AALBORG UNIVERSITET

At træffe sine valg i en usikker verden - eller den statistiske modellerings rolle.

Af E. Susanne Christensen. Lektor i statistik.
Institut for Matematiske Fag. Aalborg Universitet.

I mange tilfælde og i mange forskellige faglige sammenhænge må man træffe en afgørelse eller basere en overbevisning på et ikke fuldstændigt informationsgrundlag. Dette er fx tilfældet, når man ønsker at forudsige udfaldet af et kommende folketingsvalg og kun har en opinionsundersøgelse til rådighed, eller når man ønsker at vide, om antallet af rygere er stigende blandt unge mennesker, men ikke har mulighed for at spørge alle unge, om de ryger eller ej. Resultaterne vil i sådanne tilfælde bære præg af, hvilken stikprøve[1] man lægger til grund for til sin undersøgelse.

Eksempel:
Vi vil undersøge, om holdningen til skattelettelser i Danmark er den samme for gymnasieelever som for gymnasielærere. For at finde ud af det, må vi indsamle nogle data, som vi kan basere vores konklusioner på: Vi skal lave en stikprøve!

Hvis vi vidste, det var så heldigt, at alle gymnasieelever i landet var enige med hinanden, og at også alle gymnasielærere var indbyrdes enige i spørgsmålet om skattelettelser, så kunne vi hurtigt blive færdige. Vi skulle bare spørge én person fra hver gruppe, hvad de mente om sagen, og så fastslå, om de to grupper også var enige; Og vi ville således være helt sikre på, at vi havde draget den rigtige konklusion. Men så nemt går det sjældent.

Der kan være ret stor forskel på meningerne inden for en gruppe personer. Og selvom det faktisk skulle forholde sig sådan, at de fleste gymnasieelever går ind for skattelettelser, så kan vi jo godt være uheldige – til vores stikprøve - at udvælge de elever, der er imod … og lige så uheldige kunne vi være med vores udvælgelse blandt lærerne. Hvis det er tilfældet, så vil vi ende op med den forkerte konklusion om gruppernes mening om spørgsmålet.

Det, vi kan gøre for at mindske risikoen for at lave forkerte konklusioner, er at tage ”fornuftige” stikprøver, der er ”store nok” til, at risikoen for at komme frem til en forkert konklusion bliver ”ac-

[1] Ord markeret med rød farve indikerer, at der findes en specifik matematisk definition af ordet. Når det ikke defineres her skyldes det, at den almindelige brug og opfattelse af ordet normalt får en til at agere i overensstemmelse med den korrekte matematiske definition. I dette tilfælde kan finde den korrekte definition i [2] (stikprøve=sample).

ceptabel". Statistik går (blandt andet) ud på at præcisere alle de begreber, der her står i anførselstegn. Hvad sandsynligheden er for at drage en forkert konklusion, kan beregnes. I hvert tilfælde hvis man følger nogle enkle regler, når man indsamler data. Hvilke udvælgelsesmåder man kan bruge i praksis, er beskrevet nærmere i [2] og [3].

I denne note skal vi se på, hvordan man kan sætte tal på ens usikkerhed i et par specifikke tilfælde.

Statistisk test for uafhængighed mellem to inddelingskriterier.

Hvis vi vil undersøge, om det at bruge "mange penge på tøj" er lige udbredt blandt unge kvinder og unge mænd, er den mest objektive måde at forholde sig på at lave en empirisk undersøgelse. Dvs. man indsamler data og drager sine slutninger på baggrund af dem.
Allerførst er der et par ting, der skal præciseres!

Hvad er det for nogle unge, vi interesserer os for? I den statistiske terminologi hedder det at fastlægge *populationen*. Lad os sige, at det er unge mellem 15-20 år og bosiddende i Danmark, vi er interesseret i.

Så skal vi have præciseret, hvad vi egentligt forstår ved at "bruge mange penge på tøj". I den statistiske terminologi siger man, at vi skal have formuleret *modellen og hypoteserne.*

Modellen kan hér være, at andelen af kvinder, der bruger mere end 1500 kr. om måneden på tøj er p_k og den tilsvarende for mændene er p_m. (Andelen p_k svarer til sandsynligheden[2] for at en kvinde tilfældigt udvalgt[3] fra populationen bruger mere end 1500 kr. på tøj om måneden). Grænsen for, hvornår man "bruger mange penge på tøj", er jo hér sat subjektivt og kan selvfølgelig gøres til genstand for diskussion ☺. Vi kommer tilbage til hvilke *matematiske krav,* der skal stilles til denne grænse.

Hypotesen er, at $p_m = p_k$, eller sagt i ord: Andelen af unge mænd, der bruger mange penge på tøj, er den samme som den tilsvarende andel for unge kvinder.

For at undersøge vores hypotese kan vi fx gennemføre følgende forsøg:

Vi udvælger et antal unge mellem 15 og 20 år tilfældigt og spørger dem om, hvor mange penge, de bruger på tøj om måneden. Så tæller vi op, hvor mange kvinder, der bruger mere end 1500 kr., og hvor mange mænd.

