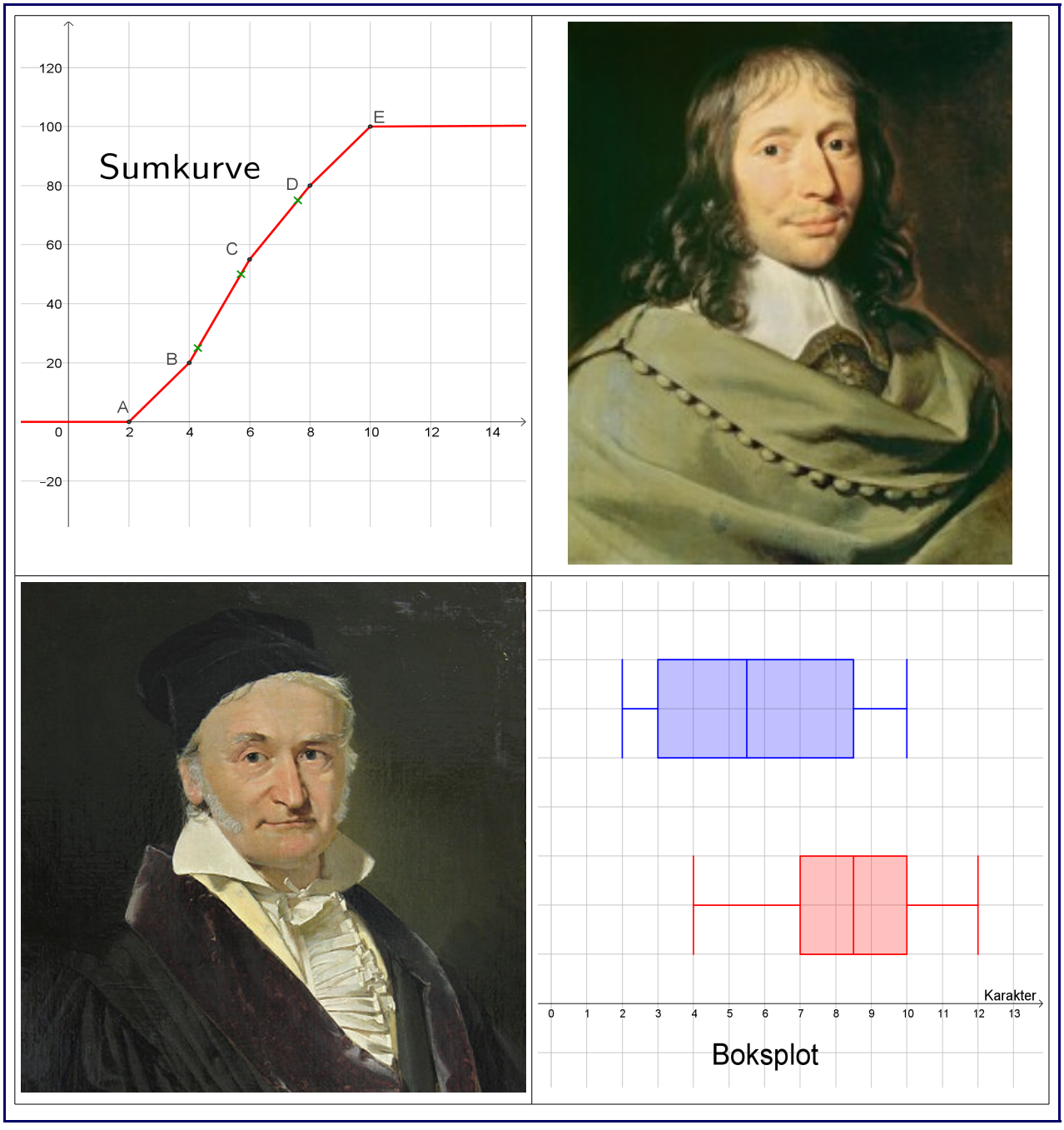
Statistik og sandsynlighedsteori



Ikast 2018

Ib Michelsen

Ib Michelsen

Matematik A: Statistik og Sandsynlighedsteori

Sidst ændret: 30-3-2018

Udskrevet:

C:\Users\IbM\Dropbox\statistik\SS2018\Statistik og Sandsynlighedsteori.odt

59 sider

Indholdsfortegnelse

Introduktion....5
Deskriptiv statistik....7
- Statistik....9
- Ugrupperede data....9
  - Observationer og deskriptorer (Definitioner)....12
  - Boksplot....13
- Grupperede observationer....15
  - Biografgængerne....15
  - Histogrammer og flere deskriptorer....17
  - De kumulerede hyppigheder....18
  - De kumulerede intervalfrekvenser....19
  - Sumkurven....19
  - Fraktiler....20
  - Kvartilsæt mv. for grupperede data....21
  - Middeltal for grupperede data....23
  - Middeltallet for grupperede data....23

Sandsynlighedsteori....25
- Forhistorien....27
- Frekvenser og sandsynligheder....27
- Grundlaget for sandsynlighedsteorien....28
- Binomialfordelingen....30
  - Binomialforsøg....30
  - Binomialfordelingen....30
- Normalfordelingen....35

Estimat, Hypotese og Test....37
- Estimat....39
- Hypoteser....40
- Test....41

X^2-fordelingen....43
- Definition....45
- Anvendelsen....45
- Hvordan....45
  - X2 test....46
  - Goodness of fit....47
- Eksempler på testberegninger....48
  - Opgave 11 4. maj 2013 stxB....48
  - Opgave 11 29. maj 2013 stxB....50
- 7 eksamensopgaver....52
  - Opgave 1....52
  - Opgave 2....53
  - Opgave 3....54
  - Opgave 4....55

Opgave 5....56
Opgave 6....56
Opgave 7....57
Stikordsregister....58

Introduktion

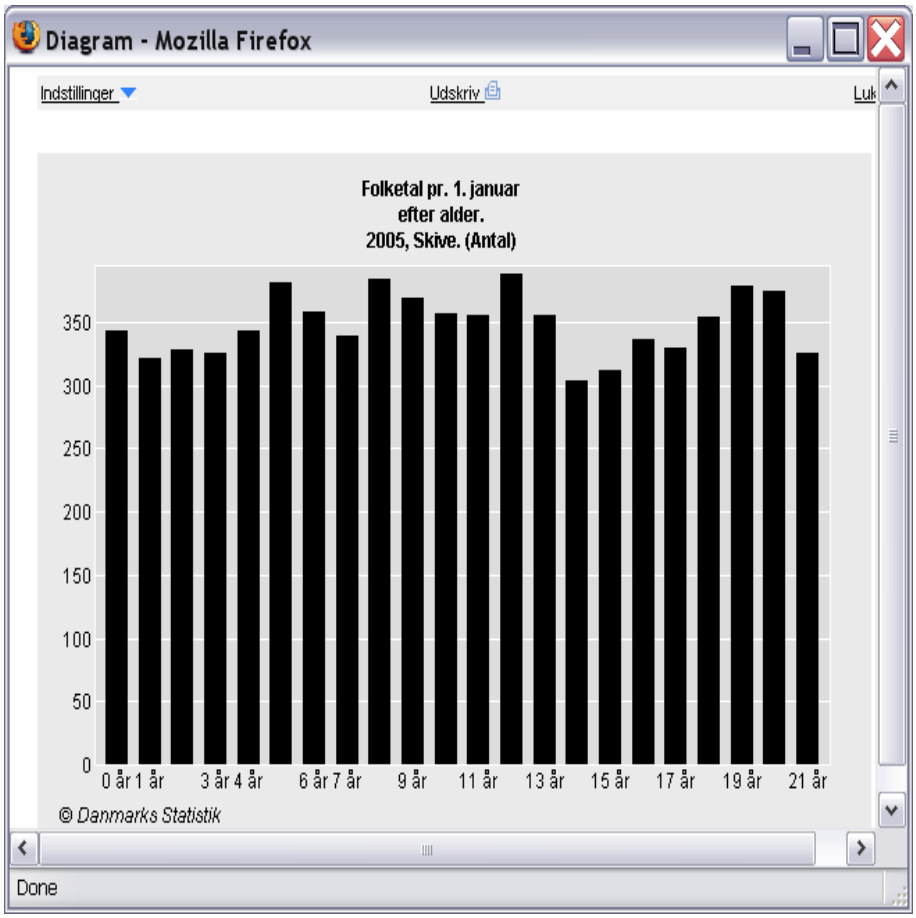
Dette skrift fokuserer på 3 emner: Deskriptiv statistik, sandsynlighedsteori og test.

Deskriptiv statistik drejer sig om at beskrive ofte store talmængder overskueligt med nogle få tal eller en figur.

Sandsynlighed er et begreb hasardspillere har interesseret sig for i hundreder af år. Ved præcisering af matematikken er vi nu blevet langt bedre til at opstille modeller, der ikke kun er af interesse for spillere, men indenfor forsikringvidenskab, kvalitetskontrol, meningundersøgelser, medicinsk forskning, politiske beslutningsmodeller og i bedømmelse af beviser i retssager.

Endelig handler **test** om at vurdere påstande (kaldet *hypoteser*) om *populationer*, dvs. store mængder af *individer* som: mennesker, begivenheder, genstande ... uden kendskab til hvert enkelt element i populationen, men udelukkende til en lille brøkdel af denne: *stikprøven*. Eller mere præcist: Kendskab til værdierne af en *variabel* knyttet til individerne.

Deskriptiv statistik



Statistik

Og hvad betyder så ordet *statistik* her?
Statistik er bl.a. videnskaben om at tilvejebringe statistikker, det vil sige: en oversigt over en stats eller staters indre forhold – især hvad angår talmæssige forhold.[1] I dag bruges ordet "statistikker" dog også om andre talmæssige forhold end statens som for eksempel i "firmaets salgsstatistik". Betydningen af ordet "statistik" er blevet bredere og dækker nu næsten en hvilken som helst (talmæssig) beskrivelse af et fænomen: vejrstatistik, vildtstatistik ...
Talmæssige forhold for en stat består tit af overvældende mange tal. Statistikerens arbejde i dette område er derfor ofte at koncentrere sådanne mængder af tal til et eller nogle få tal eller en overskuelig figur, for at læseren ikke skal drukne i information. Sådanne tal kaldes statistiske *deskriptorer*.
Danmarks Statistik [2] er en statsinstitution, der har til opgave at indsamle og formidle oplysninger om alt mellem himmel og jord. I *Statistikbanken*, der er gratis og offentlig tilgængelig på Internettet, får man et indtryk af bredden ved at se på emneoversigten*: Miljø og energi, Befolkning og valg, Uddannelse og kultur, Arbejdsmarked, Løn, Sociale forhold, Sundhed og retsvæsen, Indkomst, Forbrug og priser, Generel erhvervs-statistik, Landbrug, Industri, Byggeri og boligforhold, Serviceerhverv, Transport, Udenrigshandel, National-regnskab og betalingsbalancen, Offentlige finanser, Penge og kapitalmarked.*
De *data*, der benyttes i beskrivelserne, kan være *ugrupperede* eller *grupperede*. I det første tilfælde kendes størrelsen af hver observation, i det andet tilfælde vides alene, hvor mange *observationer* der falder i et bestemt interval. Tidligere men også i mange situationer i dag vil det være praktisk ved indsamling og behandling af data at gruppere dem. Er der kun få data eller kan *datafangsten* ske med moderne IT-teknologi, kan data være ugrupperede.

Ugrupperede data

Ved en eksamen har Josephine opnået følgende karakterer: 7, 4, 4, 7, 10, 7, 7, 10, 7, 02, 4, 7, 10, 4, 7, 10, 7 , 4, 12, 10. Hvis der som her er tale om 20 enkeltkarakterer, kan det virke lidt uoverskueligt. Derimod synes beskeden:

"Josephine bestod X-eksamen med gennemsnittet 7"

at være en klar besked og ofte lige så god som alle enkeltkaraktererne.
Her vil vi i stedet for ordet *gennemsnit* bruge ordet *middelværdi* eller bedre: *middeltal*. Hver af Josephines karakter er en *observation* og alle karaktererne kaldes *observationssættet*.
Middeltallet beregnes således:

$$Middeltal = \frac{7+4+4+...+4+12+10}{20} = 7{,}0$$

1 jævnfør "Ordbog over det Danske Sprog", Bind 21, Gyldendal, 1943 (1982)
2 http://www.dst.dk

Generelt betegnes den i'te observation x_i , den sidste observation x_N (hvor N er *observationssættets størrelse*) og middeltallet betegnes $\bar{x}$.
Formlen til beregning af middeltallet er derfor:

$$\bar{x} = \frac{\sum_{i=1}^{i=N} x_i}{N} \qquad (1)$$

Ofte vil man også være interesseret i, hvilke karakterer Josephine har fået flest af: dvs. hendes *typiske karakter*.
"Josephine har typisk fået 7"
Når vi har en række observationer, kaldes den (evt. de) observation(er) (her karakter), der er flest af, *typetallet*.

