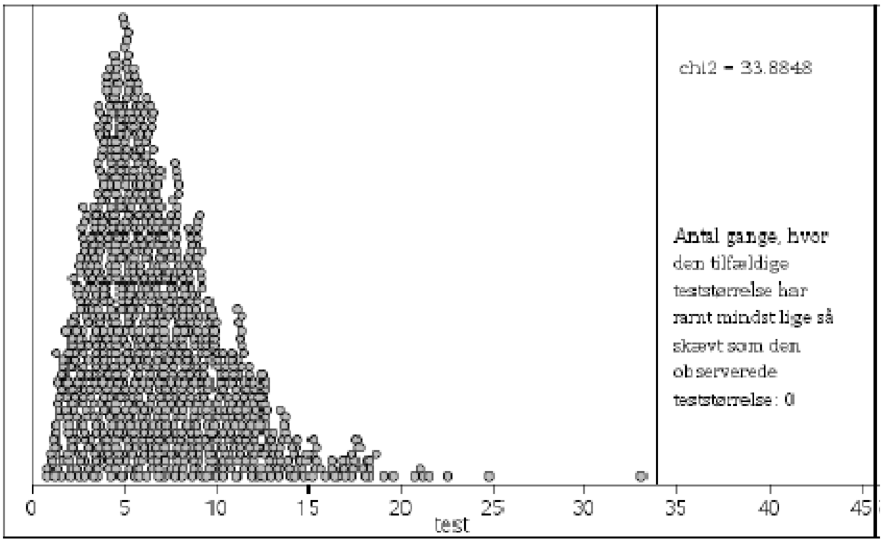
KURSUSMATERIALE TIL DET NYE STATISTIKPENSUM



Det foreliggende udkast til kursusmateriale er lagt ud til orientering for kollegerne med henblik på at indhente kommentarer til materialet. Sammen med Susanne Christensens elevnote (se forordet til kursusmaterialet) definerer kursusmaterialet det kommende obligatoriske pensum i χ^2-test for niveauerne MatB og MatA, der forventes at træde i kraft for elever, der starter sommeren 2010.

Eventuelle kommentarer sendes til fagkonsulent Bjørn Grøn:
Bjorn.Gron@uvm.dk

(redaktionen afsluttet 7-9-2009)

Forord

Lektor i statistik Susanne Christensen, Institut for Matematiske fag ved Aalborg Universitet har på min opfordring skrevet en note om χ^2 (chi-i-anden). Noten[1] dækker de emner, der påtænkes inddraget som kernestof i matematik stx på B- og på A-niveau, nemlig test for uafhængighed mellem to inddelingskriterier og test for fordeling af en given stikprøve (test for goodness of fit). Læreplansjusteringerne vil gælde for de elever, der starter i gymnasiet sommeren 2010.

Noten er blevet læst og kommenteret af en baggrundsgruppe nedsat af fagkonsulenten: Aksel Bertelsen, Helge Gram Christensen, Steen Dilling, Bjørn Felsager, Olav Lyndrup og Jan Sørensen. Det er vores vurdering at noten kan danne grundlag for undervisning i emnet, og materialet stilles frit til rådighed til kopiering til ens elever.

I tilknytning til denne note er der udarbejdet et supplerende materiale som bl.a. inddrager en eksperimentel tilgang til hypotesetest. I appendiks er demonstreret hvorledes dette kan gøres med brug af forskellige værktøjsprogrammer. Den eksperimentelle tilgang kan både indgå som en del af den mundtlige prøve og som en metode til besvarelsen af spørgsmål ved den skriftlige prøve. Fuldt pointtal i en besvarelse vil kunne opnås både med anvendelse af en eksperimentel tilgang og med anvendelse af indbyggede funktioner i et værktøjsprogram.

Der er endelig lagt flere opgaver ind i dette materiale, lige som litteraturlisten er udbygget.