(En anden statistisk undersøgelse kunne foretages med udgangspunkt i svarene om størrelsen af de bruge beløb og fx undersøge, om det gennemsnitlige forbrug for kvinder er større end det gennemsnitlige forbrug for mænd. I så fald skulle man bruge den statistiske metode, der hedder sammenligning af to middelværdier. Det kan man læse mere om i fx [2])

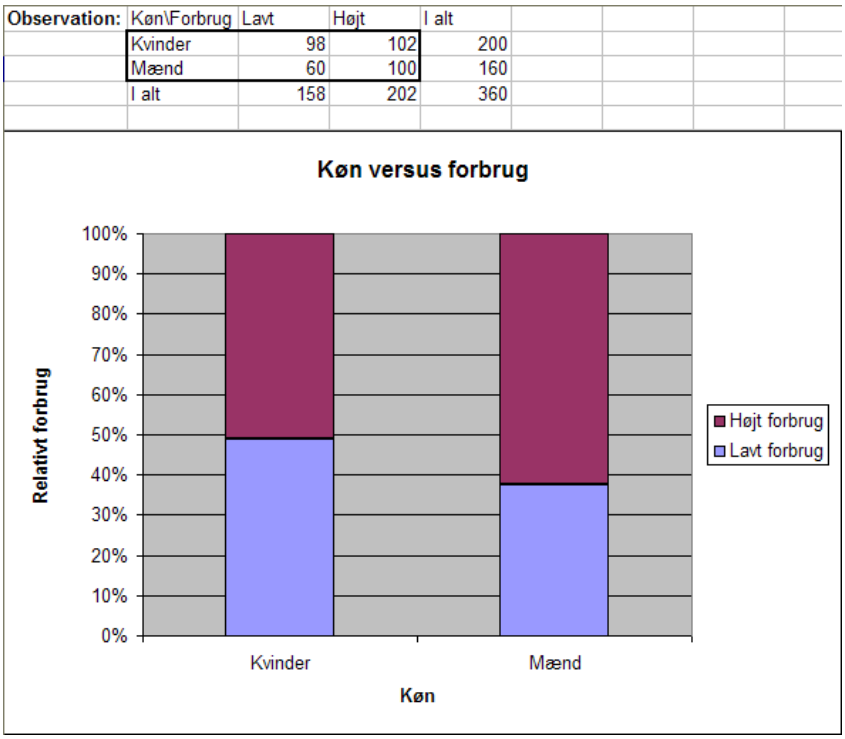
[2] For en introduktion til basal sandsynlighedsregning se fx [1] eller [2].

[3] Tilfældigt udvalgt betyder, at alle individer i populationen har samme sandsynlighed for at blive udvalgt. Vores beregninger forudsætter, at stikprøven er udvalgt på denne måde.

Vores resultat af undersøgelsen kan vi organisere i en tabel som nedenfor:
(Tallene er rent gætværk - ikke resultat af en virkelig undersøgelse).

	<1500 kr./måned	≥ 1500 kr./måned	I alt
Kvinder	98	102	200
Mænd	60	100	160
I alt	158	202	360

Vi kan eventuelt fremstille tallene fra stikprøven grafisk som nedenfor:



Vi kan nu gætte på andelen af unge kvinder, der bruger mange penge på tøj, simpelthen ved at regne ud, hvor stor andelen er i stikprøven. Vi siger, vi $estimerer\ p_k$. I dette tilfælde får vi $et\ estimat$ på 102/200= 0.51, hvilket vil sige, at vi tror, at 51 % af de unge kvinder bruger mange penge på tøj. Tilsvarende estimerer vi andelen af mænd til 100/160=0.625. Vi tror altså, at der var 62.5 % af de unge mænd, der bruger mange penge på tøj.

I den **stikprøve** vi har, er andelen af kvinder, der bruger mange penge på tøj, altså mindre end den tilsvarende andel for mænd.

Umiddelbart kan det altså se ud som om, at vores hypotese om samme andel for de to køn af personer, der bruger mange penge på tøj, IKKE holder stik. Resten af øvelsen går ud på at afgøre, om dette blot er en tilfældig følge af en uheldig stikprøve, ELLER om det vi har set, er så markant, at vi

tør tage det som udtryk for en forskel mellem de to køn generelt, altså noget vi tror, der gælder for hele populationen.

For at kunne regne på sagen og lave et statistisk test er det matematisk set vigtigt, at de enkelte svar på, hvor mange penge man bruger, er uafhængige[4] af hinanden. Hvis man fx i sin stikprøve har valgt en gruppe venner med stor indbyrdes påvirkning, så vil denne gruppe sagtens kunne have en adfærd, som er atypisk for populationen som helhed, og derved påvirke undersøgelsens resultat uhensigtsmæssigt.

Statistisk hypotesetest minder logisk set om det, du måske kender fra din matematikundervisning som et modstridsargument. Man antager en ting, gennemfører en række logiske argumenter og ender op med en konklusion, der klart er forkert. Heraf slutter man, at den oprindelige antagelse IKKE kan være rigtig. I statistik tager man hensyn til, at verden ikke er deterministisk, så hér kan man ikke konkludere, at udgangsantagelsen ikke er sand, men man kan eventuelt slutte, at det, man har set i sit forsøg vil være USANDSYNLIGT, hvis udgangsantagelsen er sand. Dermed tyder forsøget på, at antagelsen ikke er rigtig.

Vores antagelse om, at forbrugsmønsteret er ens for de to køn formuleres som vores udgangshypotese. Hvis vores test kan afvise den hypotese, så har vi et vist belæg for at påstå, at der er en signifikant forskel[5] mellem de to køn. Vi har således en udgangshypotese, som per tradition kaldes H_0 og en alternativ hypotese H_1 givet som:

H_0 $p_k = p_m$

H_1 $p_k \neq p_m$

En anden måde at udtrykke H_0 på er, at der er uafhængighed mellem det at bruge mange penge på tøj og ens køn . H_1 svarer så til, at der er afhængighed mellem de to inddelingskriterier - forbrug og køn, dvs.

H_0 *Der er uafhængighed mellem de to kriterier.*

H_1 *Der er ikke uafhængighed mellem de to kriterier.*

Vi starter med at antage, at H_0 udtrykker den sande tilstand af verden. I så fald kan vi estimere andelen af unge, der bruger mange penge på tøj, uden hensyntagen til, om de tilhører det ene eller det andet køn. Andelen af "storforbrugere" estimeres så til 202/360=0.5611, altså 56.11 %. Så hvis vi har en gruppe på 200 unge, vil vi forvente, at 200*0.5611= 112.22 af dem er storforbrugere, og 200*(1-0.5611)= 89.78 af dem var ikke-storforbrugere, uanset hvilket køn de har.
Her har vi brugt regelen, at hvis sandsynligheden for at være storforbruger er givet ved p, så er sandsynligheden for det modsatte, nemlig at være ikke-storforbruger, givet ved $(1-p)$. Ud over at være logisk er dette også en regneregel fra den basale sandsynlighedsteori.

