen skal være 1.

Når grafen er bredere, ligger tallene mere spredt.

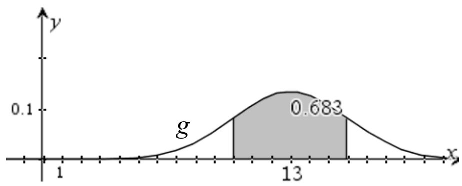
18 Middelværdi for tal der er normalfordelt.

Middelværdien af de målte tal er 13, da grafen er symmetrisk om x=13. Dette gælder både for f og g.

19 Spredning for tal der er normalfordelt.



På f-grafen ser vi at hvis vi fra middelværdien går 2 ud til begge sider, så får vi 68,3 % af tallene med.



Vi siger at tallenes spredning er 2 fordi vi skal gå 2 ud fra middelværdien for at få 68,3 % af tallene med.

På g-grafen ser vi at spredningen er 3.

20 Forskrift for funktionen der viser fordelingen af tal der er normalfordelt.

Når tal er normalfordelt med middelværdi m og spredning s, så vil funktionen f der viser fordelingen, have følgende forskrift:

$$f(x) = \frac{1}{s \cdot \sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-m}{s}\right)^2}$$

Denne forskrift skal du IKKE huske!

21 En anvendelse af normalfordeling.

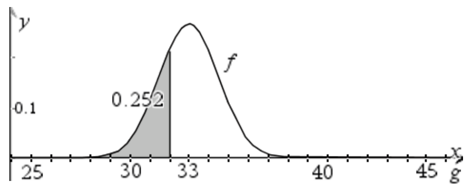
Ved fremstillingen af en vare er der på grund af tilfældigheder stor forskel på vægtene af varerne.
Vægtene er normalfordelt med spredning 1,5 gram.
Varer der vejer under 32 gram skal kasseres.
(Jo tungere varen er, jo dyrere er det at fremstille den).

Opgave: Hvor stor en del af varerne skal kasseres hvis maskinen indstilles til at fremstille varer med middelværdi 33 gram ?

Svar: Lad f være funktionen der viser fordelingen af tal der er normalfordelt med middelværdi $m = 33$ og spredning $s = 1{,}5$.
Arealet under f-grafen i intervallet $-\infty < x < 32$ er

$$\int_{-\infty}^{32} f(x)\,dx = 0{,}252$$ *Udregnet af Nspire.*

så 25,2 % af varerne skal kasseres.

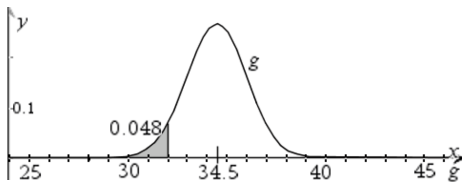


Opgave: Hvor stor en del af varerne skal kasseres hvis maskinen indstilles til at fremstille varer med middelværdi 34,5 gram ?

Svar: Lad g være funktionen der viser fordelingen af tal der er normalfordelt med middelværdi $m = 34{,}5$ og spredning $s = 1{,}5$.
Arealet under g-grafen i intervallet $-\infty < x < 32$ er

$$\int_{-\infty}^{32} g(x)\,dx = 0{,}048$$ *Udregnet af Nspire.*

så 4,8 % af varerne skal kasseres.

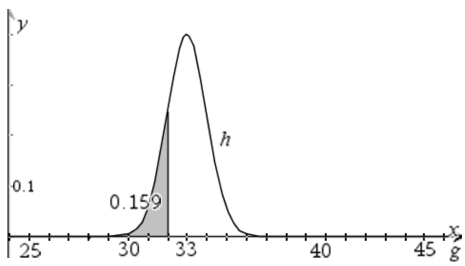


Opgave: En ny maskine laver varer der er normalfordelt med spredning 1 gram. Hvor stor en del af varerne skal kasseres hvis maskinen indstilles til at fremstille varer med middelværdi 33 gram ?

Svar: Lad h være funktionen der viser fordelingen af tal der er normalfordelt med middelværdi $m = 33$ og spredning $s = 1$.
Arealet under h-grafen i intervallet $-\infty < x < 32$ er

$$\int_{-\infty}^{32} h(x)\,dx = 0{,}159$$ *Udregnet af Nspire.*

så 15,9 % af varerne skal kasseres.



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.