Dette supplerende materiale kan enten anvendes direkte af den enkelte lærer eller danne grundlag for et skolebaseret kursus i emnet

Bjørn Grøn, fagkonsulent

[1] Noten findes på adressen: http://www.math.aau.dk/highschool/chi2JULI09.pdf

Indholdsfortegnelse:

Forord	side 1
Indledning	side 3
1. Statistisk test for uafhængighed mellem to inddelingskriterier (χ^2–test for uafhængighed)	side 4
Om teststørrelsens fordeling under H_0	side 9
Opgaver	side 12
2. Hvad gør vi, hvis vi har flere niveauer på hvert af inddelingskriterierne?	side 14
Om teststørrelsens fordeling under H_0	side 18
Opgaver	side 21
3. Statistisk test for fordeling af en stikprøve (χ^2–test for goodness of fit)	side 24
Om teststørrelsens fordeling under H_0	side 26
Opgaver	side 29
Større opgaver, der kan indgå i projekter sammen med biologi	side 31
Litteraturliste	side 34
Noter	side 35

Appendiks 1: Eksperimentel tilgang til χ^2-test med forskellige værktøjsprogrammer

Appendiks 2: Standardværktøjer til udførelse af χ^2-test med forskellige programmer

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

Et kursusmateriale baseret på en note 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.

Det er fx tilfældet, når man ønsker at gætte på udfaldet af et kommende folketingsvalg og kun har en opinionsundersøgelse, eller når man ønsker at afgøre, om antallet af rygere er stigende blandt unge mennesker, men ikke har mulighed for at spørge alle unge, om de ryger eller ej.

Ens afgørelse eller resultat kommer i sådanne tilfælde til at være præget af, hvilken *stikprøve*[1] man får fat i til sin undersøgelse.

Et eksempel:
Vi vil undersøge om holdningen til skattelettelser i Danmark er den samme for gymnasieelever som for gymnasielærere. For at være sikre på svaret skulle vi spørge hele *populationen*, dvs. samtlige gymnasieelever og gymnasielærere. Men en sådan strategi er normalt både praktisk og økonomisk umuligt. Så for at finde ud af det, må vi indsamle nogle data, som vi kan basere vores konklusioner på: Vi skal have en stikprøve!

Hvis vi var så heldige at vide, at alle gymnasieelever i landet var enige med hinanden og 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 afgøre om de to grupper også var enige. Og vi ville 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 indenfor gruppen. Og selvom det faktisk skulle forholde sig sådan at de fleste gymnasieelever går ind for skattelettelser, så kan vi jo godt være uheldige at udvælge de elever, der er imod og omvendt med lærerne til vores stikprøve. Hvis det er tilfældet, så vil vi ende op med den forkerte konklusion om gruppernes mening om spørgsmålet. Vi skal altså ikke bare have en stikprøve, men en stikprøve, der er udvalgt, så vi har en begrundet tro på, den er *repræsentativ*[2] for populationen.

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 til en forkert konklusion bliver ”acceptabel”. 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.

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

1. Statistisk test for uafhængighed mellem to inddelingskriterier (χ^2 – test for uafhængighed)

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 til problemstillingen, at lave en *empirisk* undersøgelse, dvs. ved at man indsamler data, og drager sine slutninger på baggrund af disse.

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 og 20 år 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 kunne her 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 er lig med *sandsynligheden*[3] for, at en kvinde tilfældigt udvalgt fra populationen bruger mere end 1500 kr. på tøj om måneden). Grænsen for hvornår man bruger mange penge på tøj er jo her sat subjektivt, og kan selvfølgeligt gøres til genstand for diskussion[4]. 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 eksperiment:
Vi udvælger et antal unge mellem 15 og 20 år *tilfældigt* og spørger, hvor mange penge de bruger på tøj om måneden. Derefter tæller vi op, hvor mange kvinder, der bruger mere end 1500 kr, om måneden på tøj og tilsvarende for mænd.

Vores resultat af undersøgelsen kan vi organisere i tabellen nedenfor (hvor tallene er fiktive - ikke resultat af en virkelig undersøgelse. Sådan en kan holdet selv lave!).