[4] At to hændelser A og B er uafhængige betyder at $P(A \cap B)=P(A)*P(B)$. For regneregler for sandsynligheder se [2] eller [3].

[5] Begrebet statistisk signifikans er relateret til statistisk testteori. Se fx [2].

Ved at regne på den måde kan vi udfylde skemaet med de værdier, som vi ville have forventet at se, hvis verden opførte sig som vores H_0 foreskriver.

Forventede værdier under antagelse af at der er uafhængighed:

	<1500 kr./måned	≥1500 kr./måned	I alt
Kvinder	$\frac{158}{360} * 200 = 87.78$	$\frac{202}{360} * 200 = 112.22$	200
Mænd	$\frac{158}{360} * 160 = 70.22$	$\frac{202}{360} * 160 = 89.78$	160
I alt	158	202	360

Afvigelserne mellem det resultat, vi fik i forsøget, og de hér udregnede forventede værdier er et udtryk for, hvor langt forsøgets virkelighed er fra den verden, der er modelleret i H_0.

Imidlertid er det sådan, at summen af afvigelserne (98-87.77)+(102-112.23)+(60-70.23)+(100-89.77) = 0, og sådan vil det altid være. Så at lægge afvigelserne sammen giver os ikke noget samlet billede af, hvor stor afvigelsen er. I stedet viser det sig at være smart at udregne en χ^2 teststørrelse. Man udregner differens mellem det observerede antal og det forventede antal i hver celle, sætter denne differens i anden og dividerer med det forventede antal. Til sidst summeres disse tal for alle celler, altså:

$$\chi^2 = \sum \frac{(obs.antal - forv.antal)^2}{forv.antal}.$$

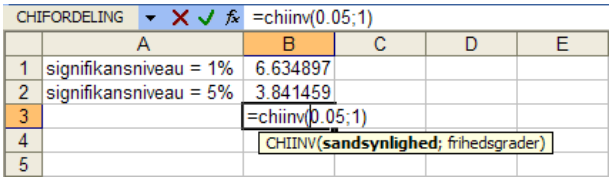
En stor værdi af teststørrelsen tyder i denne sammenhæng på, at udgangshypotesen om uafhængighed IKKE er opfyldt. Altså: store værdier af χ^2-teststørrelsen får os til at tro mere på H_1. Vi har så bare det problem tilbage, at vi skal afgøre, HVOR stor en teststørrelse skal være, før vi mener, den er så stor, at vi ikke vil tro på H_0. Til det brug skal vi vide, hvor store værdier teststørrelsen normalt vil antage, når H_0 er sand.

Hvis hypotesen om uafhængighed er rigtig, og hvis man har en stor nok stikprøve (sådan at alle de forventede værdier er større end 5), så ved man – takket være nogle matematikeres arbejde – hvilke værdier denne teststørrelse ville antage, hvis man lavede en uendelig række af forsøg som det skitserede. Den statistiske terminologi er, at man kender teststørrelsens fordeling[6] under H_0, idet den nemlig vil følge det, der hedder en χ^2-fordeling med 1 frihedsgrad. (Udtales "ki i anden"-fordelingen.)
(Og det er her, vi for en kort stund kan vende tilbage til vor subjektivt fastsatte grænse for, hvornår man bruger mange penge på tøj. Havde vi sat den grænse så højt eller så lavt, at der i en af cellerne med de forventede værdier var kommet et tal mindre end 5, så skulle vi enten have lavet grænsen om, eller være gået over til en anden statistik metode.[7])

[6] Begrebet fordeling kræver introduktion af begrebet stokastisk variabel for at formaliseres. Se [2] eller [3].
[7] Fx Fishers eksakte test.

Matematisk kan man vise, at i en verden, hvor køn og forbrugsmønster er uafhængige størrelser, så vil man i 5% af de gange, hvor man udvælger en stikprøve på 360 personer, få en teststørrelse, der er større end 3.84. I 1% af ville man få en teststørrelse, der er større end 6.63. [8]

Disse tal – også kaldet kritiske værdier - kan findes i tabeller, på moderne lommeregnere, i Excel og i statistiske værktøjsprogrammer. Med Excel ser det fx således ud:



Teststørrelsen fra vores stikprøve bliver χ^2 = (98-87.78)^2/87.78+(102-112.22)^2/112.22 +(60-70.22)^2/70.22+(100-89.78)^2/89.78= 1.19+0.93+1.49+1.16= 4.77. Den er jo altså er større end 3.84. Så HVIS antagelsen om uafhængighed mellem køn og forbrug skal holde stik, så har vi hér set et forsøg, der vil optræde med en sandsynlighed, der er betydeligt mindre end 0.05. Den statistiske terminologi er, *at testsandsynligheden er mindre end 5%.* Det er vist lettere at tro på, at antagelsen IKKE holder.

Vi forkaster vores udgangshypotese og siger: "Forsøget har påvist en sammenhæng mellem køn og forbrug på tøj, der er signifikant på 5% niveau."

Men vores teststørrelse er IKKE større end de 6.63. Det betyder at *testsandsynligheden er større end 1 %.* Vi kan derfor ikke påvise en signifikant sammenhæng på 1% niveau.

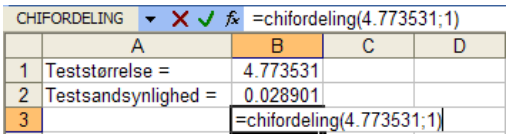
Hvis man, som det ofte er tilfældet, har en fast grænse for, hvornår man vil vælge at forkaste sin udgangshypotese, fx når testsandsynligheden er mindre end 5%, siger man, at man arbejder med et *signifikansniveau* på 5%. Ved at bruge et fast signifikansniveau på fx 5% i en hel række af forsøg og test ved man altså, at man i 5% af testene fejlagtigt vil forkaste en sand udgangshypotese. Vi ved, hvor sikker metoden er, men vi ved ikke, om den enkelte beslutning om at tro på udgangshypotesen er rigtig eller ej.

