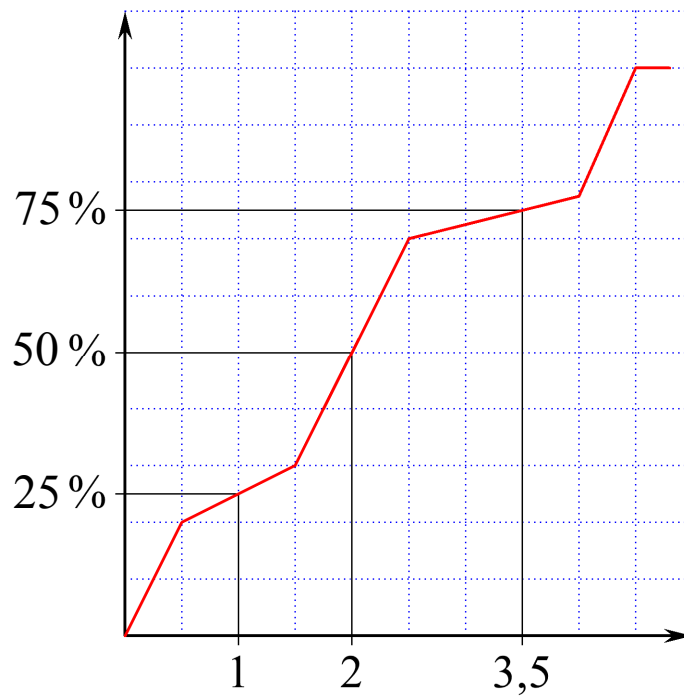


Statistik

for
matematik på B-niveau i hf



2016 Karsten Juul

GRUPPEREDE DATA

1.1	Hvad er deskriptiv statistik?	1
1.2	Hvad er grupperede og ugrupperede data?	1
1.21	Eksempel på ugrupperede data	1
1.22	Eksempel på grupperede data	1
2.1	Hvordan tegner vi et histogram?	1
2.2	Et grupperet datasæt er en model af virkeligheden der er meget forenklet.	2
2.3	Hvordan tegner vi en sumkurve?	3
2.31	Kumuleret frekvens og sumkurve	3
2.32	Hvis der er oplyst procent for hvert interval	3
2.33	Hvis der er oplyst antal for hvert interval	3
2.34	Hvis start-tal og/eller slut-tal mangler	3
2.4	Hvordan aflæser vi på en sumkurve?	4
2.41	Hvor mange procent af rørene er UNDER 3,7 meter?	4
2.42	Hvor mange procent af rørene er OVER 5,5 meter?	4
2.43	Hvor mange procent af rørene er MELLEMLIG 3,7 og 5,5 meter?	4
2.44	Hvor mange procent af rørene er LIG 3,7 meter ELLER DERUNDER ?	4
2.5	Hvordan finder vi medianen for grupperede data?	5
2.6	Hvordan finder vi kvartilsættet for grupperede data?	5
2.61	Nedre kvartil	5
2.62	Øvre kvartil	5
2.63	Kvartilsæt	5
2.7	Middeltal for grupperede data når antal (hyppighed) er oplyst	6
2.8	Middeltal for grupperede data når procent (frekvens) er oplyst	6

TEST

3	Stikprøver	7
3.1	Hvad er populationen?	7
3.2	Hvad er stikprøven?	7
3.3	Systematiske fejl ved valg af stikprøven	7
3.4	Tilfældige fejl ved valg af stikprøven	7
3.5	Er der skjulte variable?	8
4	Hvad er sandsynlighed?	8
4.1	Eksempel	8
4.2	Eksempel	8
5	Test af hypotese	8
5.1	Signifikansniveau	8
5.2	Hvornår har vi vist noget med en test?	8
5.3	Tankegangen bag det at forkaste eller ikke forkaste en hypotese	8
6	Test for uafhængighed i 2x2 tabel	9
6.1	Sådan udregner vi FORVENTEDE TAL	9
6.2	Sådan udregner vi χ^2	10
6.3	Sådan udregner vi p	10
6.4	Sådan skriver vi konklusionen	10
6.5	Misforstå ikke procenterne	10
7	Nulhypotese	11
8	Fordeling af teststørrelse	12

Tidligere udgaver af dette hæfte (med et andet indhold) har skiftet adresser til

[http://mat1.dk/statistik for matematik på b-niveau i hf_2013.pdf](http://mat1.dk/statistik%20for%20matematik%20p%C3%A5%20b-niveau%20i%20hf_2013.pdf)

[http://mat1.dk/statistik for matematik på b-niveau i hf_2014.pdf](http://mat1.dk/statistik%20for%20matematik%20p%C3%A5%20b-niveau%20i%20hf_2014.pdf)

Statistik for matematik på B-niveau i hf © 2016 Karsten Juul

4/4-2016

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

Hæftet må benyttes i undervisningen hvis læreren med det samme sender en e-mail til kj@mat1.dk som oplyser at dette hæfte benyttes (angiv fulde titel og årstal), og oplyser hold, niveau, lærer og skole.