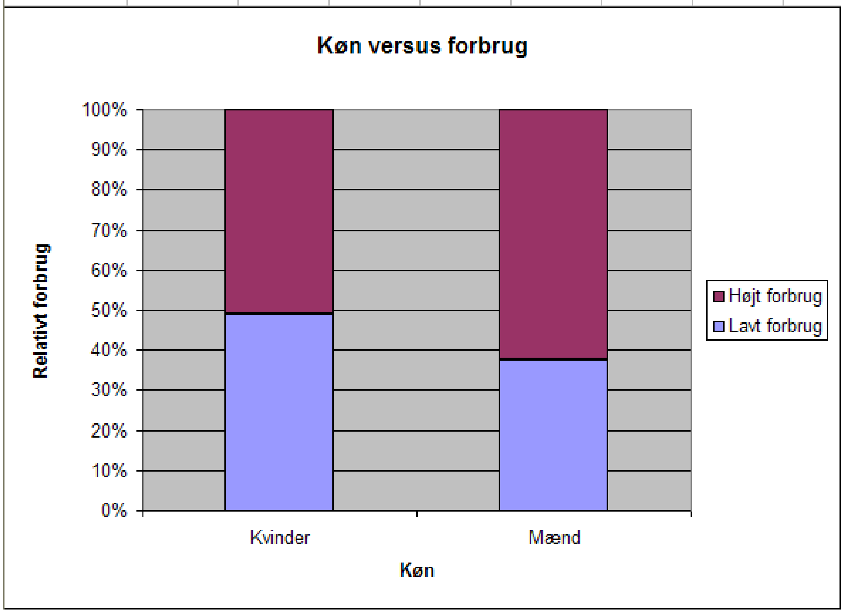
Køn\Forbrug	<1500 kr./måned	≥ 1500 kr./måned	i alt
kvinder	98	102	200
mænd	60	100	160
i alt	158	202	360

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. Det vil sige, vi tror, at 51 % af de unge kvinder bruger mange penge på tøj[5].

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.

Når man skal undersøge et talmateriale som det ovenstående for at finde sammenhænge er det ofte en god ide at illustrere talmaterialet grafisk. I dette tilfælde kan man fx afbilde tabeldataene som stablede søjlediagrammer (her vist i Excel – for andre programmer se appendiks).

Observation:	Køn\Forbrug	Lavt	Højt	I alt
	Kvinder	98	102	200
	Mænd	60	100	160
	I alt	158	202	360



Af diagrammet fremgår, at i den *stikprøve* vi har, er andelen af kvinder, der bruger mange penge på tøj, mindre end den tilsvarende andel for mænd.

Umiddelbart kan det altså se ud som om vores hypotese ikke holder stik, nemlig at der skulle være samme procentdeldel unge mænd som unge kvinder, der bruger mange penge på tøj.

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*[6] 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.[7]

Statistisk hypotesetest minder logisk set om det, du måske kender fra din matematikundervisning som et modstridsargument. Man antager én ting - gennemfører en række logiske argumenter og ender op med noget, 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 her kan man ikke konkludere at udgangsantagelsen ikke er sand, men man kan eventuelt slutte, at det man har set i sit forsøg er USANDSYNLIGT, hvis udgangsantagelsen skulle være sand. Dermed peger 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*[8] forskel mellem de to køn. Vi har således en udgangshypotese, som pr tradition kaldes H_0 (*nulhypotesen*) og en *alternativ hypotese* H_1 givet som:

H_0 $\quad p_k = p_m$

H_1 $\quad p_k \neq p_m$

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 forbrug og køn.*
H_1 *Der er afhængighed mellem forbrug og køn.*

Hypoteserne H_0 og H_1 kan illustreres med situationen ved en retssag. H_0 svarer til, at den anklagede er uskyldig, mens H_1 svarer til, at den anklagede er skyldig. Som i en retssag, lader vi tvivlen kommen den anklagede til gode, så med mindre vi har stærke beviser for, at den anklagede er skyldig, så vælger vi at acceptere H_0 (frikende den anklagede). Vi forkaster altså kun H_0, hvis vi har rigtigt gode beviser for, at H_0 er forkert.

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 uden hensyntagen til om de tilhører det ene eller det andet køn. Andelen af unge med et højt forbrug estimeres så til 202/360 ≈ 0,5611, altså 56,11 %.

Så ud af en gruppe på 200 unge, vil vi derfor forvente, at 200·0,5611 ≈ 112,22 af dem har et højt forbrug, og tilsvarende at 200·(1-0,5611) ≈ 87,78 af dem har et lavt forbrug. Her har vi brugt, at hvis sandsynligheden for at have et højt forbrug er givet ved p, så er sandsynligheden for det modsatte, nemlig at have et lavt forbrug, givet ved (1-p). Ud over at være logisk er dette også en regneregel fra den basale sandsynlighedsteori.