På dit CAS-værktøj kan du få den præcise sandsynlighed for at få en χ^2–teststørrelse, der er større end de 4.78, selvom H_0 er sand. (Ved at slå værdien 4.78 op i en χ^2-fordeling med 1 frihedsgrad.) Denne sandsynlighed kaldes *p-værdien eller testsandsynligheden* for testet.

En *p-værdi* er altså sandsynligheden for at få en teststørrelse, der får os til at tvivle mindst lige så meget på H_0, som den, vi lige har set, selvom H_0 faktisk er den rigtig hypotese. I det aktuelle eksempel får vi en p-værdi på p = 0.029.
Også p-værdien kan vi finde ved hjælp af Excel eller lignende hjælpemidler.

[8] De konkrete tal er fraktiler fra χ^2 –fordelingen med 1 frihedsgrad, der er den approximative fordeling af teststørrelsen. For en intuitiv forklaring af frihedsgrader, se [4].

Med Excel ser det således ud:



Hvad gør vi, hvis vi har flere niveauer på hvert af inddelingskriterierne?

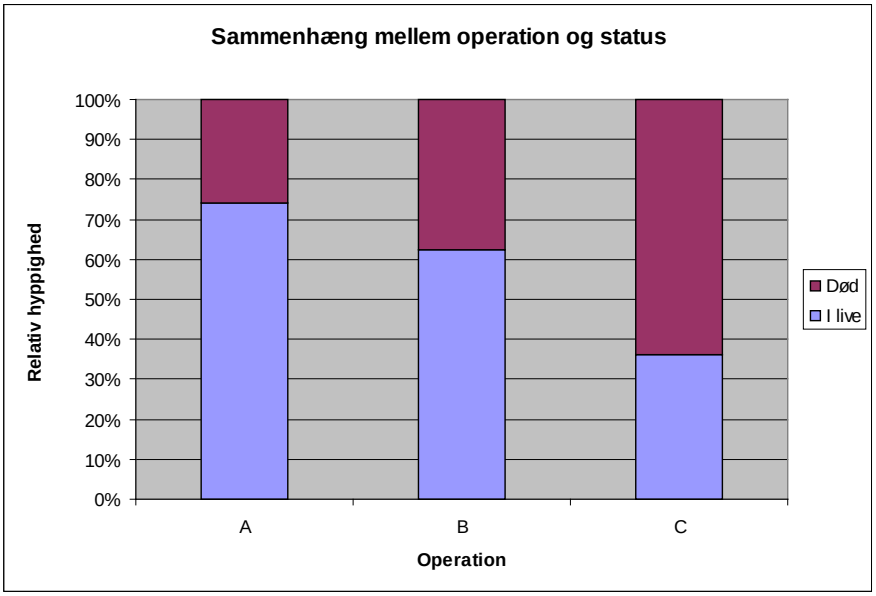
Et andet eksempel: En hjerneforsker undersøger forskellige måder at operere for en bestemt form for hjernetumor. Der er tre forskellige operationstyper: Type A, hvor man kun fjerner selve tumoren; Type B, hvor man også tager lidt af det nærmeste omkringliggende væv bort, og Type C, hvor såvel tumor som en større del af det omkringliggende væv fjernes.

Lægen ønsker at vide, hvordan operationstypen indvirker på chancen for at overleve et halvt år efter operationen. Over en årrække har lægen indsamlet følgende data over resultaterne af operationerne:

(data er fiktive, problemstillingen er autentisk)

	I live	Død	I alt
Operation A	40	14	54
Operation B	10	6	16
Operation C	9	16	25
I alt	59	36	95

En grafisk fremstilling af tallene kunne se sådan ud:



I **stikprøven** er der altså forskel på andelen af overlevende efter de forskellige typer af operation. Vi skal forsøge at undersøge, om det er en så stor forskel, at man kan sige, resultaterne kan generaliseres ud over denne stikprøve, altså: er det statistisk signifikant? Vi skal altså forsøge at regne ud, om den sete stikprøve er meget ekstrem, hvis vi antager, at der ikke er forskel på overlevelseschancerne efter de forskelige operationer.

Vi lader p_A være sandsynligheden er at være i live et halv år efter at have gennemgået en operation af type A, og tilsvarende for p_B og p_C.

Udgangshypotesen er, at der er samme sandsynlighed for overlevelse efter alle tre typer operation, dvs

H_0 $p_A = p_B = p_C$
H_1 *Ikke alle tre sandsynligheder ens.*

Hvis udgangshypotesen holder, estimerer vi sandsynligheden for overlevelse ved $\frac{59}{95} = 0.6211$, og vi kan udregne de forventede værdier efter samme princip som før:

Forventede værdier	I live	Død	I alt
Operation A	$0.6211*54 = 33.54$	$(1-0.6211)*54 = 20.46$	54
Operation B	$0.6211*16 = 9.94$	$(1-0.6211)*16 = 6.06$	16
Operation C	$0.6211*25 = 15.52$	$(1-0.6211)*25 = 9.48$	25
I alt	59	36	95

Alle de forventede størrelser er større end 5, så vi kan bruge χ^2 testet igen, nu bare med nogle lidt andre tal.[9]

Teststørrelsen udregnes som før ved at summe $\frac{(obs.antal - forv.antal)^2}{forv.antal}$ over alle celler.