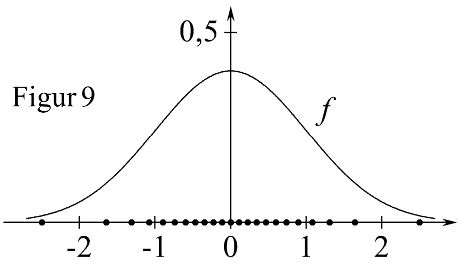
Hvert af tallene opløfter vi til anden:

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

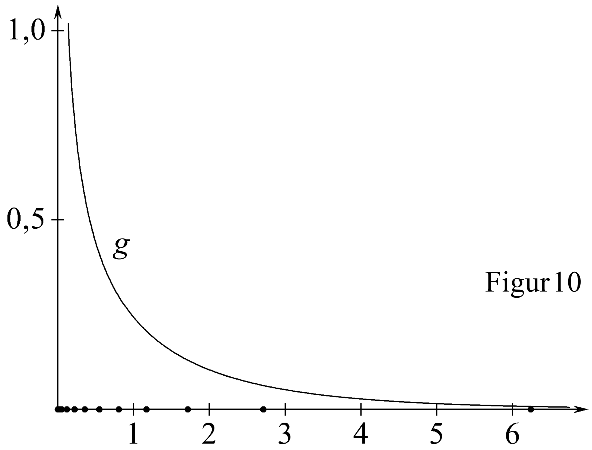
$$2{,}5^2 = 6{,}25$$

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:

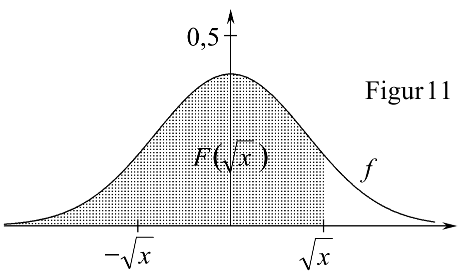
$$p = \int_{3,60}^{\infty} g(x)\,dx = 0{,}058 \ .$$

Her har vi brugt forskriften for g som står i linje (8) nederst i afsnit 23 .

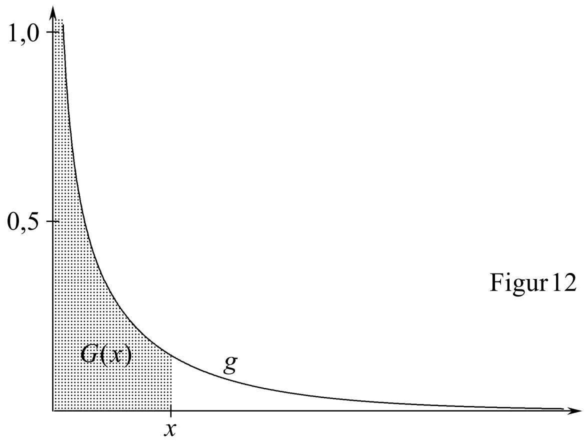
23 Forskrift for g.

f og g er funktionerne fra afsnit 22. 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



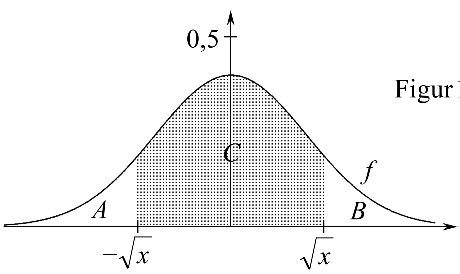
Figur 12

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

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

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

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

Vi reducerer (4) og får:

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

Da G er arealfunktion for g, er

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

Af (5) og (6) får vi

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

Nspire udregner højresiden og får

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

I rammen ses hvordan vi taster for at få (8).
Du skal IKKE kunne denne indtastning.

Define $\mathbf{f}(x)=\dfrac{1}{\sqrt{2\cdot\pi}}\cdot \mathbf{e}^{\frac{-x^2}{2}}$ ▸ *Udført*

fs står for **F**

Define $\mathbf{fs}(x)=\int_{-\infty}^{x} \mathbf{f}(t)\,dt$ ▸ *Udført*

Define $\mathbf{g}(x)=\dfrac{d}{dx}\left(2\cdot \mathbf{fs}(\sqrt{x})-1\right)$ ▸ *Udført*

$\mathbf{g}(x) = \dfrac{\sqrt{2}\cdot \mathbf{e}^{\frac{-x}{2}}}{2\cdot\sqrt{\pi\cdot x}}$ ⚠

$p=\int_{3.6}^{\infty} \mathbf{g}(x)\,dx = 0.05778$

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

Forskriften (8) omskriver vi til

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

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

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

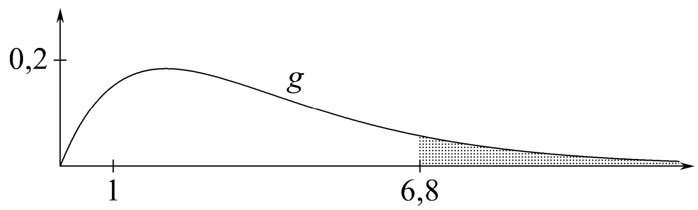
25 χ^2-fordeling og test.

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