Vi kunne også sortere alle Josephines karakterer i størrelsesorden begyndende med 2, så 4, 4 ... og til sidst 12. Den karakter, der står midt i rækken er *medianen*. Er der et lige antal observationer, benytter vi middeltallet af de to midterste observationer.
"Josephine har karakteren 7 som median"
Både middeltal, typetal og median beskriver Josephines eksamen; de er *deskriptorer*. At de her blev samme tal er ikke sikkert, men at de har omtrent samme værdier er heller ikke unormalt. Hvad der er vigtigt (for os) er, at deskriptoren fortæller det vigtigste uden at vildlede. Det har nok ikke altid været tilfældet ...[3]

Et eksempel

Dette er en undersøgelse af, hvor mange timer en 8. klasse har set tv på et døgn.

Observationer

Nr.	1	2	3	4	5	6	7	8	9	10	11
TVtid	0	0	½	½	½	1	1	1	1	1	1½

Nr.	12	13	14	15	16	17	18	19	20	21	
TVtid	1½	1½	1½	1½	1½	1½	2	2	2	2½	

Eleverne er nummererede (øverst), der er altså 21 elever og nederst står det antal timer, hver af dem har set tv på et døgn.

3 " There are three kinds of lies: lies, damned lies, and statistics." Benjamin Disraeli *British politician (1804 - 1881)* (http://www.quotationspage.com/quote/487.html)

Observationssættets størrelse N=21. Bemærk at observationerne er sorteret således at mindste observation kommer først, største sidst.

Her er **medianen** observation nummer 11 i ovenstående tabel, altså 1½ time. Den midterste observation. Hvis der kun havde været 20 observationer, ville medianen have været middeltallet af observation 10 og 11.
Denne undersøgelse har **typetal** 1½ time, fordi det forekommer flest gange.
Middeltallet er i dette tilfælde:
Middeltallet = (0+0+½+½+½+1+1+1+1+1+1½+1½+1½+1½+1½+1½+1½+2+2+2+2½):21 = 25,5:21=1,21
Middeltallet af børnenes tv-kiggeri er 1,21 timer/døgn

Hyppighedstabel for antal timer

Observation	Hyppighed	Frekvens	Frekvens om-regnet til %
0 timer	2	2/21=0,095	0,095*100=9,5%
½ time	3	3/21=0,143	0,143*100=14,3%
1 time	5	5/21=0,238	0,238*100=23,8%
1½ time	6	6/21=0,286	0,333*100=28,6%
2 timer	4	4/21=0,190	0,143*100=19,0%
2½ time	1	1/21=0,048	0,048*100=4,8%
I alt:	21	21/21=1	100%

Bemærk, at

- observationen er det målte, her længden af det enkelte barns tv-kiggeri
- hyppighed er det antal observationer, der har en bestemt størrelse
- frekvensen er hyppigheden beregnet som brøk eller procent af det samlede antal observationer

Du noterer, at

- *mindste observation* er 0
- *største observation* er 2,5
- *medianen* er den 11. observation: 1,5
- *nedre kvartil* (medianen for observationerne før den 11. observation): (½+1)/2 = 0,75
- *øvre kvartil* (medianen for observationerne efter den 11. observation): (1½+2)/2=1,75

Her er de 5 *deskriptorer* alle forskellige. Flere af dem og eventuelt alle kan dog være ens.

Observationer og deskriptorer (Definitioner)

En *observation* er resultatet eller svaret på en undersøgelse – ofte et tal. Symbolsk skrives det ofte x_i, hvor i står for et eller andet nummer 1, 2, 3 ... og x_i så er værdien af den i'te observation.

Et *observationssæt* er alle de samhørende tal: $\{x_1 ; x_2 ; x_3 ; ... x_N\}$

Observationssættets størrelse (N) er antallet af observationer; ofte gives de enkelte observationer numre fra 1 til N som ovenfor.

Observationerne kan antage en række mulige værdier; for hver af disse kan vi opgøre, hvor mange af vore observationer, der er lig med denne værdi. Dette er *hyppigheden* (af denne værdi.)

Hyppigheden (h) omregnes ofte til en *frekvens* (f) således:

$$f = \frac{h}{N}$$

Det normale er at angive brøken som decimalbrøk med 2 eller 3 cifre efter kommaet.

Frekvensen kan også angives i procent:

$$f = \frac{h}{N} \cdot 100 \quad \%$$

Middeltallet $\quad \bar{x} = \frac{(x_1 + x_2 + ... + x_N)}{N} = \frac{\sum x_i}{N}$

Typetallet er den hyppigste observation

*Median*en er den midterste observation, når observationerne er ordnet efter størrelse. Er der to i midten (når n er lige), benyttes middeltallet af disse to.

Nedre kvartil findes som medianen, men kun i den første halvdel af observationerne (sorteret i voksende rækkefølge.) Ved ulige antal ses der bort fra midterste observation.

Øvre kvartil findes på tilsvarende måde blandt de største observationer.

Kvartilsættet er mængden bestående af nedre kvartil (= 1. kvartil), median (= 2. kvartil) og øvre kvartil (=3. kvartil).

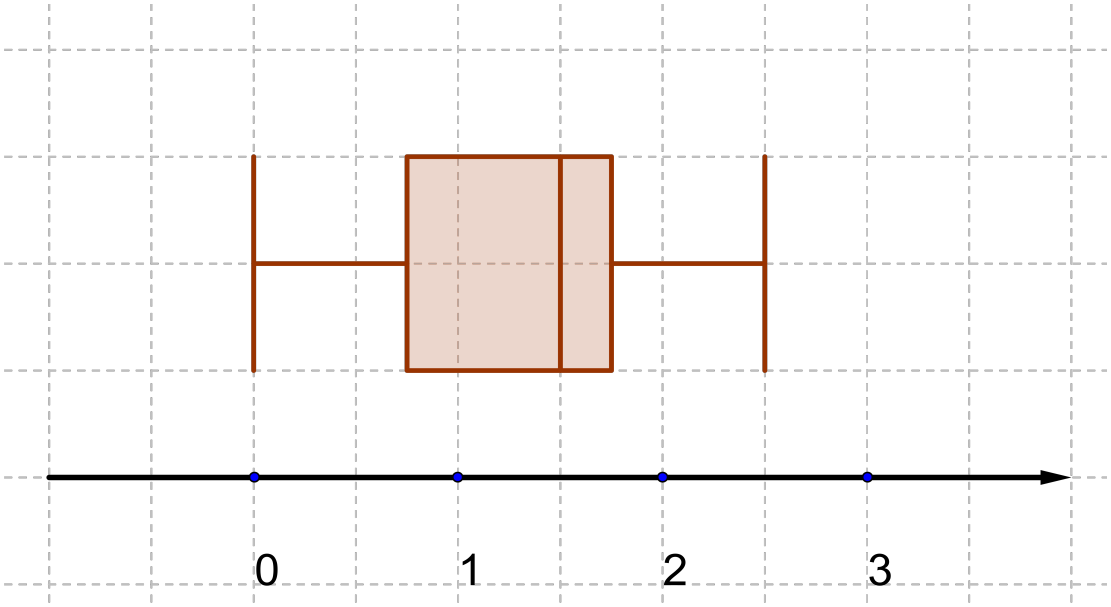
Boksplot

Et boksplot er en figur, der med en boks (= rektangel) viser kvartilerne samtidig med største og mindste observation.

Bemærkning: Der er markeret en sædvanlig x-akse, men der er ingen y-akse. Kassens "højde" har ingen betydning.

Læsbarheden er nok bedst, hvis tegningen placeres lidt over x-aksen; evt. kan der placeres flere boksplots over hinanden for at kunne sammenligne observationssæt. Sammenligningen kan støttes af lodrette hjælpelinjer for bedre at kunne aflæse tal på x-aksen.

Eksempel



Ved hjælp af deskriptorerne fra TV-eksemplet (se side 10) tegnes ovenstående boksplot:

- Der er markeret en almindelig x-akse
- Observationerne er i eksemplet børnenes ”TV-tid” målt i timer
- Den mindste observation er 0 og den største 2,5
- Kvartilsættet {0,75 ; 1,50 ; 1,75} er benyttet til at tegne den midterste firkant, hvor den lodrette linje gennem firkanten markerer medianen, forreste og bagerste lodrette kanter de to øvrige kvartiler.
- Det fremgår fx, at nogle børn ikke har set TV. De udgør dog højst en fjerdedel af alle børnene.
- Tegningen laves nemt med kommandoen Boksplot i GeoGebra; bemærk at figuren kan laves både med en liste med "rå data" eller med mindsteværdi, kvartilsæt og størsteværdi.

Beregn karaktergennemsnit

Øvelse 1

- Beregn karaktergennemsnit (middeltallet) for Ida, der fik karaktererne : 4, 7, 7, 10,10

- Beregn også middeltallet for eleverne til en prøve, der fik karakterer som vist herunder

Karakterer	00	02	4	7	10	12	
Antal karakterer	13	25	45	68	33	7	191

Beskrivelse af issalget

Øvelse 2

- En sommer er der i højsæsonen solgt følgende antal vaffelis (XTR-Mega) pr. dag:

  98, 320, 398, 265 , 225, 175, 195, 285, 267, 165, 138, 185, 168, 241, 350, 375, 300, 280, 264, 277, 239, 258, 285.
- Beregn middeltallet
- Find kvartilsættet (se definitionen side 12)
  - Dvs: Find først medianen – og så "medianen" for hver af de to dele af materialet over hhv. under medianen.
- Tegn det tilsvarende boksplot
  - i hånden
  - med GeoGebra (hvor du benytter mindsteværdi, kvartilsæt og størsteværdi)
  - med GeoGebra (hvor du laver en liste med de rå data)
- Kontroller om alle tegninger principielt er ens

Grupperede observationer

I dag vil man nemt kunne håndtere tusindvis af data opsamlet af computere eller indtastet på computer.

Tidligere har man møjsommeligt måtte regne med papir og blyant og har derfor ved optælling grupperet sine observationer for at få en vis arbejdslettelse – uden at resultaterne for beregningerne ændredes ret meget: I stedet for at Jens var 17 år 287 dage, blev han med en pind registreret fx som én tilhørende aldersgruppen 15-20 årige. Men også i dag må man sommetider nøjes med at arbejde med et grupperet materiale. Det kan skyldes, at materialet er offentliggjort som sådant for at det skal være overskueligt, eller at der ved dataindsamlingen er indtastet afrundede størrelser.

Biografgængerne

Lad os se på et eksempel: Vi vil gerne have et overblik over aldersfordelingen for biografgængerne, der ser en bestemt film. Den information kunne fx anvendes til overvejelser om markedsføring o.l.
I stedet for at betragte alder som en kontinuert størrelse, der ændrer sig sekund for sekund, er det for vort formål nok at placere observationerne i nogle få grupper, således at der kan laves en tabel som nedenstående:

Alders-interval	"pinde"	Interval-hyppigheder
15 - 20	𝍸 𝍸 𝍸 I	16
20 - 25	𝍸 𝍸 𝍸 𝍸 II	
25 - 30	𝍸 III	
30 - 35	𝍸 I	
35 - 40		
40 - 45		
-45	III	

Øvelse 3

- Færdiggør tabellen.
- Kontroller, at der er 60 observationer!

Bemærk, at 20 er grænse i både 1. og 2. interval. Sommetider fremgår det af sammenhængen, hvorledes intervallerne skal forstås, sommetider kan det skrives eksplicit som fx [15 ; 20[som tolkes: Fyldt 15 år og men ikke fyldt 20 år. Den samme observation skal aldrig medregnes i 2 intervaller, men det er ikke hver gang, at de halvåbne intervaller er åbne i højre side.

Frekvenser

Hvis vi beregner intervalhyppighederne i procent af det samlede antal observationer (***N*** = observationssættets størrelse) findes intervalfrekvenserne.

Alders-interval	Interval-hyppigheder	Interval-frekvenser (%)
15 - 20	16	26,7
20 - 25		
25 - 30		
30 - 35		
35 - 40		
40 - 45		
-45		
	60	

$f([15\ ;\ 20[) = \quad 16/60 = 0{,}2667 = 0{,}2667 \cdot 100\,\% = 26{,}7\,\%$

Øvelse 4

- Færdiggør tabellen.
- Kontroller, at der er 60 observationer!