Det vil sige, at vi her får

$$\frac{(40-33.54)^2}{33.54} + \frac{(14-20.46)^2}{20.46} + \frac{(10-9.94)^2}{9.94} + \frac{(6-6.06)^2}{6.06} + \frac{(9-15.52)^2}{15.52} + \frac{(16-9.48)^2}{9.48} = 10.5.$$

Denne gang skal vi bruge en χ^2-fordeling med (2-1)*(3-1)=(antal_rækker -1)*(antal_søjler-1)=2 frihedsgrader. Matematikeren, der er i besiddelse af de relevante tabeller, kan her fortælle os, at vi med en sandsynlighed på 5% vil få en teststørrelse større end 5.99, og med sandsynlighed 1% en teststørrelse større end 9.81, NÅR udgangshypotesen er sand.

Vi kan også selv finde en p-værdi via tabel, lommeregner eller i Excel med kommandoen CHIFORDELING(10.5;2), og vi får en testsandsynlighed på p = 0.0052.

[9] Den approximerende χ^2-fordeling har denne gang 2 frihedsgrader. Antallet af frihedsgrader er (antal rækker-1)*(antal søjler -1).

Så i dette tilfælde er vores valgmuligheder:
Vi fastholder troen på, at de tre operationstyper giver samme overlevelseschance, og vi har set et forsøg, der har mindre end 1 % sandsynlighed for at indtræffe.
ELLER
Vi forkaster hypotesen om ens chancer for overlevelse og siger:

Der er fundet en sammenhæng mellem overlevelseschance og operations-type, der er statistisk signifikant på 1% niveau.

Imidlertid skal man tænke sig om, inden man foreslår operationstype C forbudt. Hvis der er en tredje faktor der influerer billedet, så kan det give misvisende konklusioner, når man kun tager to af dem i betragtning. Vi kigger lidt nærmer på tallene fra før:

	I live	Død	I alt
Operation A	40	14	54
Operation B	10	6	16
Operation C	9	16	25
I alt	59	36	95

Eller i procenter:

	I live	Død	I alt
Operation A	74%	26%	100%
Operation B	63%	37%	100%
Operation C	36%	64%	100%

Efter en nærmere inspektion af journalerne viser det sig, at også patientens alder er noteret, og ved at opdele i to grupper efter alder får vi billedet:

50 år eller derunder:

	I live	Død	I alt
Operation A	27	1	28
Operation B	2	0	2
Operation C	1	0	1
I alt	30	1	31

eller i procent

	I live	Død	I alt
Operation A	96%	4%	100%
Operation B	100%	0%	100%
Operation C	100%	0%	100%

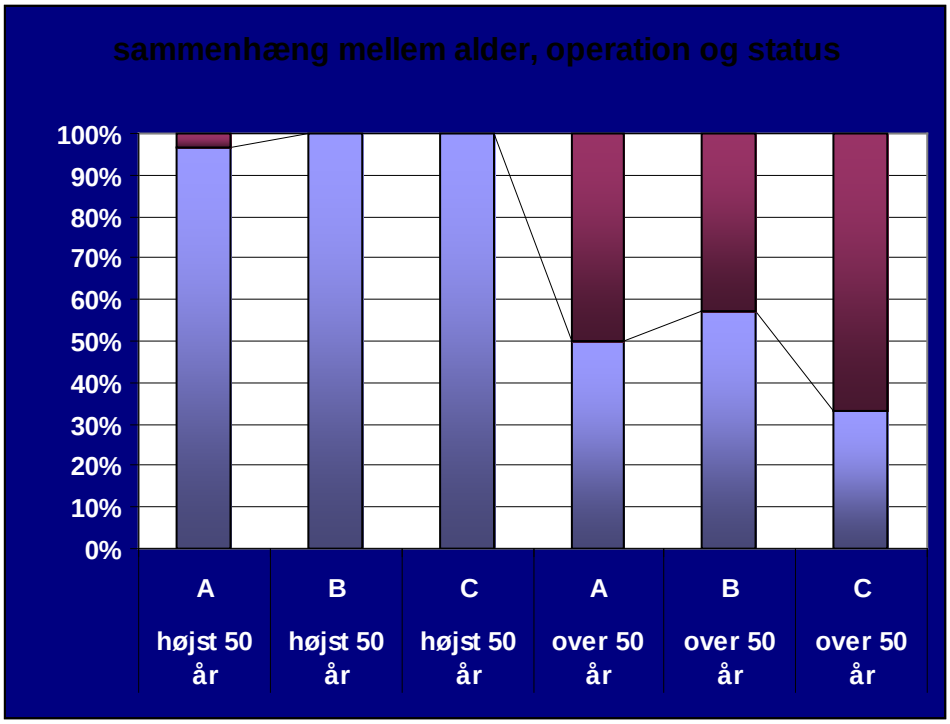
Over 50 år:

	I live	Død	I alt
Operation A	13	13	26
Operation B	8	6	14
Operation C	8	16	24
I alt	29	35	64

eller i procent

	I live	Død	I alt
Operation A	50%	50%	100%
Operation B	57%	43%	100%
Operation C	33%	66%	100%

Grafisk ser det nu sådan ud:



Prøv at lave et nyt test for uafhængighed mellem overlevelse og operationsform for aldersgruppen over 50!

Når man tager alderen med i betragtning, er operation C pludselig ikke længere så stor en skurk.

Her sker der det, at operationstypen og alderen ikke er uafhængige af hinanden eller af overlevelseschancen. Når der er flere vigtige faktorer, der spiller ind på en gang, så bør de alle tages med i analysen, der så bliver noget mere kompliceret. En statistisk metode, man kan anvende, hedder loglineære modeller. Men den sag vil vi ikke komme ind på hér - det må vente til universitetet ☺.

Opgaver:

Opgave 1: En amerikansk undersøgelse af bilisters brug af sikkerhedsseler resulterede i følgende stikprøve:

	Brug af sikkerheds sele			
Køn	Altid	Som regel	Af og til	Aldrig
Mænd	37	60	54	64
Kvinder	39	58	49	39

Spørgsmål a: Opstil den relevante nulhypotese og den alternative hypotese for at undersøge, om der er uafhængighed mellem køn og brug af sikkerhedssele?