GRUPPEREDE DATA

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.

Der er to slags deskriptiv statistik: "Ugrupperede data" og "Grupperede data".

1.2 Hvad er grupperede og ugrupperede data?

Hvis der er mange forskellige data, så grupperer vi dem i intervaller. (Hvis vi kaster en terning 1000 gange, er der mange data, men kun 6 forskellige, så disse skal ikke grupperes).

1.2.1 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

1.2.2 Eksempel på grupperede data.

Vi har vejlet 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 tegner vi et histogram?

Tabellen viser fordelingen af nogle frugters vægt.

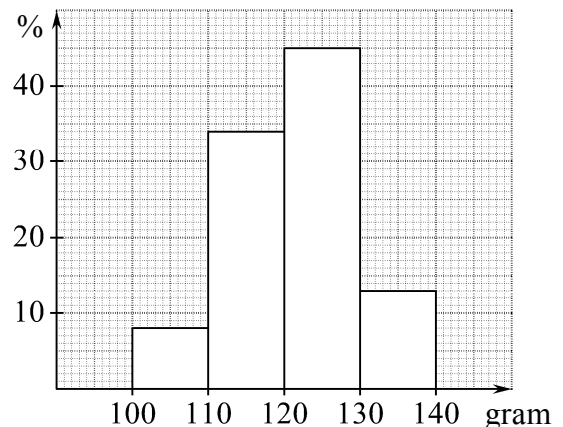
Vægt i gram	100-110	110-120	120-130	130-140
Procent	8	34	45	13

Histogrammet til højre viser oplysningerne i tabellen.

Rektanglet over intervallet 100-110 har højden 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. Du skal kun kende denne måde.



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

RIGTIGT:

FORKERT:

FORKERT:

2.2 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.

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.](#)

2.3 Hvordan tegner vi en sumkurve?

2.31 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.

2.32 Hvis der er oplyst procent for hvert interval.

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

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

For at tegne en sumkurve, udregner vi kumulerede frekvenser.

Vi har skrevet dem i tabellen, og vi har udregnet dem sådan:

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

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

For at tegne sumkurven gør vi sådan:

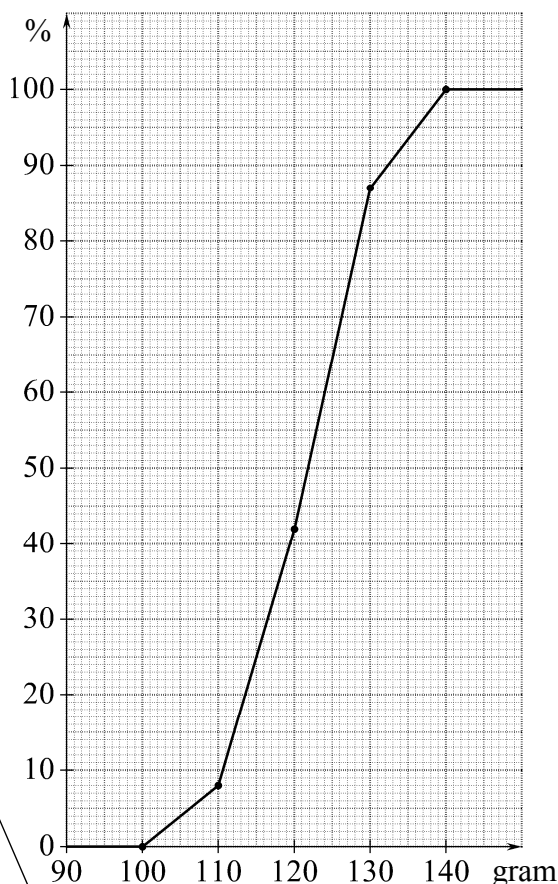
0 % er mindre end 100, så ved $x = 100$ afsætter vi et punkt ud for 0 % på y -aksen.