Histogram

Histogrammet minder om søjlediagrammet, men må ikke forveksles hermed.
I søjlediagrammet står der (i princippet) under hver søjle, hvad der måles. Dvs. der er ikke tale om en alm. x-akse!
I histogrammet har man et normalt retvinklet koordinatsystem, hvor x-aksen inddeles **som sædvanligt**!.
I eksemplet her vil vi have en x-akse hvor tallene 0, 5, 10, ...95, 100 er markerede. Den første "søjle" svarende til intervallet [15 ; 20[tegnes mellem 15 og 20, den næste mellem 20 og 25 osv. Bemærk, at her støder nabosøjler sammen. Højden på søjlen svarer til intervalhyppigheden.[4]

4 Hvis alle intervallerne har samme bredde, kan højden på søjlerne afsættes som intervalhypighederne. Men hvis et interval med hyppigheden 15 er 3 gange så bredt som de andre, svarer det jo til 3 søjler med gennemsnitshyppigheden 5. I det tilfælde skal søjlehøjden være 5!
Hvis du på y-aksen angiver procent som måleenhed, bør alle søjlerne være lige brede – eller hvis du har en søjle med 3-dobbelt bredde, kan du opdele den i 3 søjler med stiplede linjer.
Alternativt kan du opgive at benytte y-aksen som målestok for hyppighederne (frekvenserne), men i stedet tegne et lille kvadrat som svarer til fx. hyppigheden: 10 observationer (eller hvis histogrammet er baseret på frekvenser: 10 %)

Histogrammer og flere deskriptorer

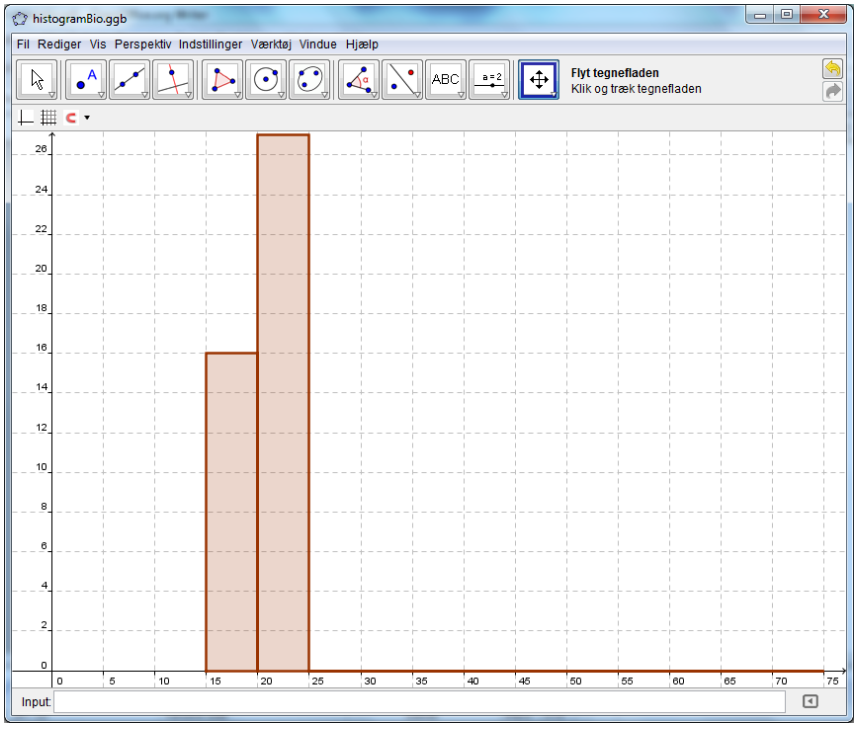
En *intervalhyppighed* er antallet af observationer i et bestemt interval

En *intervalfrekvens* er den hertil svarende brøk eller procent

Et *histogram* er en tegning, der viser disse hyppigheder eller frekvenser som søjler over de respektive intervaller; søjlerne er placeret ved siden af hinanden på en sædvanlig x-akse og deres arealer svarer til hyppighederne (frekvenserne.) Ofte er bredden af søjlerne ens og i det tilfælde svarer søjlernes højde også til hyppigheden (frekvensen.)

Et *typeinterval* er det (de) interval(ler), hvor histogrammets højeste søjle(r) findes.

Histogramøvelse



Øvelse 5

- Færdigggør histogrammet om biografgængerne ovenover. Tegn i bogen!
- Overvej hvorledes de 3 ældste bigrafgængere skal vises i histogrammet.
- Tegn histogrammet med GeoGebra, idet du slal lave to lister: liste1, som indeholder grænserne 15, 20, 25 , … , 75 og liste2 som indeholder dine frekvenser (Se side 17). Sammenlign så histogrammet i bogen med dette.

De kumulerede hyppigheder

Tabellen (side 15) med intervalhyppigheder, kan suppleres med en ekstra kolonne, som det delvist er vist herunder: Kumule**rede** [5] **intervalhyppigheder**:

Ud for intervallet: [15 ; 20[skrives "16", fordi:
antallet af biografgængere med en alder, der er ligger i dette interval er 16 (her: intervalhyppigheden.) - og fordi der ikke findes biografgængere, der er yngre. Dette var jo det første interval!
Ud for intervallet: [20 ; 25[skrives "43", fordi:
antallet af biografgængere med en alder, der er mindre end 25 er:
atallet der er yngre end 20 (dvs. 16)
og dem, der ligger i intervallet [20 ; 25[(dvs 27);
i alt 16+27=43.
Opstillet i en tabel foregår beregningen ret mekanisk:

Alders-interval	Intervalhyppigheder	Kumulerede intervalhyppigheder
15 - 20	16	16
20 - 25	27	43
25 - 30		
30 - 35		
35 - 40		
40 - 45		
-45		

Øvelse 6

- Udfyld resten af tabellen

De kumulerede intervalfrekvenser

Fuldstændigt tilsvarende kan kan intervalfrevenserne summeres. Ved sammenligninger vil de ofte foretrækkes frem for de kumulerede intervalhyppigheder.

Øvelse 7

- Udfyld tabellen på næste side

5 Kumulerede eller summerede

Alders-interval	Intervalfrekvenser	Kumulerede intervalfrekvenser
15 - 20		
20 - 25		
25 - 30		
30 - 35		
35 - 40		
40 - 45		
-45		

Den *kumulerede intervalfrekvens* er brøken (eller procenten) af observationer, der er mindre end eller lig med en mulig observationsstørrelse x. Den er en ikke aftagende funktion af den observerede størrelse og kaldes $F(x)$.

Bemærk: Alders-intervallets højre endepunkt og den tilhørende værdi blandt de kumulerede intervalfrekvenser (i tabellen herover) er hhv. x- og y-værdi for støttepunkter til grafen. Grafen tegnes normalt som en *stykkevis lineær funktion*, idet man forudsætter en jævn fordeling i intervallerne. Punktet med første intervals venstre endepunkt som x-værdi og 0 som y-værdi medtages! (Her: (15 , 0).)

Sumkurven

Sumkurven er defineret som grafen for funktionen $F(x)$.

Øvelse 8

- Tegn sumkurven på almindeligt millimeterpapir (med data fra tabellen herover)
- Tegn sumkurven med GeoGebra

Fraktiler

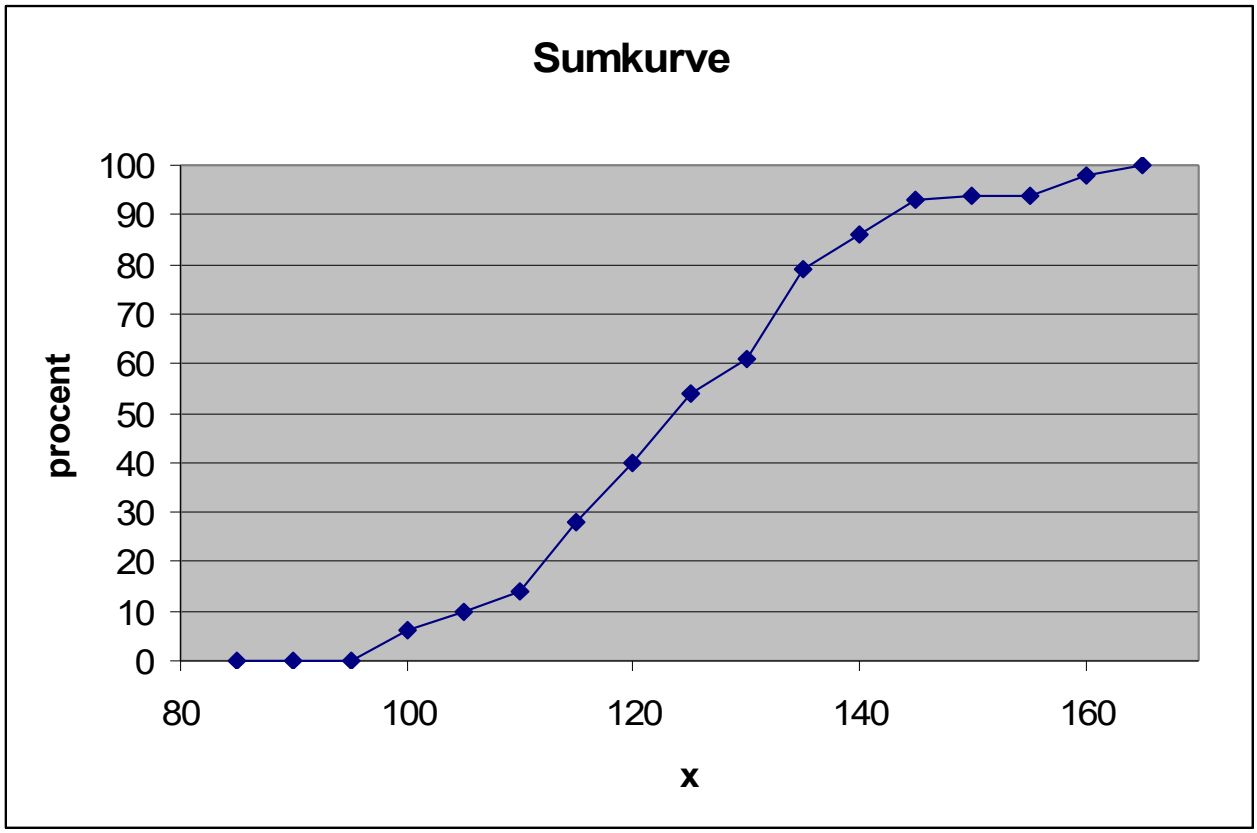
Definition: p %-fraktilen

Vi har en givet sumkurve og tilsvarende funktionen $F(x)$
p %-fraktilen er den (mindste) x-værdi, hvor $F(x) = p$ %

Bemærk: p er en y-værdi. Derfor starter du på y-aksen, tegner en vandret linje ud til grafen og fra første skæringspunkt tegner du lodret ned til x-aksen, hvor du aflæser p%-fraktilen.

Øvelse 9

- Kontroller på nedenstående figur, at 54%-fraktilen er 125.
- Aflæs på nedenstående figur 10 %-fraktil, 25 %-fraktil, 50 %-fraktil, 75 %-fraktil, 90 %-fraktil, 94 %-fraktil, 100 %-fraktil.



Kvartilsæt mv. for grupperede data

25 %-fraktil kaldes 1. kvartil (eller: nedre kvartil)
50 %-fraktil kaldes 2. kvartil eller **medianen**
75 %-fraktil kaldes 3. kvartil (eller: øvre kvartil)
Tilsammen kaldes de 3 kvartiler "*kvartilsættet*".
Skrivemåden for kvartilsættet er fx: Kvartilsættet = {12 ; 17; 21}, idet 1. kvartil = 12 osv.
Bemærk, at denne definition bruges for grupperede observationer – hvor ikke alle enkelte observationer kendes. [6]

6 Det er bestemt ikke normalt, at det samme begreb defineres på to forskellige måder, der vel at mærke giver forskellige resultater. I Danmark er der en gammel tradition for at anvende en definition, der bygger på fraktiler, men denne tradition er ved at blive erstattet af den mere intuitive definition, du tidligere har set.
For ikke at blive misforstået, når du anvender kvartilsæt, må du derfor tilføje, hvordan kvartilsættet er fundet: Enten ved hjælp af de tre fraktiler eller ved hjælp af de tre observationer, der deler materialet i 4 lige store dele.

Hos lægevagten

Øvelse 10

- Hos lægevagten i X-køb̈ing har man lavet statistik over antallet af henvendelser, som det fremgår af skemaet under øvelsen.
- Tegn et histogram, der svarer til tabellens oplysninger
- Beregn middeltallet for antal henvendelser pr. døgn (Antag, at middeltallet for antallet af henvendelser indenfor et interval er intervallets midtpunkt)
- Hvad er typeintervallet?
- Beregn frekvenserne, der svarer til hyppighederne
- Beregn de kumulerede frekvenser
- Tegn sumkurven
- Find kvartilsættet
- Hvor mange procent af døgnene var der færre end 125 henvendelser?
- Hvor mange procent af døgnene var der flere end 200 henvendelser?
- Tegn et "boksplot"

Antal henvendelser pr. døgn	Hyppighed
0 – 50	5
51 – 100	37
101 – 150	98
151 – 200	165
201 – 250	201
251 – 400	45

Middeltal for grupperede data

Vi vender tilbage til biografeksemplet. Vi kender ikke den enkelte observation, men "snyder" ved at antage, at de er jævnt fordelt i hvert interval. Derfor bliver middeltallet for observationerne i et interval intervallets midtpunkt.