Spørgsmål b: Udregn tabellen med de forventede værdier og χ^2 teststørrelsen

Spørgsmål c: Vil det være rimeligt at bruge en χ^2-fordeling til at vurdere teststørrelsen her?

Opgave 2: En forretningskæde vil undersøge, om farven på indpakningen af nye kartofler påvirker salget. Butikken sælger derfor i en periode poser med samme slags kartofler, alle med 2.5 kg/pose og til samme pris, men i poser med forskellig farve.

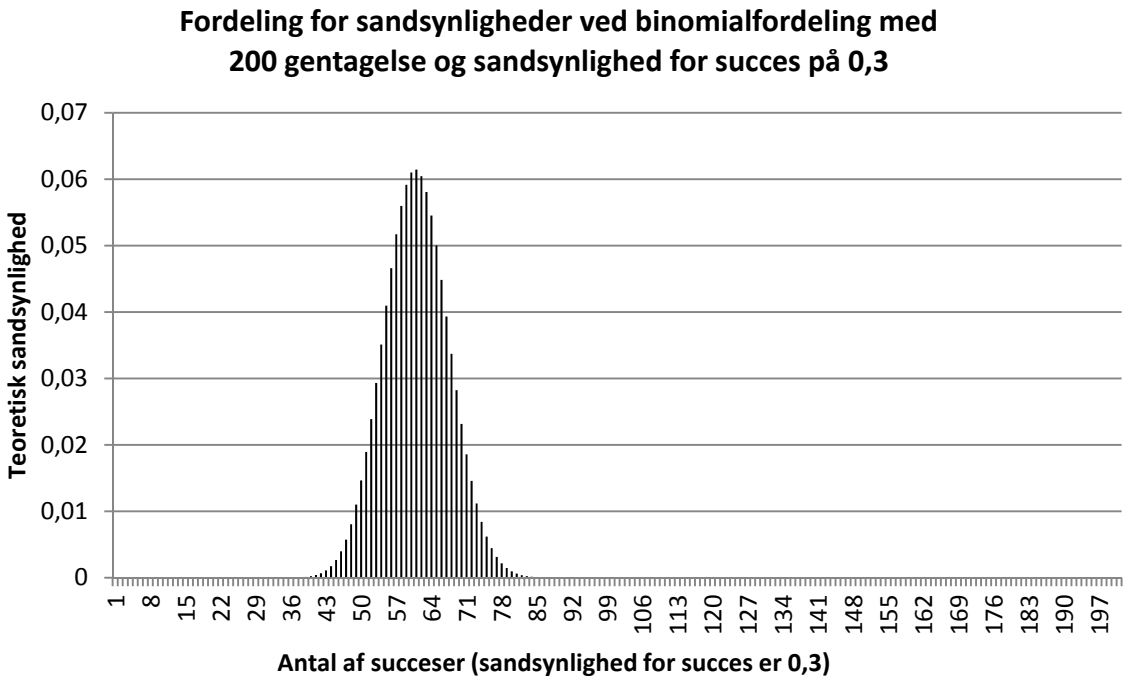
Der bliver i alt sendt 600 poser kartofler ud i butikkerne, hvoraf 520 poser bliver solgt.
Af de solgte poser er de 375 gule, og der er 55 gule poser tilbage. De øvrige poser er blå.

Undersøg, om der er grundlag for at påstå, at farven på posen påvirker salget af kartofler.
Undervejs skal du formulere de relevante hypoteser, kommentere på begreber som signifikansniveau og/eller p-værdi og forklare den anvendte metode.

Litteratur:

1. J. Burt & G. Barber. Elementary Statistics for Geographers. The Guildford press.
2. P. Newbold. Statistics for Business and Economics. Prentice Hall International Editions.
3. P. Mortensen. Repræsentative undersøgelser. Systime.
4. H.J. Beck, H.C. Hansen, A. Jørgensen, L.Ø. Petersen, P. Bollerslev. Matematik i læreruddannelsen. Gyldendals Uddannelse.

Følgende figur viser en fordeling for en anden binomialfordeling:



Her er søjlerne placeret så tæt så det nærmest ser ud som om der er en kurve. Hvad der egentlig er en diskret fordeling, tager sig nærmest ud som en kontinuert fordeling.

Opgave 4. Kast med 12 terninger – teoretisk

I denne opgave skal du undersøge et kast med 12 terninger hvor man *spørger om hvor mange 1'ere og 2'ere der tilsammen* ***forventes - teoretisk set***.

a) Forsøget kan beskrives ved binomialfordelingen *b(12,1/3)*. Forklar hvorfor det er binomialfordelingen med sandsynligheden 1/3 der skal anvendes.

b) Åbn *"Excelark til illustration af binomialsandsynligheder"* – det ligger i Fronter. Sæt parameterværdierne til *n=12* og *p=1/3*. Kopier det pindediagram som du får frem, ind i din besvarelse. Hvad kan du aflæse af diagrammet?

c) Alle de forskellige sandsynligheder kan beregnes ved hjælp af binomialformlen $P(k) = \frac{n!}{k! \cdot (n-k)!} \cdot p^k \cdot (1-p)^{n-k}$. Vis hvordan man udregner $P(2)$ ved at anvende formlen – skriv op hvad du indsætter i formlen (resultatet skal give *0,1272* hvilket fremkommer af regnearket).
(Bemærk at der er en trykfejl i binomialformlen i bogen s. 194)

d) Forklar hvad der skal forstås ved: *"forventes - teoretisk set"*?

Opgave 5. Spørgeskema til 50 personer - simulering

I denne opgave skal du arbejde med et spørgeskema til 50 personer. Det kunne eksempelvis være at man spurgte 50 elever på et gymnasium om de kunne tænke sig at der blev spillet livemusik til den næste fest. Svarene kan være interessante i sig selv, men matematisk set er det mest interessante om de adspurgte i virkeligheden svarer i overensstemmelse med hele populationens mening. Popula-

tionen er samtlige elever på det pågældende gymnasium. Opgaven går derfor ud på at finde svar på spørgsmålet: Med hvilken sikkerhed er det forsvarligt kun at adspørge et lille udsnit af populationen?