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).



2.33 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

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.

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

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 %

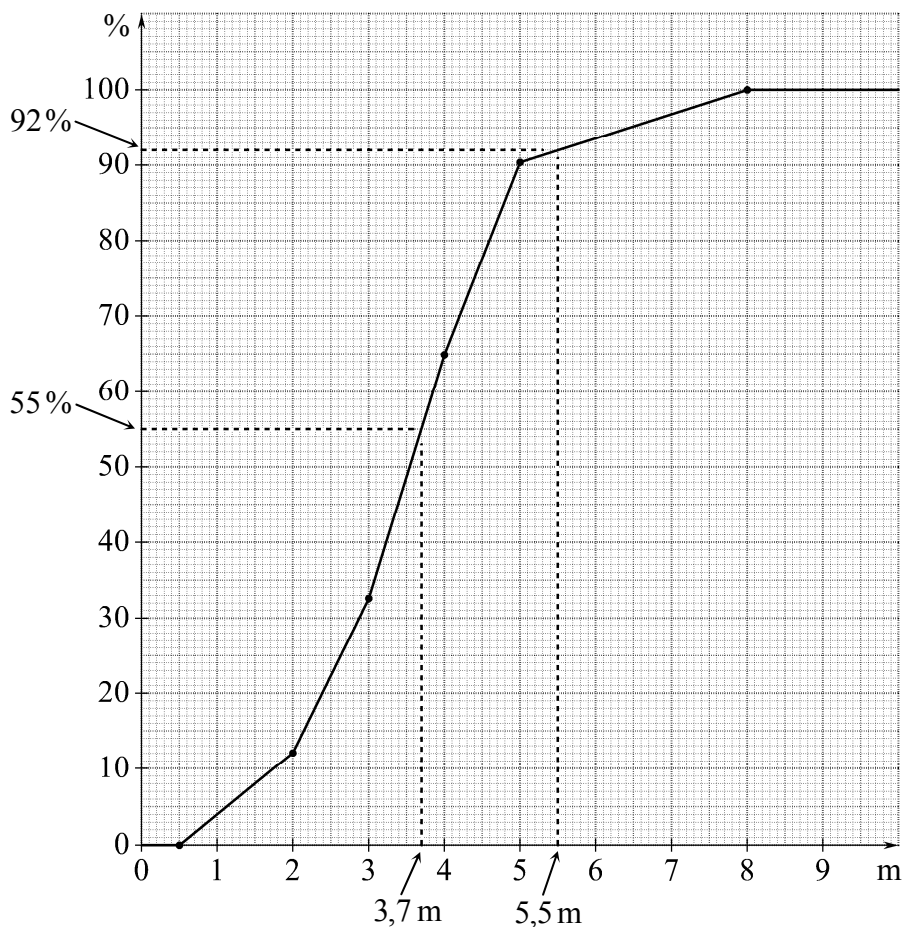
2.34 Hvis start-tal og/eller slut-tal mangler.

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.

2.4 Hvordan aflæser vi på en sumkurve?

Figuren viser sumkurven for rørene fra tabellen på foregående side.



2.41 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.

2.42 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.

2.43 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.

2.44 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... 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).

2.5 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.

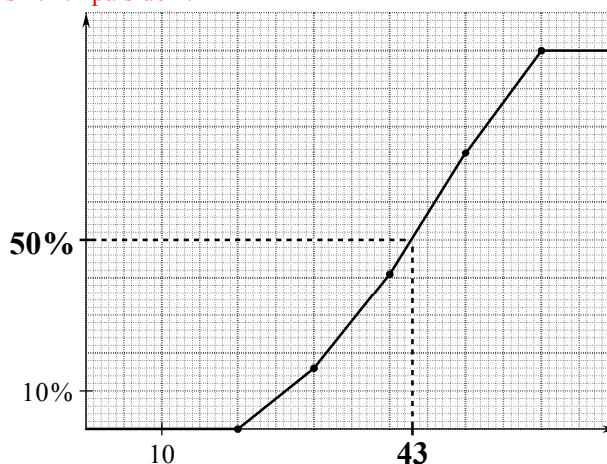
Vi starter i 50% på y-aksen, går vandret hen til sumkurven, går lodret ned på x-aksen, og aflæser x-værdien.

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.



2.6 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.