Alders-interval	hyppigheder h	midtpunkt m	m*h
15 - 20	16	17,5	280,0
20 - 25	27	22,5	607,5
25 - 30	8	27,5	220,0
30 - 35	6	32,5	195,0
35 - 40	0	37,5	0,0
40 - 45	0	42,5	0,0
45 -	3	60	180,0
	60		1.482,5
Middeltal	= 1.482,5/60		24,7

Denne forudsætning er oftest kun omtrent rigtig, men ofte vil fejlene omtrent udligne hinanden. Du bedes bemærke, at i det sidste interval kendes højre intervalendepunkt ikke. For at løse opgaven, må vi lave et intelligent gæt på dette, her benyttes 75 år. Du kan prøve at erstatte det med 50 eller 80: I værste fald ændres middeltallet til 24,1

Middeltallet for grupperede data

For en række observationer $x_1, x_2, ..., x_N$
kan middeltallet beregnes som:

$$\overline{x} = \frac{\sum_{i=1}^{n} h_i \cdot m_i}{N} = \sum_{i=1}^{n} \frac{h_i}{N} \cdot m_i = \sum_{i=1}^{n} f_i \cdot m_i \qquad (2)$$

hvor det græske bogstav $\sum$ (stor sigma) betyder, at man skal addere. Her er det en række produkter svarende til rækkerne i tabellen. Antag at der er n rækker: i skal så have alle værdierne fra 1 til n, således at der beregnes et produkt pr. række. Her antages det samlede antal observationer at være N, som summen derfor divideres med. Omskrivningen viser, at du i stedet for at bruge hyppighederne kan benytte frekvenserne.

Sandsynlighedsteori

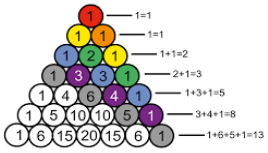
Christiaan Huygens

af Caspar Netscher

Roulette

Snake eyes

Pascals trekant



♖	1	1	1
1	2	3	4
1	3	6	10
1	4	10	20

Forhistorien

Spil med udfald, der afhænger af tilfældet, har været kendt i årtusinder. I Indusdalen har arkæologer fundet 5000 år gamle terninger. Terninger har været brugt til hasardspil i Romerriget eller som gudernes vejledning for Caesar ved floden Rubicon: "Jacta alea est", dvs. Terningen er kastet.

Sådanne spil er også kendt i Danmark fra folkeviserne:

"Og hør, du favre gangerpilt,
leg tavlebord med mig!"

Men først sent begynder man at spekulere over, hvorledes man kunne regne på chancerne for gevinst eller tab. *Gerolamo Cardano* (1501 - 76) er et uægte barn af en italiensk jurist med mange talenter spændende fra astrologi, astronomi, biologi, kemi, fysik, matematik og hasardspil blandt meget andet. Han var blandt de første, der regnede på gevinstchancer. Senere følges hans arbejder op af *Blaise Pascal* (1623 - 1662) og *Pierre de Fermat* (1607 - 1665). Mest kendt er måske deres brevveksling fra 1654, hvor de diskuterer løsninger til spilproblemer.
Men vi skal helt frem til 1933 før *Andrey Kolmogorov* (1903 - 1987) formulerer det moderne grundlag for sandsynlighedsteorien med anvendelse af aksiomer, mængdeteori og en sandsynlighedsfunktion.

Øvelse 11
Kolmogorov's first mathematical discovery was published in this journal[8]: at the age of five he noticed the regularity in the sum of the series of odd numbers:

- $1=1^2; 1+3=2^2; 1+3+5=3^2; 1+3+5+7=4^2; ...$
- Bevis sætningen ...

Frekvenser og sandsynligheder

Når en række identiske eksperimenter har fundet sted, kan vi med hyppigheder og frekvenser beskrive eksperimenternes udfald. Dertil benyttes teknikken fra den deskriptive statistik.
Når vi ser, hvorledes frekvenser (for et bestemt udfald) ser ud til at nærme sig et bestemt tal mellem 0 og 1, når serien af eksperimenter bliver længere og længere, kunne man forestille sig, at der fandtes et sådant bestemt tal, som vi kunne kalde **P** [8].

På den baggrund er der opbygget en teori til beskrivelse af *stokastiske* eksperimenter (det vil sige med et tilfældigt udfald), der tager udgangspunkt i:
at der findes en funktion, der til enhver begivenhed (svarende til delmængder eller enkelte udfald) kan angives som en funktionsværdi mellem 0 og 1 (dvs. sandsynligheden for begivenheden.)
Teorien hviler på nogle grundsætninger (jævnfør og sammenlign med Euklids Elementer) som ikke (kan) bevises, men som er inspireret af frekvenser. Herfra udledes så teorien om sandsynligheder. Men hvor Euklids aksiomer har langt over 2000 år på bagen, har

7 Skoleavisen: "Forårssvalen"
8 P anvendes som generel forkortelse for sandsynlighed (fra engelsk Probability.)

grundlaget for sandsynlighedsteorien udviklet sig gennem de seneste århundreder.[9]

Men hvorledes findes sandsynlighederne? Tidligt gives udfaldet "6 øjne" i et terningeslag sandsynligheden 1/6. Ikke fordi man havde eksperimenteret med (næsten) uendelig lange serier af terningekast, men "fordi det synes logisk". Af symmetrigrunde. Det kaldes en **a priori** viden.
På et senere tidspunkt interesserer statistikere sig for demografi (befolkningslære.) I den sammenhæng kan en fødsel betragtes som et stokastisk eksperiment med to mulige udfald: dreng eller pige. Hvis man brugte et symmetriprincip kunne sandsynligheden for at få en dreng være 0,50 eller 50%. Men erfaringen viser, at der fødes flere drenge end piger, og at frekvensen af drenge i forhold til samtlige fødte er uhyre konstant. Sandsynligheden for at få en dreng kan så opfattes som den ukendte brøk, den målte frekvens er tæt på og – påstås det – nærmer sig. I modsætning til a priori kaldes dette **a posteriori** viden.
Men om den sandsynlighedsteoretiske model, der følger på næste side, beskriver virkeligheden mere eller mindre godt, kan kun erfaringen vise.

Find delmængder

Øvelse 12
Tænk på kast med en terning. Skriv udfaldsrummet op.
En delmængde kan være: {1, 5, 6}.

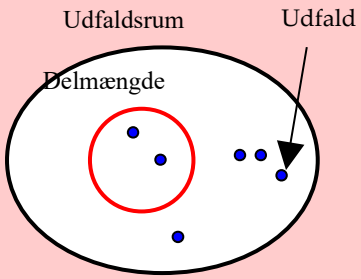
- Skriv så mange delmængder op, som du kan, i løbet af 2 minutter.
- Beskriv nogle af dem med ord og nogle med X og tegn som <, >, =.

9 Eller som formuleret af Harald Cramer i Sannolikhetskalkulen: "... kan sannolikhetskalkulens uppgift i störste allmänhet sägas vara att *ge ett schema för matematisk beskrivning och behandling av en vis klass av iagttagna företeelser* "

Grundlaget for sandsynlighedsteorien

Lad os præcisere og opsummere grundlaget:

- Vi beskæftiger os med *stokastiske eksperimenter* og benytter her terningekastet som eksempel. Herved forstås, at eksperimentet udføres på "nøjagtig samme måde" hver gang,, men resultatet varierer uforudsigeligt.
- I forbindelse med eksperimentet defineres en *stokastisk variabel*: X. X er det vi måler. Ved et terningekast tælles terningens øjne. Udfaldet "der vises 2 øjne", skrives ofte "X=2".
- Vi definerer en funktion *P*. Definitionsmængden for *P* er enhver delmængde *D* af *udfaldsrummet U*.
  - Til sammenligning med tidligere funktioner:
    - Her bruges altså bogstavet *P* i stedet *f* eller *g* eller ....
    - *D* erstatter *x*; istedet for bogstavet *D* kan der være en beskrivelse af delmængden med almindelige ord. I eksemplet kan delmængden fx være "Terningekast med et lige antal øjne".
- $f(x)$ erstattes så af $P(D)$



Den røde ring markerer en delmængde med 2 udfald - som et eksempel – og for hver ring, der overhovedet kan tegnes, har vi en delmængde, der må have en sandsynlighed.

- For *P* gælder:
  - $P(U) = 1$
    - Reglen siger, at det er "100 % sikkert", at vi får et af de mulige udfald.
    - Anvendt på eksemplet er $U = \{1; 2; 3; 4; 5; 6\}$
  - Påstanden læses: Sandsynligheden for at få et af udfaldene i *U* (altså fra 1 til 6 øjne) er 1 eller 100%.
- $0 \leq P(D) \leq 1$
  - Påstanden læses: Sandsynligheden for et udfald i delmængden *D* (som kan være udfald med 5 øjne eller udfald med et lige antal øjne eller udfald med højst 2 øjne ..) ligger mellem 0 og 1.

- $P(\varnothing) = 0$
  - $\varnothing$ er den tomme mængde; den indeholder ikke nogen udfald fra udfaldsrummet. X kan derfor ikke få en værdi fra $\varnothing$.
  - Hvis $D_1 \cap D_2 = \varnothing$, gælder: $P(D_1 \cup D_2) = P(D_1) + P(D_2)$
    - Forudsætningen er: De to delmængder har ingen fælles udfald.
    - I eksemplet kunne de to delmængder være hhv. $D_1 = \{1; 3\}$ og $D_2 = \{4; 5\}$. Så er $D_1 \cup D_2 = \{1; 3; 4; 5\}$

Binomialfordelingen

Lad os tænke på et eksperiment, hvor der kun er to mulige udfald: succes og fiasko. Det kan være:

1. spillet: plat eller krone
2. terningekast, hvor et bestemt antal øjne udnævnes til succes
3. fødsel, hvor fødsel af et udvalgt køn (ikke nødvendigvis det samme i alle subkulturer) kaldes succes.

Binomialforsøg

Et eksperiment med 2 mulige udfald kaldes et binomialforsøg.
Kaldes sandsynligheden for det første udfald (= succes) p, giver det sig selv, at sandsynligheden for det andet udfald (= fiasko) er 1-p.

Udfald	P(Succes)	P(Fiasko)
X= S/F	p	1-p

Forestil dig, at du gentager et sådant forsøg 10 gange eller n gange. De 10 eller n eksperimenter udgør også et samlet eksperiment, hvor man kan tælle antallet af successer. Dette tal er en stokastisk variabel X, som siges at følge binomialfordelingen [10]:

Binomialfordelingen

X følger en binomialfordeling, hvis X er en stokastisk variabel, der tæller antal succeser i n binomialforsøg.
Det skrives: $X \sim b(n;p)$
hvor $U = \{0; 1; 2; 3; \ldots ; n\text{-}1; n\}$
og b minder om, at det er en binomialfordeling; n er antallet af binomialforsøg og p er sandsynligheden for succes i det enkelte binomialforsøg.
Hvis $X \sim b(10;0,5)$ tæller X antal succeser, når vi laver et binomialforsøg 10 gange; i hvert af forsøgene er sandsynligheden for succes 0,5 = 50%.
P(X=3) betyder: sandsynligheden for at få 3 succeser blandt de 10 binomialforsøg.

10 "Bi" betyder noget med dobbelt /to; kendes fra: Ur**bi**no (italiensk bynavn: to byer), **bi**gami og **bi**seksuel.

Binomialfordeling I

Øvelse 13
Lad X~b(10 ; 0,5).

- Hvor uheldige kan vi være? hvad er det mindste antal succeser, vi kan få?
- Hvad er det største antal?
- Hvad er udfaldsrummet U = ?
- Hvis du skulle gætte på et typisk antal succeser når du laver de 10 forsøg: hvad ville du så gætte på?
- Prøv at slå plat og krone 10 gange; tæl kroner (=succes.)
- Fik du hvad du gættede på?
- Hvis ja: sker det også næste gang?
- Hvis nej: hvad vil du nu gætte på, hvis du gentager forsøget?

Binomialfordeling II

Øvelse 14

- Forklar med egne ord, hvad der kan læses i skemaerne til højre.
  - Øverste skema
  - Midterste skema
  - Nederste kema
- Beregn følgende sandsynligheder, idet vi forudsætter X~b(12; 0,3)
  - P(X=0)
  - P(X<4)
  - P(X > 10)
  - P(3,5≤X≤10)
  - P(X>0)
  - P(X er ikke 11)

P(X=a)	
n=	10
p=	0,2
a=	4
P(X=a)	0,0881

P(X<=a)	
n=	10
p=	0,2
a=	4
P(X<=a)	0,9672

P(a<=X<=b)	
n=	10
p=	0,2
a=	3
b=	6
P(a<=X<=b)	0,3213

Både punktsandynlighederne (P(X=*a*)) og de andre sandsynligheder kan beregnes med lommeregner, regneark eller GeoGebra (med Sandsynlighedslommeregneren.)