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:

Køn\Forbrug	<1500 kr./måned	≥1500 kr./måned	i alt
kvinder	$\frac{158}{360}\cdot 200 = 87{,}78$	$\frac{202}{360}\cdot 200 = 112{,}22$	200
mænd	$\frac{158}{360}\cdot 160 = 70{,}22$	$\frac{202}{360}\cdot 160 = 89{,}78$	160
i alt	158	202	360

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

For umiddelbart at kunne sammenligne de to tabeller med de observerede værdier og de forventede værdier samler vi dem i en enkelt tabel med observationer til venstre og forventede værdier til højre:

Køn\Forbrug	<1500 kr./måned	≥ 1500 kr./måned	i alt
kvinder	98 87,78	102 112,22	200
mænd	60 70,22	100 89,78	160
i alt	158	202	360

Det er nu altid sådan, at summen af afvigelser er 0. I dette tilfælde fås således:

$$(98-87,78) + (102-112,22) + (60-70,22) + (100-89,78) = 0 .$$

Så at lægge afvigelserne sammen giver os ikke noget samlet billede af, hvor stor afvigelsen er.

I stedet kan man udregne et mål for, hvor stor afvigelsen på følgende måde:Først udregnes for hver celle kvadratet på differensen mellem det observerede antal og det forventede antal (hvorved vi kun får positive bidrag), og derefter divideres med det forventede antal. Dermed fås for den enkelte celle et relativt mål for afvigelsen. Til sidst summeres disse tal fra alle cellerne. Denne størrelse kaldes χ^2-*teststørrelsen.* Som samlet formel ser det sådan ud[9]:

$$\chi^2 = \sum \frac{(observeret\ antal - forventet\ antal)^2}{forventet\ antal}$$

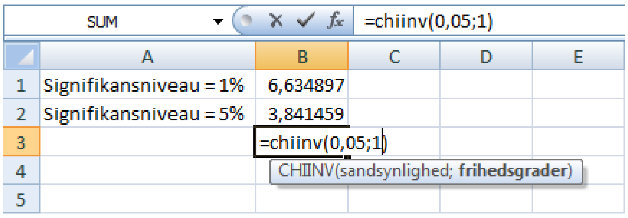
.

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 H1. 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 test-størrelsen normalt vil antage, når H_0 er sand.

Hvis hypotesen H_0 om uafhængighed er rigtig, og hvis man har en stor nok stikprøve (dvs. alle de forventede værdier er mindst 5)[10], så kan man i den matematiske statistik undersøge, hvilke værdier denne teststørrelse ville antage, hvis man lavede en uendelig række af forsøg som det skitserede. Det viser sig da at teststørrelsen med stor tilnærmelse har en opførsel, der *ikke* afhænger af hvor mange vi i alt har spurgt, dvs. i vores tilfælde 360 personer. Den statistiske terminologi er, at man kender teststørrelsens fordeling under H_0. Den vil nemlig med stor tilnærmelse følge det, der hedder en χ^2-*fordeling med* 1 *frihedsgrad* (hvilket vi vil uddybe i et følgende afsnit).

Hvis H_0 er rigtig, dvs. de to inddelingskriterier (her køn og forbrugsmønster) er uafhængige størrelser, så ville 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 tilfældene ville man tilsvarende få en teststørrelse, der er større end 6,63.

Disse tal – der også kaldes *kritiske værdier* hørende til signifikansniveauerne - kan findes i tabeller, på moderne lommeregnere, i Excel og i statistiske værktøjsprogrammer. (*Man skal anvende den inverse χ^2-fordeling til at finde de kritiske værdier hørende til et givet signifikansniveau*). I Excel, hvor den inverse χ^2-fordeling hedder chiinv(), ser det fx således ud (se appendiks for andre værktøjer):



Teststørrelsen fra vores stikprøve bliver:

$$\chi^2 = \frac{(98-87{,}78)^2}{87{,}78} + \frac{(102-112{,}22)^2}{112{,}22} + \frac{(60-70{,}22)^2}{70{,}22} + \frac{(100-89{,}78)^2}{89{,}78}$$
$$= 1{,}19 + 0{,}93 + 1{,}49 + 1{,}16 = 4{,}77$$

Den er altså større end 3,84. Så HVIS antagelsen om uafhængighed mellem køn og forbrug skulle holde stik, så har vi her set et forsøg, der ville optræde med en sandsynlighed, der er mindre end 5%.