Da spørgeskemaet kun rummer svarmulighederne ja eller nej, er dette problem matematisk set det samme problem som kast af 12 terninger hvor man tæller hvor mange gange man får både 1'ere og 2'ere i samme kast. I denne opgave skal vi se på hvordan svarene kan variere fra én meningsmåling til en anden. Bemærk at vi i regnearket der omtales nedenfor, spørger 50 personer i alt 1000 gange.

Åbn regnearket "Spørgeskema til 50 personer" - det ligger i Fronter. Orienter dig i regnearkets konstruktion. Vælg den udgave der passer til din udgave af Excel. Hvis du åbner arket med Excel 2003, bliver du spurgt om en bestemt makro skal aktiveres, og du skal svare ja (der kan opstå problemer pga. sikkerhedsniveauet på din pc – spørg i så fald). Hvis du åbner med Excel 2007, skal du først svare "Ja" til at åbne filen. Derefter skal du klikke på "Indstillinger" over regnearket. Her skal du vælge "Aktiver indholdet".

Når du har åbnet Excelarket, skal du vælge "Gem som" og gemme arket på din egen pc (fx med navnet "Statistikopgave 5").

a) Hvad forstås der ved tallene i celle D11, H11 og J11?
b) Kør simuleringen to gange. En simulering køres ved at skrive et tilfældigt tal i cellen B4 og trykke ENTER.
c) Gør følgende i hver af de to simuleringer:
    1. Tegn et pindediagram (se tastevejledningen på sidste side i materialet her) som viser hyppigheden af de forskellige antal ja-svar, og kopier pindediagrammet over i din besvarelse.
    2. Aflæs og noter gennemsnittet fra regnearket.
    3. Aflæs og noter 'afvigelse i procent', dvs. hvor meget gennemsnittet afviger fra den sande (teoretiske) middelværdi.

    Vi kan oplyse at middelværdien i regnearket bliver udregnet ved $\mu = n \cdot p$. For at få regnearket til at beregne afvigelsen benyttes formlen $\frac{|\bar{x}-\mu|}{\mu} \cdot 100\%$ hvor $\bar{x}$ er det teoretiske gennemsnit.
d) Gør rede for hvad skal der laves om i regnearket hvis du skal simulere en situation hvor 20 % af populationen svarer ja. Foretag ændringen, og kør denne simulering to gange. Kopier et pindediagram for hver simulering ind i din besvarelse, og kommentér kort resultaterne.
e) Hvordan vil det se ud hvis kun 4 % svarer ja? Her skal du igen ændre i regnearket, og du skal derefter køre simuleringen to gange. Kopier et pindediagram for hver simulering ind i din besvarelse, og kommentér kort resultaterne.

Vurdering af data

Man får et meget mere sikkert resultat når man gennemfører et eksperiment 1000 gange end hvis man gennemfører det 20 gange. Hvis man kaster terninger, så vil tilfældige udsving udligne hinanden når man gentager et eksperiment mange gange, og man kommer tættere på den sande middelværdi. I det følgende skal vi se på hvordan ***usikkerheden*** i bestemmelse af sandsynligheder ud fra stikprøver knytter sig til antal stikprøver og spredningen.

Boks 1

Beregning af spredning

Som du har set i ovenstående øvelser, får man ikke det samme antal 1'ere og 2'ere hver gang man kaster med tolv terninger. Vi skulle forvente fire 1'ere eller 2'ere.

For at angive hvor langt væk fra gennemsnittet vore resultater kan forventes at ligge, definerer man en størrelse der hedder *spredningen*. Denne betegnes med det græske bogstav σ (sigma) og beregnes ved

$$\sigma = \sqrt{n \cdot p \cdot (1-p)}$$

hvor *n* her er antallet af terninger og *p* er sandsynligheden (i decimaltal, eksempelvis 33 % = 0,33) for at få det ønskede resultat.

Boks 2

95%-konfidensinterval – usikkerheden for μ

Lad os antage at vi udfører et *binomialforsøg* som et terningekast. Ud fra dette forsøg bestemmes et gennemsnit $\bar{x}$ og spredning σ. Hvis man har *tilstrækkeligt mange* observationer (*tilstrækkeligt mange* kast), gælder med god tilnærmelse at 95 % af observationerne vil ligge i intervallet

$$[\bar{x} - 2 \cdot \sigma \,;\, \bar{x} + 2 \cdot \sigma]$$

Med andre ord er det 95 % sikkert at den sande middelværdi μ ligger i $[\bar{x} - 2 \cdot \sigma \,;\, \bar{x} + 2 \cdot \sigma]$.

Dette interval kaldes 95%-*konfidensintervallet* for μ.

Man præciserer ikke hvad man mener med "tilstrækkeligt mange" – men det viser sig at være en rimelig tilnærmelse når $n > 30$, forudsat at p ikke er for tæt på 0 eller 1.

Boks 3

Usikkerheden for *p*

Usikkerheden for sandsynlighedsparameteren p er defineret som $\Delta p = \pm\frac{\sigma}{n}$.