Tordenskjold

Øvelse 15
Tordenskjold havde 11 brødre og 6 søstre. Hvis hver fødsel er et eksperiment med sandsynligheden 0,51 for at få en dreng:

- Hvad er så hr. og fru Wessels sandsynlighed for at få netop denne kombination af piger og drenge (givet antal fødsler)?
- Hvad er sandsynligheden for at få mindst 12 sønner ved 18 enkeltbarnsfødsler?

Gallup

Øvelse 16
Ved en Gallupundersøgelse udspørges 1200 vælgere, om de vil stemme på partiet W. 70 vælgere svarer, at de vil stemme på partiet. Partiet hævder, at det vil få 7% af stemmerne ved næste valg.

- Hvad mener du? Overdriver partiet ...?

Sprogbrug

Øvelse 17

- Hvorfor er det rimeligt at sige, at sandsynligheden er 2/3 for at få mindst 3 i et terningekast med en sædvanlig terning?
- Hvad er så sandsynligheden for at få "mindst 3" mindst 6 gange af 10 kast?
- Hvis det lyder knudret skal du prøve at slå 10 slag; for hver slag noterer du, om du fik mindst 3? Så var det slag en succes! Hvor mange succeser fik du ialt? Var det mindst 6 gange?
- Prøv så at besvare spørgsmålet igen.

Cola

Øvelse 18
Når man vil teste, om en gruppe mennesker kan smage forskel på 2 produkter (som for eksempel forskellige varianter af cola), kan man lave følgende binomialforsøg:
Hæld cola op i 3 glas således at indholdet i de 2 glas er det samme, men det tredie glas indeholder den anden cola. Forsøgspersonen ved ikke, hviket glas der har et andet indhold end de to andre.
Forsøgspersonen bliver nu bedt om at udpege det glas, der "smager anderledes."

- Hvorfor er dette et binomialforsøg?
- Hvis forsøgspersonen kan pege på det rigtige glas, er udfaldet "en succes".
- Hvad er p i binomialforsøget, hvis forsøgspersonen bare gætter?
- Hvad er p i binomialforsøget, hvis forsøgspersonen sikkert kan smage forskel?
- Da binomialforsøget blev gentaget 12 gange fik vi 5 rigtige svar.
- Kan forsøgspersonen (eller forsøgspersonerne) sikkert smage forskel på de 2 colaer? (Meget let spørgsmål!)
- Kan det tænkes, at han bare har gættet?

Praktiske anvendelser

Du kommer ofte ud for "eksperimenter", hvor udfaldet ikke er sikkert, og hvor du møder påståstande som:

- frøene har en spiredygtighed på 90 %
- 60 % af lodsedlerne giver gevinst
- 9 ud af 10 filmskuespillere spiser "Lux"
- Efter indførelse af den nye operationsmetode overlever 31 % af patienterne ...
- Smitterisikoen var 80 % ved fysisk kontakt

Det første problem er selvfølgelig at forstå præcist hvad der menes med påstanden, fx om frøenes spiredygtighed. En naturlig forståelse vil være, at sælgeren har lavet undersøgelser (har erfaringer), der godtgør, at han forventer at 90 % af alle hans frø vil spire.
Når du som køber køber n frø (altså nogle, men ikke alle) og sår dem, kan du forvente, at ca. 90 % af dem vil spire – med de tilfældige udsving, man kan forvente.

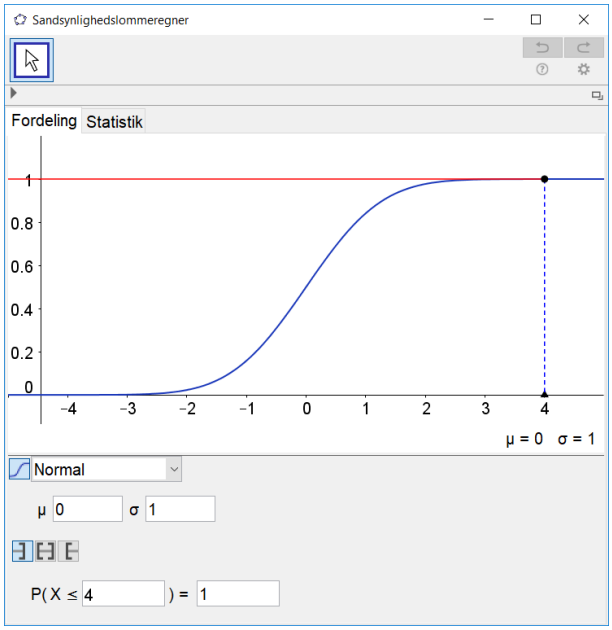
Det næste problem er så at undersøge, om påstanden er rigtig. Dels kan sælgeren have overdrevet spiredygtigheden, dels kan frøene have været udsat for forhold, der forringer spiredygtigheden. Afhængig af vigtigheden og økonomien kan du undersøge en større eller mindre stikprøve og vurdere om resultatet af stikprøven er i overensstemmelse med sælgerens påstand. Og forstå baggrunden for at afvise en påstand. (Se næste de afsnit.)

Det tredje problem er (her) sælgerens problem: hvorledes kan han garantere, hvorvidt frøene vil spire? hvilke målemetoder har han benyttet? og hvor sikre er de? hvor langt fra det rigtige kan de 90 % være?

Det fjerde og ikke mindst vigtige problem er – når man arbejder med stikprøver, hvad man oftest gør – at stikprøven skal ligne helheden: den skal være repræsentativ. Det kan ikke nytte noget kun at udtage frø fra én sæk; denne kan have ligget meget udsat, have fået fugt, kulde osv. eller den kan modsat have været opbevaret korrekt og bedre end alle de andre. Den rigtige udvælgelsesmetode skal sørge for, at ethvert frø har samme sandsynlighed for at indgå i stikprøven.

Normalfordelingen

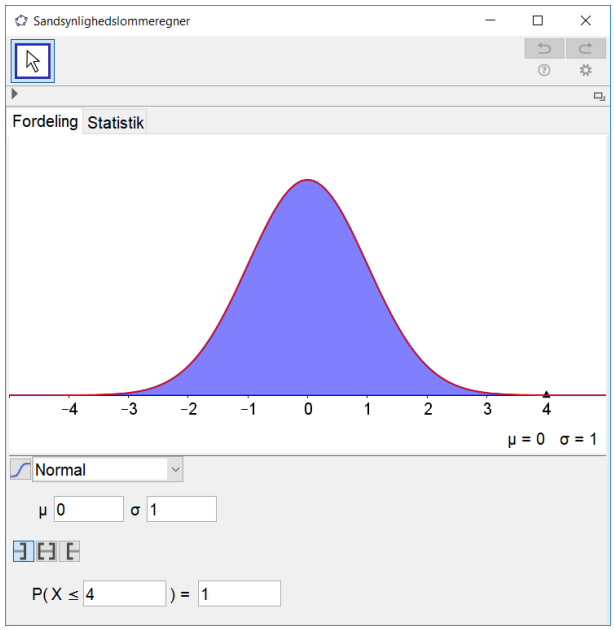
Stokastiske variable kan også antage andre værdier end de hele tal. Antag fx at det stokastiske eksperiment består gentagne gange at måle en bestemt længde. På grund af måleusikkerheden vil udfaldene falde i et større eller mindre interval og principielt er udfaldsrummet alle de reelle tal.
Svarende til sumfunktionen fra den deskriptive statistik kan vi med de teoretiske sandsynligheder lave en tilsvarende funktion:
$F(x)=P(X<x)$ kaldet *fordelingsfunktionen*.
For standardnormalfordelingen (X ~ N(0,1)) ses fordelingsfunktionens graf til højre.



Svarende til histogrammet fra den deskriptive statistik kan der tegnes en graf kaldet *tæthedsfunktionen*: Søjlehøjderne fås som $F(x+h)-F(x)$; lader vi h gå mod 0 (nul) fås et histogram med mange søjler, men med højder, der også går mod 0. Derfor skaleres søjlehøjden ved division med h, så grænseværdien bliver

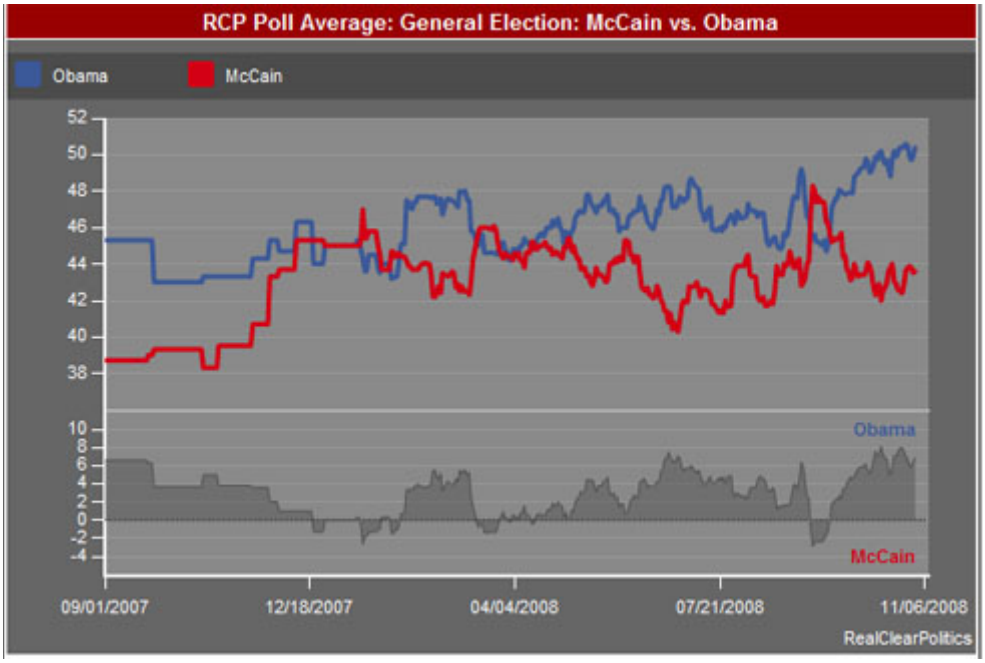
$$f(x)=\lim\left\{\frac{F(x+h)-F(x)}{h}\right\}=F'(x) \qquad (3)$$

f kaldes tæthedsfunktionen og har grafen som vist her med rødt:
For yderligere information om normalfordelingen se Erik Vestergaards hjemmeside. [11]



11 http://www.matematikfysik.dk/mat/noter_tillaeg/normalfordeling.pdf

Estimat, Hypotese og Test



Estimat

Eksempel: Opinionsundersøgelser og estimater

Når Gallup, Observa, Sonar eller andre spørger 1400 udvalgte personer **(*stikprøven*)**, hvilket parti de vil stemme på ved næste folketingsvalg (og hvilken sæbe , de vasker sig med - og hvilken filmstjerne de drømmer om og mange andre interessante forhold), er det ikke fordi man er interesseret i netop disse menneskers meninger og vaner. Man spørger, fordi man tror, at man ved hjælp af svarene kan få noget at vide om hele befolkningens **(*populationens*)** holdning til fx de politiske partier.
Interessant for partierne er det at få oplyst, hvilken procentdel af befolkningen, der pt. vil stemme på dem. Det kan man ikke få at vide uden at spørge alle, men det er for dyrt.
Ved hjælp af stikprøven, kan vi gætte - og forhåbentligt fornuftigt.
I vores eksempel vil vi gerne kende procenten (eller brøken) p, som fortæller om den brøkdel af stemmerne et parti vil opnå. p er altså et ganske bestemt tal - i denne sammenhæng.
Ved hjælp af stikprøven kan vi regne mange forskellige tal ud; for eksempel den brøk **i stikprøven**, der har villet stemme på det pågældende parti. Vi gætter på, at dette tal er tæt på den brøkdel af populationen, der ville stemme på partiet.
Sådanne gæt kaldes ***estimater***, og i princippet kan de vise det fuldstændigt rigtige svar, et omtrent rigtigt svar og noget fuldstændigt forkert. Men estimatet er ikke det samme hver gang og der er ingen garanti for, at det er "tæt" ved tallet vi søger.
Estimatet er et resultat af et stokastisk eksperiment, hvor vi konstruerer en (særlig) stokastisk variabel (kaldet ***estimatoren*)**: den tilfældige værdi af denne er estimatet.