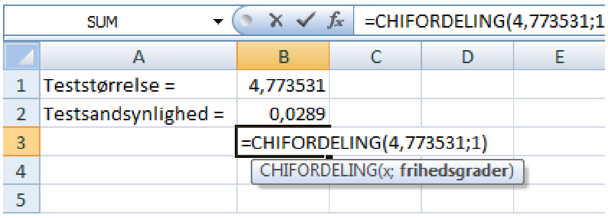
Den statistiske terminologi er, at *testsandsynligheden er mindre end* 5 %. Dette kan naturligvis indtræffe, men når sandsynligheden er så lille, vælger vi at forkaste vores udgangshypotese og siger:

"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 forsøget ville optræde med en sandsynlighed, der er større end 1%. Vi siger, 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*[11] 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[12]. Vi ved, hvor sikker metoden er, men vi ved ikke, om den enkelte beslutning om at tro på udgangshypotesen er rigtig eller ej. Ved at fastlægge et *lavere* signifikansniveau, fx på 1%, kræver vi en betydelig *større* forskel på de to køns forbrug af tøj, før vi siger, der er signifikant forskel. Til gengæld er det nu kun i 1% af tilfældene, at vi forkaster en sand hypotese.

På en lommeregner, i et regneark eller et statistisk værktøjsprogram kan man få den præcise sandsynlighed for at få en χ^2-teststørrelse, der er større end de observerede 4,77, under antagelse af at H_0 er sand. Det sker ved at slå værdien 4,77 op i en χ^2 fordeling med 1 frihedsgrad. (*Man skal anvende den kumulerede χ^2-fordeling til at finde testsandsynligheder eller p-værdier*). I Excel ser det fx således ud:



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 teststørrelse vi lige har observeret, under antagelse af at H_0 faktisk er den rigtig hypotese. I det aktuelle eksempel får vi således en *p*-værdi på

$$p = 0.0289 = 2.89\%$$

jfr. celle B2. Da *p*-værdien er lille, dvs. under 5%, er forskellen signifikant på signifikans-niveauet 5%. Men den er stadigvæk større end 1%, dvs. forskellen er ikke signifikant på signifikansniveauet 1%.

Vi kan altså nå til de samme konklusioner som før uden at udregne størrelsen af de kritiske teststørrelser hørende til signifikansniveauerne.

Om teststørrelsens fordeling under H_0.

I det ovenstående eksempel har vi spurgt 360 unge om deres køn og deres tøjforbrug. Hvis vi først ser på køn og forbrug hver for sig fordeler de sig således: I undersøgelsen deltog 200 kvinder og 160 mænd. Tilsvarende fandt vi at 158 unge havde et lavt tøjforbrug, mens 202 havde et stort tøjforbrug. Det svarer til randværdierne i krydstabellen:

Køn\Forbrug	<1500 kr./måned	≥ 1500 kr./måned	i alt
kvinder			200
mænd			160
i alt	158	202	360

Men det der interesserer os er i virkeligheden koblingen mellem køn og forbrug, dvs. hvor mange kvinder har et lavt tøjforbrug osv. Det svarer til tallene i cellerne inde i krydstabellen. Disse tal afhænger nu af hinanden på den følgende måde: Summen af alle tallene for cellerne i første søjle (unge med lavt forbrug) skal give 158. Summen af alle tallene i første række (unge kvinder) skal give 200 osv. Da et antal ikke kan være negativt følger det nu at antallet af kvinder med et lavt tøjforbrug i princippet kan antage alle værdier fra 0 til 158. Så snart vi har fastlagt antallet af kvinder med lavt tøjforbrug følger resten af cellerne imidlertid automatisk, idet summen af tallene i første søjle *skal* give 158, summen af tallene i første række *skal* give 200 osv.[13] Hvis vi fx ser på en mulig krydstabel med **100** kvinder, der har et lavt tøjforbrug, følger det derfor at den må have følgende opbygning:

Køn\Forbrug	<1500 kr./måned	≥ 1500 kr./måned	i alt
kvinder	**100**	100	200
mænd	58	102	160
i alt	158	202	360