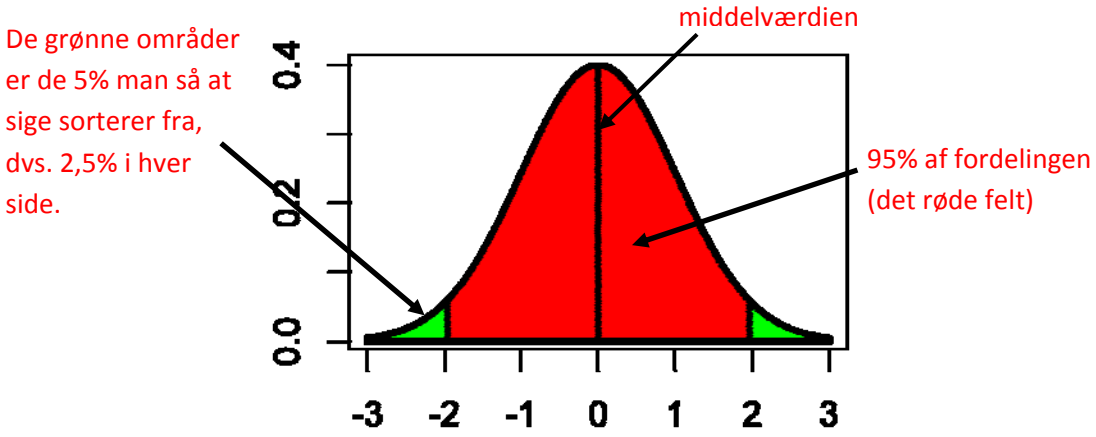
Konfidensintervallet for sandsynlighedsparameteren p er direkte knyttet til konfidensintervallet for μ fordi $\mu = n \cdot p$. Dermed kan p med 95 % sikkerhed siges at ligge i intervallet

$$[p_{min}, p_{max}] = [p - 2\Delta p\ , p + 2\Delta p] = \left[\frac{\bar{x} - 2\sigma}{n}, \frac{\bar{x} + 2\sigma}{n}\right]$$

Boks 4

En geometrisk fortolkning af 95%-konfidensinterval

Hvis man afbilder fordelingen af teoretiske sandsynligheder i et koordinatsystem, får man en figur med en masse tætstående søjler – som vist s. 3. Hvis man tegner en kurve langs søjlernes øvre kanter, får man fx en kurve som nedenstående. Vi illustrerer konfidensintervallet på kurven:



Eksempel

Ved en gallupundersøgelse spørges 1500 mennesker om de ser alt kongestof på TV2. 300 rødmer dybt og indrømmer. Her har vi $n = 1500$, og det bedste bud på middelværdien er netop $\bar{x} = 300$. Det bedste bud på antal ja-sigere i hele befolkningen er derfor $p = \frac{300}{1500} = 0{,}2 = 20\%$.

For at finde **usikkerheden for *p*** beregner vi først spredningen: $\sigma = \sqrt{1500 \cdot 0{,}2 \cdot 0{,}8} \cong 15{,}5$. Usikkerheden er da

$$\Delta p = \pm\frac{\sigma}{n} = \pm\frac{15{,}5}{1500} = \pm 0{,}0103.$$

95%-sikkerhedsintervallet er derfor $[0{,}20 - 2 \cdot 0{,}0103\ ;\ 0{,}20 + 2 \cdot 0{,}0103] = [0{,}18\ ;\ 0{,}22]$

Vi kan med 95% sikkerhed konkludere at mellem ca. 18% og 22% af befolkningen ser alt kongestof på TV2.

Meningsmålinger hvor stikprøven er repræsentativ for populationen, er et *binomialforsøg* ligesom terningekastene. Antallet af adspurgte svarer til antallet af terninger ved kast. Ja-procenten svarer til sandsynligheden for at få det ønskede.

Boks 5

Gallupundersøgelser

Dette er netop det en Gallupundersøgelse af befolkningens holdning til de politiske partier (og mange andre forhold) går ud på. Gallup udspørger en repræsentativ gruppe af befolkningen, udregner gennemsnitssvaret og beregner intervallet $[\bar{x} - 2 \cdot\ \sigma\ ;\ \bar{x} + 2 \cdot \sigma]$ for svarene. Man kan nu forvente at hvis hele befolkningen blev spurgt om det samme, så ville man få et svar som ligger inden for det beregnede interval med 95% sikkerhed.

Opgave 6. Usikkerheden på en undersøgelse med 50 adspurgte

Vi ser nu på situationen hvor 50 elever på et gymnasium bliver adspurgt om de kunne tænke sig livemusik til næste fest (se opgave 5).

a) Forklar med ord hvad der skal gælde om udvælgelsen af de 50 elever.

Vi antager nu at de 50 udspurgte er repræsentative for hele gymnasiet med hensyn til spørgsmålet. Antag endvidere at 10 elever svarer ja til spørgsmålet.

b) Da eleverne netop er repræsentative, hvor mange procent af skolens elever ville da svare ja?

Spørgsmålet er nu hvor nøjagtigt vores bud er på at hele skolen vil svare dette. Antag derfor at ja-procenten er 20%.

c) Beregn usikkerheden på antal ja-sigere, og beregn konfidensintervallet. Undersøg om 20 positive svar ligger i 95%-konfidensintervallet.

Tastevejledning til Excel

Gennemsnit

Skal man beregne gennemsnit af alle tal i en kolonne, fx fra E2 til E20, markeres en tom celle, fx nedenunder, og der skrives "=middel(e2:e20)"

Kopiering af formel

Marker cellen hvor din formel står. Sæt markøren i nederste højre hjørne på cellen og markøren bliver til et plus-tegn. Klik med musens venstre knap og hold knappen nede mens du trækker musen ned til alle de celler der skal indeholde formlen og slip så museknappen. Nu beregnes formlen i alle cellerne.

Diagram – histogram

Når du skal indsætte et histogram (stolpediagram) i et regneark skal du have to de kolonner med x'erne og y'erne stående i regnearket. I nedenstående eksempel står tallene i kolonne G3 til 15 og H3 til 15.

Placer markøren i en celle til venstre for dine måledata og tast følgende

- Vælg **Indsæt**
- Vælg **Diagram**
- Vælg diagramtype **søjle** og den øverste til venstre som undertype
- Vælg **Næste**
- Marker som dataområde H3 til H15
- Vælg øverst i diagramvinduet fanen **Serie**
- Skriv i feltet Navn **Hyppighed**
- Marker som Kategoriakseetiketter G3 til G15
- Vælg **Næste**
- Skriv som diagramtitel "**Spørgeskema til 50 personer**"
- Vælg **Udfør**