Lad os antage, at vi **ved**, at partiet NN har en vælgertilslutning på 30 %, og at vi laver en miniopinionsundersøgelse ved at spørge 10 tilfældigt udvalgte personer.
Hvis stikprøven skal ligne populationen, skal der være 3 i stikprøven, der vil stemme på partiet. Men tilfældet er en skælm: Der kan jo være alt fra 0 til 10, der siger ja. Og selv om vi ved, at X~b(10; 0,3) og at p = 0,3, kan vi komme til at gætte på, at parameteren p = 0,0; 0,1; 0,2; ... ; 0,9; 1,0.

Tabel for punktsandsynlighederne X~b(10; 0,3)

Antal succes'er	0	1	2	3	4	5	6	7	8	9	10
Beregnet sandsynlighed i %	3	12	23	27	20	10	4	1	0	0	0

Øvelse 19

- Hvilke procenter (for vælgertilslutning) tror du ikke på, der vil blive gættet på? (når vi laver miniopinionsundersøgelsen.)
- Hvad bliver det typiske gæt?

I den følgende øvelse vil vi vurdere, hvor gode gæt er under forskellige omstændigheder.
En vigtig pointe ved disse undersøgelser er, at alle i befolkningen skal have nøjagtig den samme chance for at være med. Ellers bliver undersøgelsen misvisende. Og det er sket

utallige gange. Hvordan man sikrer sig dette - så godt som muligt - er en større videnskab, og behandles ikke her.

Kvalitet af estimat

Øvelse 20
Idet du benytter GeoGebras Sandsynlighedslommeregner, besvares spørgsmålene herunder.

- Vi spørger 200 personer og "ved", at der i befolkningen er 30 % (dvs. p=0,3) Tegn et histogram (med de teoretiske punktsandsynligheder som frekvenser).
- Hvad er det typiske udfald af eksperimentet?
- Hvilke værdier af p svarer det til?
- Hvad er P(43<X<50))?
- Hvor tit vil det ske, at selvom om p=0,30, vil man gætte på, at værdien er 0,25 eller mindre?
- Besvar det sidste spørgsmål igen ved forskellige stikprøvestørrelser: n=100, 400, 1600.

Hypoteser

Når nogen fremsætter en påstand, kaldes det her en ***hypotese***. Hypoteser kan være sande eller falske og vi kan tro på dem eller lade være med at tro på dem.

Øvelse: Muligheder for en hypotese

Øvelse 21

Hypoteser		
	H er sand	H er falsk
Påstand: Vi tror H er sand		
Påstand: Vi tror H er falsk		

- Hvilke af mulighederne er "gode" og hvilke er "uheldige"? Sæt X ved de gode kombinationer

Eksempler på hypoteser kan være:

1. 13 % af de stemmeberettigede vil pt. stemme på partiet Akrobaterne..
2. 95 % af sukkerposerne vejer over 2000 g
3. Der er gevinst på hver 10. lodseddel

I forbindelse med hver af hypoteserne vil der så være en alternativ hypotese: at hypotesen ikke er rigtig.

I det første eksempel er den alternative hypotese: Nej der er ikke 13 %, men færre (flere.)

Om man vil bruge færre eller flere afhænger selvfølgelig af, hvem der fremsætter påstanden; er det partiets presseafdeling eller spindoktor, vil man formode, at han vil tegne et optimistisk billede af situationen – og evt. overdrive tilslutningen. Så benyttes "færre".

Når vi afgør, om vi tror på hypotesen eller ej, er logikken:

> Vi tror altid på hypoteser (men ved godt, at der er unøjagtighed forbundet med dem) – med mindre vi får et usandsynligt resultat. Det usandsynlige tror vi ikke på. Da det usandsynlige en sjælden gang forekommer, må vi acceptere, at en sjælden gang kommer vi til at begå en fejl ...

Men selvom vi tror på hypotesen (accepterer den og lader være med at forkaste den), ved vi jo godt, at den ikke behøver at være helt rigtig. Den er bare ikke så meget forkert, at vi er overbevist om, at den må være forkert.

Test

Når man undersøger en **hypotese,** kaldes det at **teste hypotesen.** En test går ud på at gennemføre et eksperiment, regne med at den oprindelige hypotese er sand og derefter vurdere, om udfaldet af eksperimentet er sandsynligt i den givne model.

Hvis resultatet er usandsynligt; dvs. kun sker for eksempel 1 ud af 100 gange, vil vi **forkaste** hypotesen, dvs. vi tror ikke påstanden er sand. Men vi ved godt, at en gang i mellem sker det usandsynlige (som for eksempel at vinde den store milliongevinst.) Men det regner vi ikke med til daglig. Så en gang i mellem begår vi **fejl**.

Hvis resultatet ikke er usandsynligt (og det er noget andet end sandsynligt !!) **forkaster vi ikke** hypotesen. Den kan være rigtig eller nogenlunde rigtig. Vi ved i hvert tilfælde ikke nok til at afvise den. Og også her kan vi begå en **fejl** ved ikke at forkaste en forkert hypotese; hypotesen er bare ikke forkert nok til, at vi med sikkerhed kan afvise den.

Eksempel på hypotese og test

Lad partiet W hævde, at deres tilslutning blandt vælgerne er 7 %.

Vi lader nu Observa spørge 1000 tilfældigt udvalgte vælgere, om de vil stemme på W.

Vi går ud fra en masse urealistiske forhold vedrørende modellen:

- At de udspurgte er tilfældigt udvalgte blandt alle stemmeberettigede
- At alle vil svare og svarer
- At de svarer ja – eller nej
- At de svarer ærligt
- …
- At vi kan bruge en binomialfordeling som beskrivende model
- At modellen i dette tilfælde er: X~b(1000 ; 0,07)

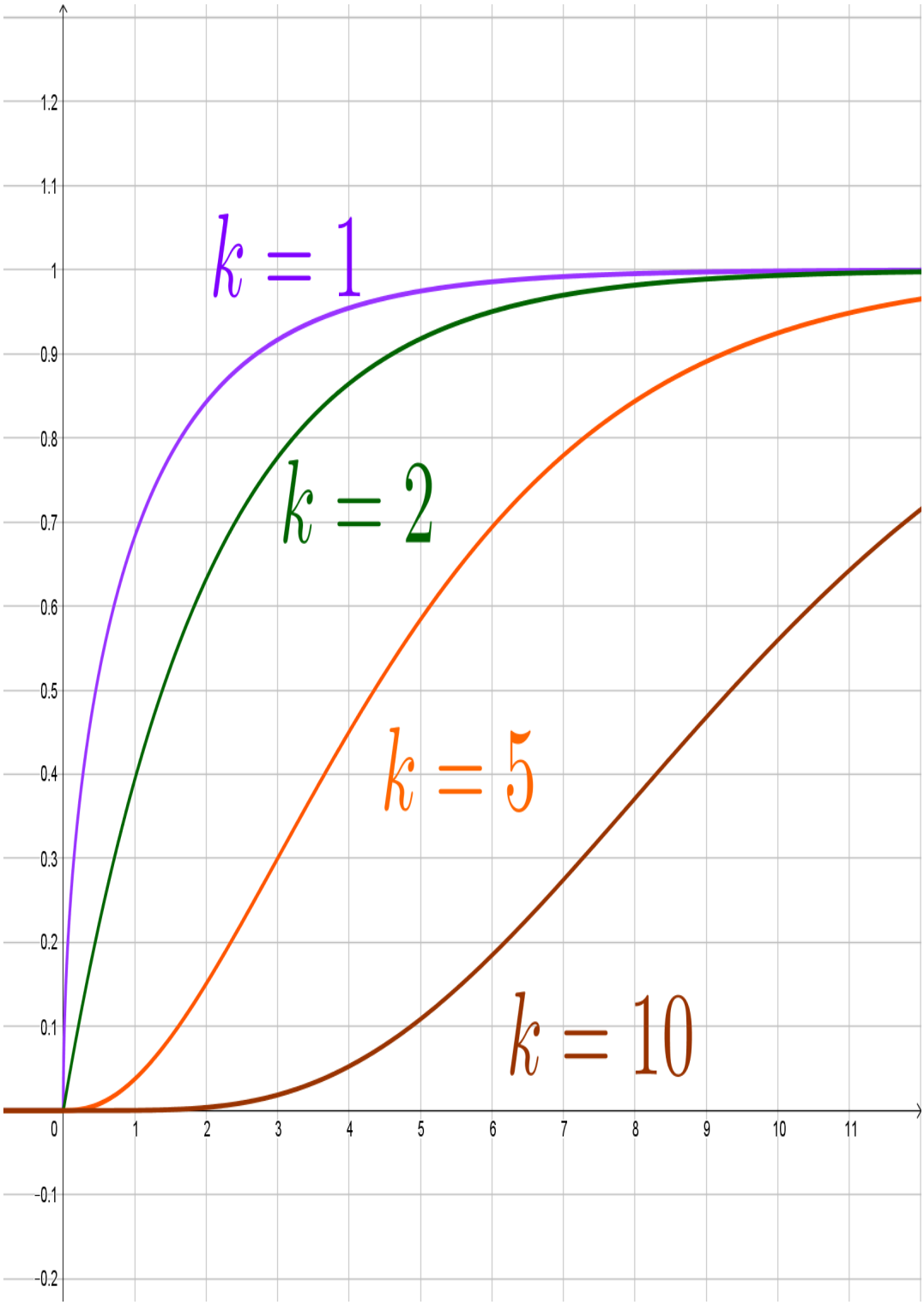
Hypotesen (kaldet Ho) er: $p = 0{,}07$

Den alternative hypotese er: $p < 0{,}07$

Vi får nu svaret fra Observa:

1. **Antag svaret er: 41**; hvis påstanden om p = 0,07 er rigtig, kan vi med binomialfordelingen beregne, at dette resultat (eller noget der er endnu ringere) kun indtræffer i 0,01 % af alle tilfælde.
2. Derfor forkastes hypotesen; vi tror ikke at Observas undersøgelse har ramt et så forkert resultat – vel vidende, at det sker en sjælden gang.
3. **Hvis svaret i stedet for havde været 55**, kunne man tilsvarende beregne, at det sker i 3,3 % af alle tilfælde; hvis W har været meget uheldig, kan deres påstand godt være rigtig – og vi forkaster ikke hypotesen. Det kan godt være, der er en smule tvivl om det nu kan passe, men tvivlen er ikke stærkt nok begrundet til at vi vil sige: det her med 7 % tror vi ikke på.

X^2-fordelingen

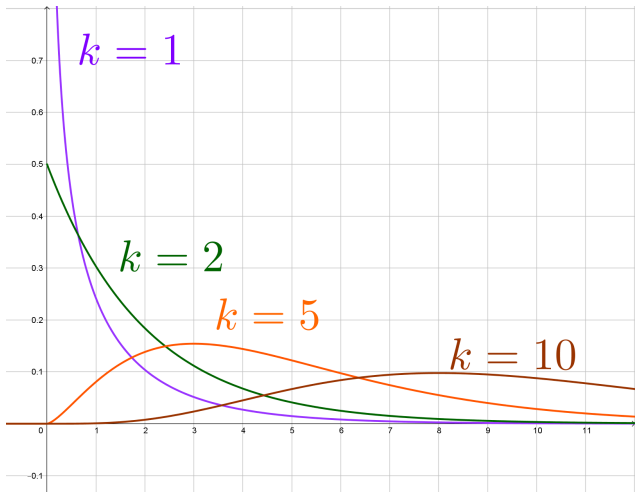


Definition

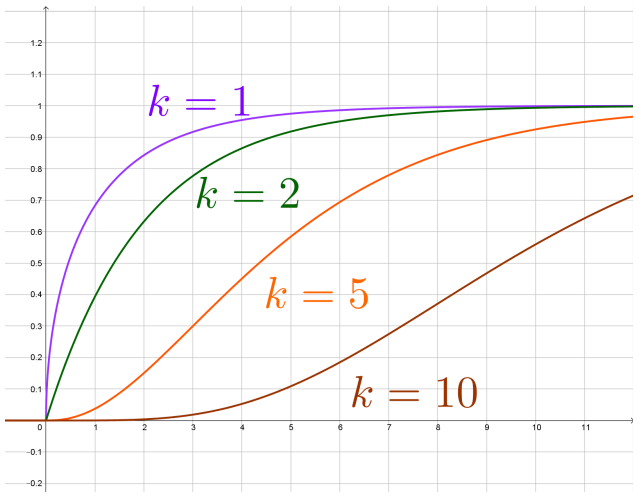
Lad $X_i \sim N(0,1), i \in \{1,2,3,\ldots,k\}$ være k uafhængige stokastiske variable (der er normalfordelte.) Så er

$\chi 2 = \sum_{i=1}^{i=k} X_i^2$ en stokastisk variabel som følger en χ2–fordeling med k frihedsgrader (χ2 udtales ki-i-anden).

De to tegninger herunder viser grafer for hhv. tæthedsfunktioner og fordelingsfunktioner.