2.61 Nedre kvartil.

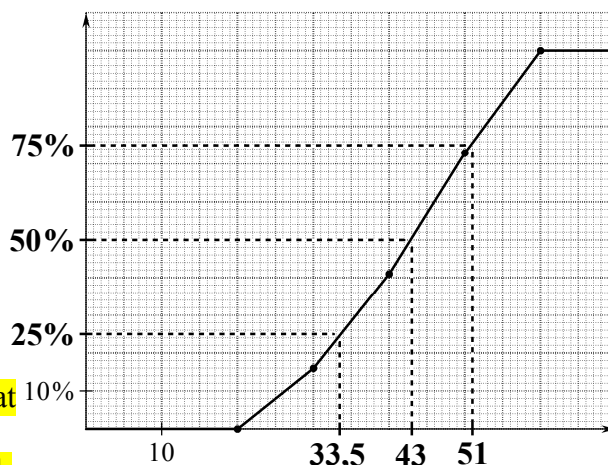
Vi starter i 25% på y-aksen, går vandret hen til sumkurven, går lodret ned på x-aksen, og aflæser x-værdien.

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.



2.62 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.

↑ 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.

2.63 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.

2.7 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:

$$\frac{1,25 \cdot 34 + 2,5 \cdot 58 + 3,5 \cdot 91 + 4,5 \cdot 72 + 6,5 \cdot 27}{282} = 3,56560 \quad \text{Middeltal for rørs længde er } \underline{\underline{3,57 \text{ cm}}} .$$

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

2.8 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:

$$\frac{1,25 \cdot 12 + 2,5 \cdot 18 + 3,5 \cdot 35 + 4,5 \cdot 25 + 6,5 \cdot 10}{100} = 3,6 \quad \text{Middeltal for rørs længde er } \underline{\underline{3,6 \text{ cm}}} .$$

TEST

3 Stikprøver.

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

3.1 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.

3.2 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.

3.3 Systematiske fejl ved valg af stikprøven.

Eksempel 1

Population: Eleverne på vores gymnasium.

Stikprøve: Eleverne i en sproglig klasse.

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.

3.4 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 3 fortsætter på næste side.

3.5 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.

4 Hvad er sandsynlighed?

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

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

5 Test af hypotese.

5.1 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 $3\% < 5\%$.

Hvis vi har valgt signifikansniveau = 1 %, så forkaster vi ikke da $3\% \geq 1\%$.

5.2 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.

5.3 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.

6. 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

6.1 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

6.2 Sådan udregner vi χ^2 .

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

For hvert af de fire felter udregner vi

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

Ved at lægge disse fire tal sammen får vi tallet χ^2 :

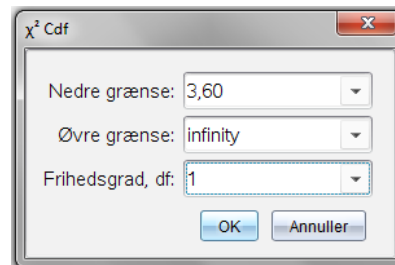
$$\chi^2 = \frac{(87 - 93,40)^2}{93,40} + \frac{(46 - 39,60)^2}{39,60} + \frac{(71 - 64,60)^2}{64,60} + \frac{(21 - 27,40)^2}{27,40}$$
$$\chi^2 = 3,60$$

Teststørrelsen 3,60 er afstanden mellem de faktiske tal og de forventede tal.

6.3 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:



Antal frihedsgrader er altid 1 i den slags test vi behandler her.

Vi kan skrive:

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

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.

6.4 Sådan skriver vi konklusionen.

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

Stikprøven giver ikke belæg for at hævde at andelen der siger ja, er forskellig for piger og drenge.

6.5 Misforstå ikke procenterne.

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

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

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

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

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

7 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.

Elevens holdning til spørgsmålet er uafhængig af køn.

Elevens 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.*

8 Fordeling af teststørrelse

I skemaet er nogle elever inddelt på to måder:

måde 1: pige eller dreng

måde 2: svarer ja eller nej

	ja	nej
pige		
dreng		

De to inddelinger er uafhængige af hinanden

hvis H_0 : sandsynlighed for ja er ens for piger og drenge

dvs. hvis H_0 : brøkdelen af piger der siger ja = brøkdelen af drenge der siger ja.

Man vælger tilfældigt en stikprøve blandt eleverne og udfylder skemaet ovenfor.