Tæthedsfunktioner



Fordelingsfunktioner

Anvendelsen

Ved to typer af test anvendes en teststørrelse, som følger en χ2–fordeling. De kaldes

- Uafhængighedstest og
- Goodness-of-fit test

Uafhængighedstesten tester, om en stikprøve kan komme fra en population opdelt i delpopulationer (fx mænd / kvinder eller unge / gamle) og hvor frekvenserne for kategorierne i den ene delpopulation er mage til frekvenserne i de andre delpopulationer.

Goodness-of-fit test tester, om en stikprøve kan komme fra en population med kategorier, for hvilke frekvenserne er kendte.

Hvordan

Vi vil gerne vide noget om en *population*, uden at spørge / undersøge alle. Altså udvælges en *stikprøve*, som altid bør være en *repræsentativ delmængde* af populationen. Hermed menes, at ved udtrækning af stikprøven skal enhver i populationen have samme sandsynlighed for at komme med i stikprøven.

Både population og stikprøve består af individer, hvortil der til individerne er knyttet en *variabel*. I disse test vil det være en kategorisk variabel som svar (ja / nej), partivalg (A / B / C ...), indkomstinterval eller hvor individet er et terningeslag antal øjne (1 / 2 / 3 / 4 / 5 / 6).

X^2 test

bruges for at vise, om nogle delpopulationer opfører sig ens eller forskelligt. *Hypoteser* (altså påstande om populationen) af denne art kan formuleres på forskellige måder:

- Der er ingen forskel på kvinders og mænds valg af bopæl
- Valget af bopæl er uafhængigt af køn
- Den samme brøkdel af kvinder som af mænd vælger at bo i byer og den samme brøkdel af kvinder som af mænd vælger at bo på landet

Alle formuleringerne udtaler sig om populationen og betyder reelt det samme. I teorien kunne man også undersøge den stik modsatte hypotese. Problemet med den er, at man i så fald ikke har gode modeller til at beregne sandsynligheder med. Derfor formuleres hypotesen om populationen på én af ovenstående måder. Hvis man i virkeligheden tror, at der er forskel på mænd og kvinder, må man så undersøge, om der er belæg for (grunde til) at forkaste ovenstående hypotese.

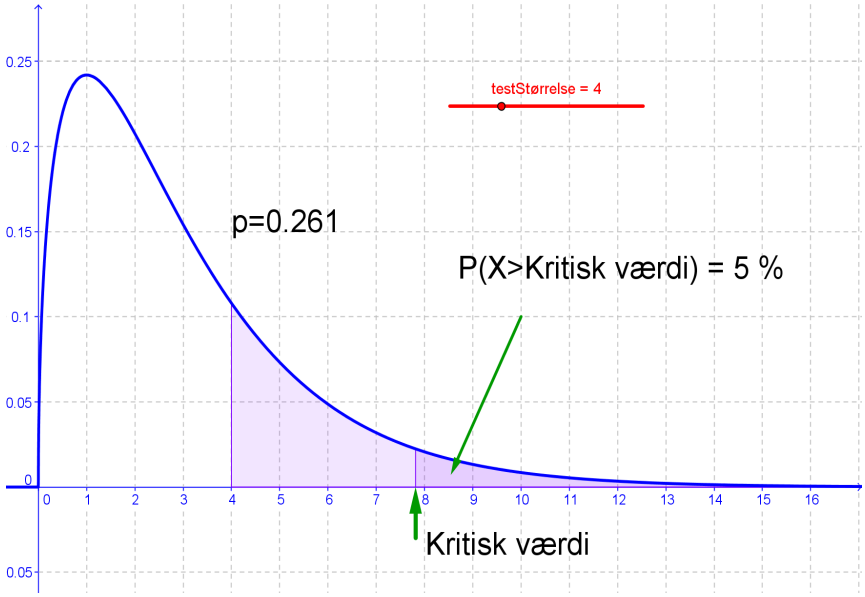
Sådan en hypotese kaldes *nulhypotesen*: H_0.[12]

Undersøgelsen bygger på beregning af en *teststørrelse*: $\chi 2$. Denne beregner summen af afvigelser mellem stikprøvens forventede tal og stikprøvens observerede tal, og denne sum varierer *stokastisk* (tilfældigt), fordi det er tilfældigheder der afgør, hvem der kommer med i stikprøven. Når denne sum er lille, *accepterer* vi, dvs. tror vi (til en vis grad) [13]på nulhypotesen, når den er stor, tror vi ikke på nulhypotesen og *forkaster* den. Selve teststørrelsen beregnes som $\sum \frac{\left(x_{observeret} - x_{forventet}\right)^2}{x_{forventet}}$, hvor forventede værdier for (her) mænd og kvinder beregnes med samme brøk som begge køn udgør inden for hver kategori

Men hvad er stor og lille? Vi fastlægger lidt tilfældigt et *signifikansniveau* på 5 % (eller evt. 1 %). En stor afvigelse betyder så en sum, der er så stor, at den eller en endnu større sum kun optræder med en sandsynlighed (*den såkaldte p-værdi*), der er mindre en signifikansniveauet (hvis H_0 er sand.) Logikken er: Optræder der en tilsyneladende usandsynlig begivenhed tror vi ikke på det! Så må modellen være forkert.

12 Nul er en ikke særlig klar oversættelse fra engelsk af "null", som snarere burde oversættes med ugyldig (eller frit: uden påviselig virkning.)

13 Du kan sammenligne med domsfældelse af forbrydere: Antagelsen er, at de mistænkte er uskyldig indtil det modsatte er bevist. Er der kun indicier eller usikre beviser accepterer vi at han går fri; kun ved stærke beviser dømmes han som skyldig.



Den sum, hvor p-værdien = signifikansniveauet, kaldes *den kritiske værdi.* Med andre ord: er teststørrelsen mindre end den kritiske værdi, accepteres nulhypotesen, er den større, forkastes nulhypotesen.

En accept er ingen garanti for, at hypotesen er rigtig. Men med de foreliggende data er det ikke tydeligt nok, at der findes en bedre model – og hvad den så skulle være.

Forkastes hypotesen, er hypotesen ofte forkert. Men er signifikansniveauet valgt til 5 % tages der fejl i 5 % af tilfældene (i det lange løb.) Dvs. at inddelingen har betydning for sandsynlighederne. I eksemplet med bopæl: der er forskel på, hvor mænd og kvinder bosætter sig.

Konklusionen i disse tilfælde kan lyde: Der er påvist en *statistisk sammenhæng* mellem køn og bopæl (med det valgte signifikansniveau.)

Tilføjelsen "statistisk" medtages for at adskille denne sammenhæng fra en sammenhæng baseret på årsag - virkning. Det er ikke (sikkert?) fordi kvinder er kvinder, at de ikke vil bo på landet. Men det kunne tænkes, at der var baggrundsvariable, der var med til at bestemme valget. Hvor et blandt mange gæt kunne være uddannelse og beskæftigelse.

Goodness of fit

bruges for at vise, om en population passer med en bestemt (kendt) teoretisk fordeling. Det kan være en undersøgelse af, om Mendells arvelighedslove gælder i forbindelse med et krydsningsforsøg eller der kan være tale om at sammenligne den nuværende population med population fra et tidligere tidspunkt. Det sidste kan eksemplificeres med vælgernes holdninger til partier, hvor resultatet fra det sidste valg er kendt og en ny stikprøve skal vise om den nuværende vælgermasse vil stemme på samme måde, som ved sidste valg.

Teststørrelsen beregnes på samme måde som før; dog beregnes forventede værdier som

absolutte tal ved hjælp af procenter fra teorien og stikprøvens størrelse. Summen af de forventede værdier = summen af observerede værdier.

Som før: Ved små afvigelser accepteres nulhypotesen (at teorien holder – eller rettere: ikke kan afvises); ved store afvigelser forkastes hypotesen (med typisk 5 % risiko for at begå en fejl.) Vi vil så konkludere, at fordelingen i stikprøven afviger signifikant fra teorien / den kendte fordeling.

Eksempler på testberegninger

Opgave 11 4. maj 2013 stxB

I en undersøgelse har et reklamebureau adspurgt to grupper unge, en fra Jylland og en fra Sjælland, om hvilken af to typer cornflakes A eller B, der smager bedst.

	A er bedst	B er bedst	Ved ikke
Jyder	55	40	30
Sjællændere	75	45	20

Foto: commons.wikimedia.org

Man ønsker at undersøge nulhypotesen:

Der er ikke forskel på unge jyders og unge sjællænderes smag for de to typer cornflakes.

a) Bestem de forventede værdier i ovenstående tabel under forudsætning af, at nulhypotesen er sand.

b) Undersøg på et 5% signifikansniveau, om nulhypotesen kan forkastes.

Forventede værdier

For jyder er den observerede værdi for kategorien "A er bedst" 55; den tilsvarende forventede værdi beregnes som

$$Forventet\ værdi = \frac{Antal\ Jyder}{Antal\ ialt} \cdot (Antal\ ialt\ ,\ der\ synes\ A\ er\ bedst) \tag{4}$$

Med de oplyste tal indsat fås:

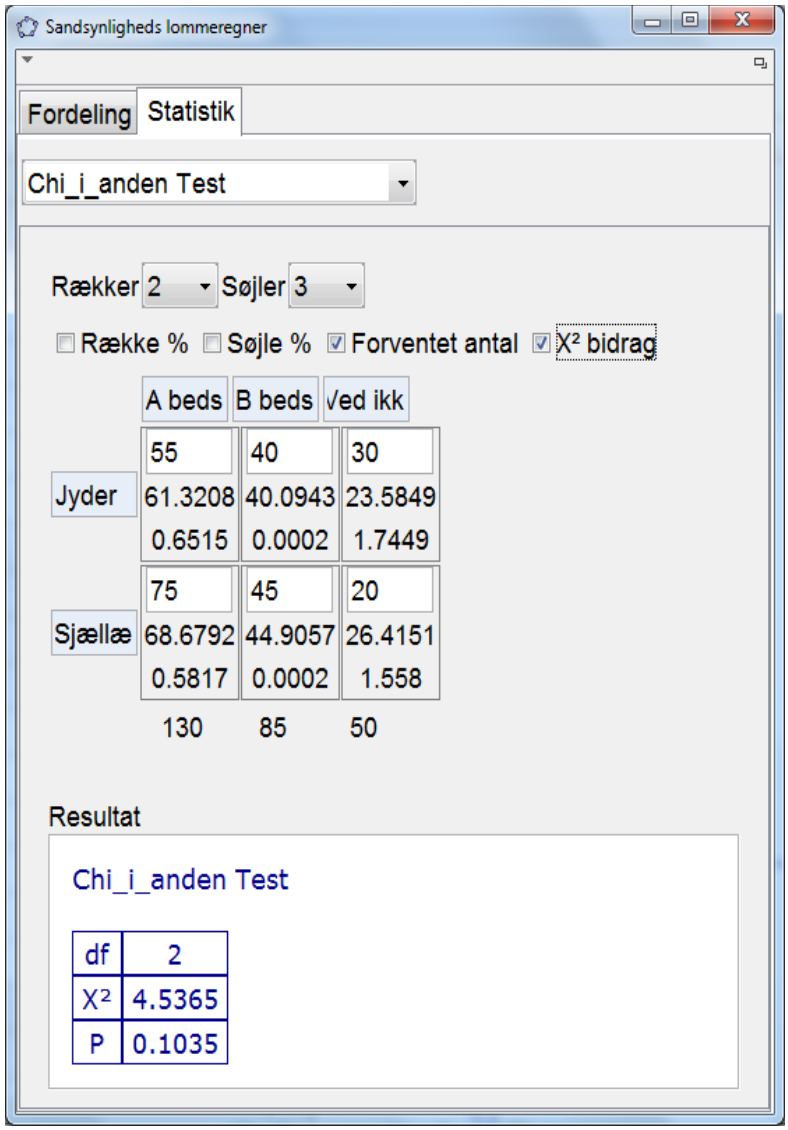
$$Forventet\ værdi = \frac{125}{265} \cdot (55+75) = 61{,}3 \tag{5}$$

Denne og de 5 øvrige værdier beregnes nemmest af GeoGebras sandsynlighedslommeregner. Da der er tale om et uafhængighedstest vælges i statistikfanen Chi_i_anden Test. De 6 oplyste observationer indtastes, der sættes flueben ved Forventet antal og alle forventede antal kan aflæses. De gengives her:

Forventede obs.	*A er bedst*	*B er bedst*	*Ved ikke*
Jyder	61,3	40,1	23,6
Sjællændere	68,7	44,9	26,4

Kan hypotesen forkastes?

Samtidigt med indtastningen beregnes automatisk i lommeregneren teststørrelsen (4,5), som er et mål for afvigelsen mellem observerede og forventede værdier,og p-værdien P=0,10. P er sandsynligheden for (forudsat at nulhypotesen er sand) at få endnu større afvigelser og da denne sandsynlighed er større en signifikansniveauet, forkastes nulhypotesen ikke.



Opgave 11 29. maj 2013 stxB

I en landsdækkende statistik fra 2010 fandt man følgende svarfordeling vedrørende internetadgang i danske husstande:

Internetadgang	ADSL	Kabelmodem	Fiberoptik	Mobilt bredbånd	Ved ikke
Andel af husstande	38%	33%	8%	11%	10%

Et år senere blev 800 tilfældigt udvalgte danske husstande spurgt om, hvilken internetadgang de benytter i deres husstand. Svarfordelingen fremgår af nedenstående tabel.

Internetadgang	ADSL	Kabelmodem	Fiberoptik	Mobilt bredbånd	Ved ikke
Antal husstande	300	235	78	106	81

a) Forklar, hvad der er stikprøve og population i ovenstående, og opstil en nulhypotese, der kan anvendes til at undersøge om internetadgangen i danske husstande har ændret sig siden 2010.

b) Undersøg på et 5% signifikansniveau, om nulhypotesen må forkastes.

Kilde: Danmarks Statistik – statistikbanken.dk

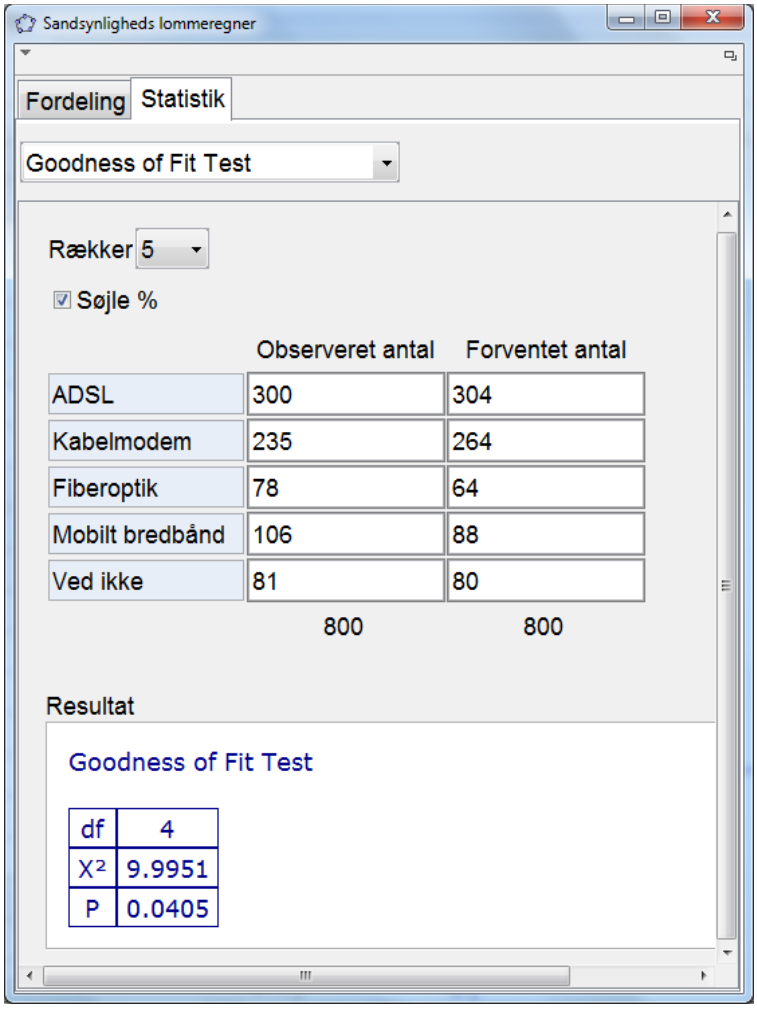
Population og stikprøve
Populationen er alle danske husstande i 2011 med internetadgang.
Stikprøven er den delmængde af disse, der blev tilfældigt udvalgt (dvs . 800 husstande.)

Nulhypotese
H_0: For internetadgang er der den samme procentvise fordeling i 2011 som i 2010

Forkastes nulhypotesen

Først beregnes de forventede absolutte størrelser af de forventede værdier:
I 2010 benyttede 38 % ADSL, så ifølge nulhypotesen forventes 38 % af 800 = 304 at benytte ADSL i 2011.
Tilsvarende beregnes de øvrige forventede værdier, som kan ses i figuren herunder.

Det ses at teststørrelsen, der er et mål for afvigelsen mellem observerede værdier og forventede værdier, er 10,0 svarende til en p-værdi på P=0,0405.
Da P < 5 % (signifikansniveauet), forkastes nulhypotesen.



7 eksamensopgaver

Opgave 1

For at afsløre om der er foregået fusk ved et valg, kan man udnytte, at første ciffer i partiernes stemmetal fra alle valgstederne fordeler sig som vist i tabellen nedenfor, hvis det er gået retfærdigt til.

Første ciffer	1	2	3	4	5	6	7	8	9
Sandsynlighed i procent	30,1	17,6	12,5	9,7	7,9	6,7	5,8	5,1	4,6

I en stikprøve fra et bestemt valg fordelte første ciffer i partiernes stemmetal fra 4430 tilfældigt udtrukne valgsteder sig som vist i tabellen nedenfor.

Første ciffer	1	2	3	4	5	6	7	8	9
Antal	1261	784	577	454	359	350	223	232	190

a) Opstil en nulhypotese, og beregn de forventede værdier med udgangspunkt i nulhypotesen.

b) Afgør, om nulhypotesen kan forkastes på et 5% signifikansniveau.

Opgave 2

En stikprøve fra en population er udvalgt på tilfældig måde. Personerne i stikprøven har svaret på spørgsmålet

> *"Bør der indføres forbud mod at bære hovedbeklædning i offentlige bygninger?"*.

Svarene på spørgsmålet er opdelt efter personernes uddannelse, og svarene fordeler sig som vist i tabellen.

	Ja	Nej
Erhvervsuddannelse	102	132
Videregående uddannelse af kort varighed	43	63
Videregående uddannelse af mellemlang varighed	80	157
Videregående uddannelse af lang varighed	30	84
Grundskoleuddannelse	72	99

På grundlag af svarene fra personerne i stikprøven skal det undersøges, om der er uafhængighed mellem uddannelse og holdning til spørgsmålet.

a) Opstil en nulhypotese, og bestem med udgangspunkt heri de forventede værdier.

b) Afgør på et 5% signifikansniveau, om nulhypotesen kan forkastes.

Opgave 3

I forbindelse med en forbrugerundersøgelse i en bestemt population vil man undersøge, om en tilfældig stikprøve på 300 forbrugere er repræsentativ med hensyn til alder. Nulhypotesen er derfor:

Aldersfordelingen i stikprøven følger aldersfordelingen i populationen.

I stikprøven er aldersfordelingen som vist i nedenstående tabel.

Alder	20-29	30-39	40-49	50-59	60-69	70-79	80-
Antal	39	42	83	54	49	21	12

Aldersfordelingen i populationen er vist i nedenstående tabel.

Alder	20-29	30-39	40-49	50-59	60-69	70-79	80-
Antal i %	15	16	19	17	17	10	6

a) Afgør, om nulhypotesen kan forkastes på et 5% signifikansniveau.

b) Angiv teststørrelsen, og redegør for, hvilken af aldersgrupperne der giver det største bidrag til teststørrelsen.

Opgave 4

Blandt en population er der på tilfældig vis udtaget en række personer. Disse personer er blevet spurgt om deres fysiske aktivitetsniveau og rygevaner. Personernes svar er sammenfattet i tabellen.

	Daglig-ryger	Fest-ryger	Ikke-ryger
Højt aktivitetsniveau	71	45	101
Middel aktivitetsniveau	67	61	115
Lavt aktivitetsniveau	98	50	93

a) Opstil en nulhypotese, og benyt et statistisk test med et signifikansniveau på 5% til at undersøge, om der er uafhængighed mellem det fysiske aktivitetsniveau og rygevaner hos individerne i populationen.

Opgave 5

Søren har fået en ny terning og vil undersøge, om den er "ærlig" (dvs. de forventede hyppigheder for antal øjne er lige store). Han kaster derfor terningen 1000 gange på tilfældig måde. Fordelingen af antal gange, de enkelte øjne vises, fremgår af tabellen.

Antal øjne	1	2	3	4	5	6
Hyppighed	172	168	150	188	185	137

a) Opstil en nulhypotese, der kan anvendes til at undersøge, om terningen er "ærlig", og opstil på baggrund heraf en tabel over fordelingen af antal gange, de enkelte øjne vises.

b) Afgør, om man på et 5% signifikansniveau kan forkaste nulhypotesen.

Opgave 6

I et afgrænset område ved et rev på kysten nær Belize har to biologer undersøgt en stikprøve af ål fordelt på art og opholdssted. Ålene i stikprøven var af henholdsvis G. moringa arten og G. vicinus arten, og ålene opholdt sig enten i græs, sand eller i en mellemting. Biologernes stikprøve gav følgende resultat:

	Græs	Sand	Mellemting
G. moringa	127	99	264
G. vicinus	116	67	161

a) Opstil en nulhypotese, der kan anvendes til at teste, om der er forskel på, hvilke opholdssteder de to ålearter foretrækker, og opstil med udgangspunkt heri en tabel over de forventede værdier.

b) Undersøg på et 5% signifikansniveau, om nulhypotesen kan forkastes.

Kilde:"Handbook of biological statistics" http://udel.edu/~mcdonald/statintro.html

Opgave 7

Tabellen viser, hvordan det samlede antal familier med bil var fordelt på de fem regioner i Danmark et bestemt år.

	Hovedstaden	Sjælland	Syddanmark	Midtjylland	Nordjylland
Forventet andel i %	26	15	24	24	11

En stikprøve på 1005 familier med bil har givet følgende fordeling på de fem regioner i Danmark:

	Hovedstaden	Sjælland	Syddanmark	Midtjylland	Nordjylland
Antal	271	161	250	201	122

a) Forklar, hvad der er stikprøve og population, og opstil en nulhypotese, som kan anvendes til at undersøge, om der er sket en ændring af fordelingen på de fem regioner.

b) Undersøg på et 5% signifikansniveau, om nulhypotesen kan forkastes.

Kilde: www.dst.dk

Stikordsregister

a posteriori....29
a priori....29
accepterer....46
Andrey Kolmogorov....27
Binomialfordelingen....30
Binomialforsøg....30
Blaise Pascal....27
Boksplot....13
boksplot....13
Cardano , Gerolamo....27
Christiaan Huygens....25
data....9
datafangsten....9
Deskriptiv statistik....5
deskriptorer....9ff.
Estimat....37, 39
estimater....39
estimatoren....39
fejl....41
Fermat , Pierre de 27
fordelingsfunktionen....35
forkaste hypotesen....41
forkaster....46
Fraktiler....20
frekvens....12
gennemsnit....9
Gerolamo Cardano 27
grupperede....9
Grupperede observationer....15
Histogram....16
histogram 17
Hypotese....37
hypotese....40
Hypoteser....40
hypoteser....5
hyppighed....12
Hyppighedstabel....11
individer....5
intervalfrekvens....17, 19
intervalhyppighed....17
Kolmogorov, Andrey....27
kritiske værdi....47
kumulerede....19
kumulerede hyppighed....18f.
kumulerede intervalfrekvens....19
kvartilsæt....14
Kvartilsæt mv. for grupperede data....21
Kvartilsættet....12
kvartilsættet....21
Median....12
median....10
medianen....11
Middeltal....12
middeltal....9
Middeltallet for grupperede data....23
middelværdi....9
mindste observation....11
Nedre kvartil....12
nedre kvartil....11
Normalfordelingen....35
nulhypotesen....46
observation....9, 12
observationer....9
observationssæt....12
observationssættet....9
Observationssættets størrelse....12
observationssættets størrelse....10
p %-fraktilen....20
p-værdi....46
Pascal, Blaise....27
Pascals trekant....25
Pierre de Fermat 27
populationens....39
populationer....5

Praktiske anvendelser....33
Roulette....25
Sandsynlighed....5
signifikansniveau....46
standardnormalfordelingen....35
Statistikbanken....9
statistisk sammenhæng....47
stikprøven....5, 39
stokastisk variabel....28
stokastiske....27
stokastiske eksperimenter....28
stykkevis lineær funktion....19
største observation....11
Sumkurven....19
søjlediagram....16
Test....37, 41
test....5
teste hypotesen....41
teststørrelse....46
typeinterval 17
Typetal....12
typetal....10
typiske karakter....10
tæthedsfunktionen....35
udfaldsrummet....28
ugrupperede....9
variabel....5
variabel, knyttet til individ....45
X~b(10; 0,3)....39
X^2-fordeling, Definition....45
Øvre kvartil....12
øvre kvartil....11