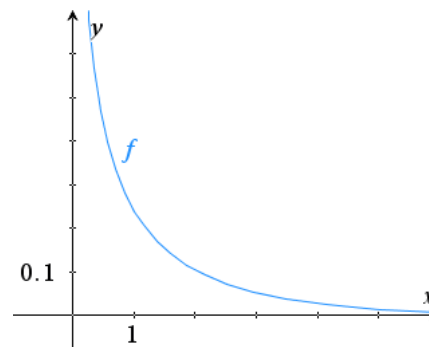
Selv om populationen opfylder brøkdelen af piger der siger ja = brøkdelen af drenge der siger ja, kan vi ikke forvente at dette gælder stikprøven (skemaet). Hvis skemaet er meget langt fra at opfylde dette, så gætter vi på at H_0 er forert. Vi siger at vi forkaster nulhypotesen.

Ud fra stikprøven (skemaet) udregner vi et tal (teststørrelsen χ^2) som er et udtryk fra stikprøves afstand fra hypotese. (Tallet udregnes ved at bruge formlen i afsnit 6.2).

Hvis hypotesen er rigtig og vi mange gange tager en stikprøve og udregner tallet χ^2 , så får vi tal der er fordelt som kurven viser.

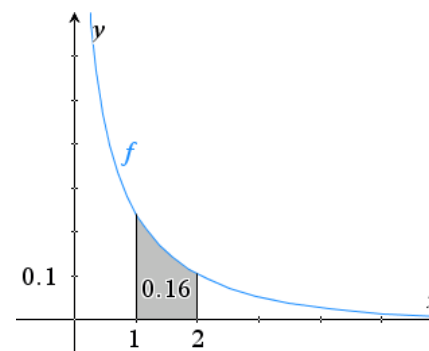
$$f(x) = \frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} \cdot e^{-\frac{x}{2}}, \quad x > 0.$$

Forskriften for f skal du IKKE huske!



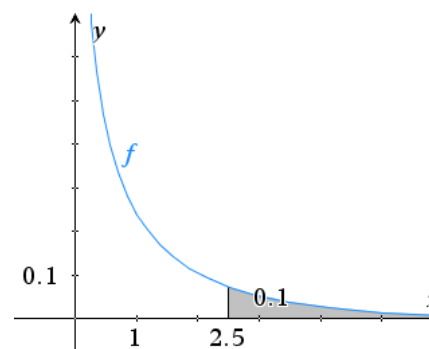
Eksempel: Areal mellem graf og x-akse i intervallet $1 \leq x \leq 2$ er 0,16. Det betyder at 16% af tallene ligger i intervallet. Hele arealet mellem graf og x-akse må være 1 (100%).

$$\int_1^2 f(x) dx = 0,16.$$



Hvis vi får en stikprøve med $\chi^2 = 2,5$, så er p lig sandsynligheden hvis hypotesen er rigtig, for at få en stikprøve hvor χ^2 er 2,5 eller større. På sidste figur ser vi at p er 0,1, altså 10%, så vi forkaster ikke hypotesen hvis signifikansniveau er 5%.

$$\int_{2,5}^{\infty} f(x) dx = 0,1.$$



Stikordsregister

#		
χ^2 -test for uafhængighed	9	
A		
alternativ hypotese	11	
D		
data	1	
deskriptiv statistik	1	
F		
faktiske tal i test for uafhængighed	9	
forkaste hypotese.....	8, 10	
forventede tal i test for uafhængighed.....	9	
frekvens	3	
G		
grupperede data	1, 2	
H		
histogram.....	1	
hypotese	8, 9	
I		
intervals frekvens	3	
K		
kumuleret frekvens.....	3	
kumuleret hyppighed.....	3	
kvartilsæt for grupperede data.....	5	
M		
median for grupperede data.....	5	
middeltal for grupperede data	6	
middelværdi for grupperede data	6	
N		
nedre kvartil for grupperede data.....	5	
nulhypotese	11	
P		
<i>p</i>	7, 10	
population	7, 9	
S		
sandsynlighed.....	8	
signifikansniveau	8	
skjult variabel.....	8	
stikprøve.....	7, 9	
sumkurve, aflæs	4, 5	
sumkurve, tegn når antal oplyst.....	3	
sumkurve, tegn når procent oplyst.....	3	
systematisk fejl	7	
T		
test af hypotese.....	8	
test for uafhængighed.....	9	
teststørrelse	10	
tilfældig fejl.....	7	
U		
uafhængig.....	11	
ugrupperede data.....	1	
Ø		
øvre kvartil for grupperede data.....	5	