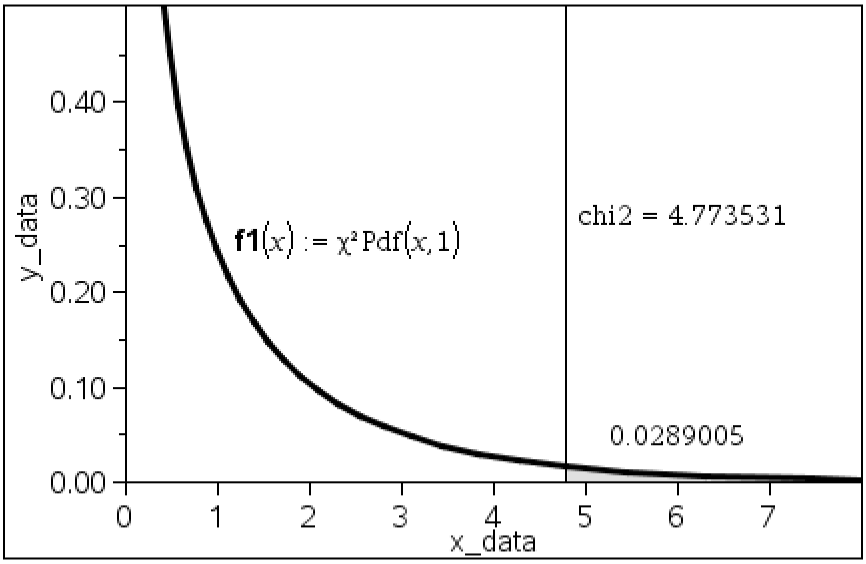
Da vi kun kan vælge en af cellerne frit (indenfor visse bånd) siger vi, at krydstabellen har netop én *frihedsgrad*.

Til hver af de mulige krydstabeller kan vi nu knytte en χ^2-teststørrelse som beskrevet ovenfor. I det ovenstående eksempel fås fx

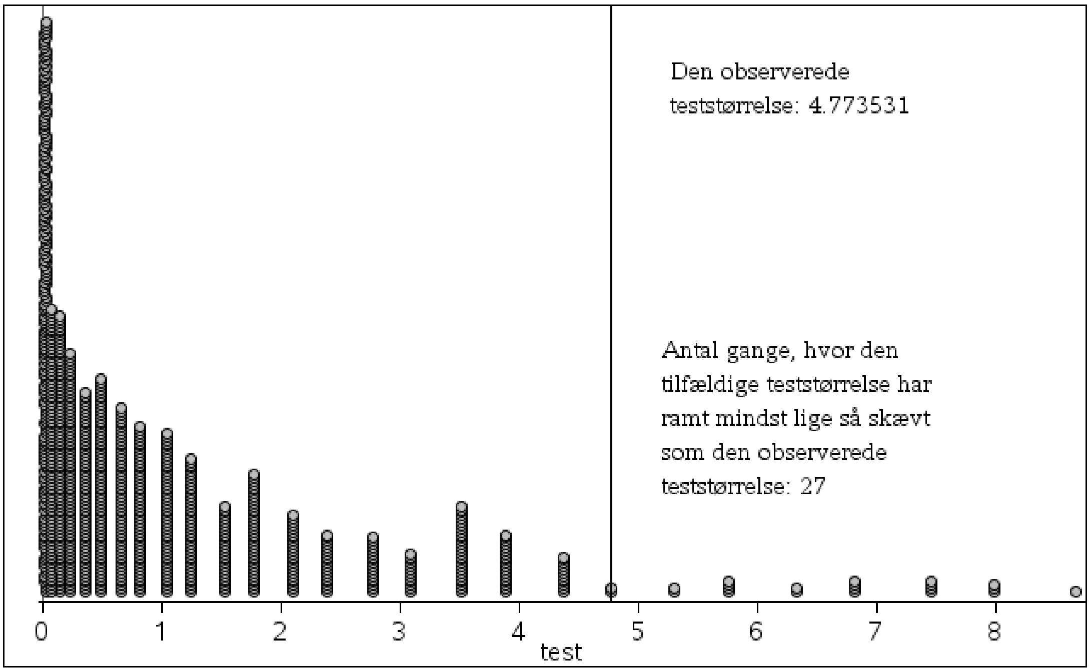
$$\chi^2 = \frac{(100-87{,}78)^2}{87{,}78} + \frac{(100-112{,}22)^2}{112{,}22} + \frac{(58-70{,}22)^2}{70{,}22} + \frac{(102-89{,}78)^2}{89{,}78}$$
$$= 1{,}70 + 1{,}33 + 2{,}13 + 1{,}66 = 6{,}82$$

I χ^2-testet antager vi nu at køn og forbrug er uafhængige, og dermed at krydstabeller, der ligger tæt på de forventede værdier (og dermed har en lille teststørrelse) har høj sandsynlighed, mens krydstabeller, der ligger langt fra de forventede værdier (og dermed har en stor teststørrelse) har lav sandsynlighed.

En præcis beregning af sandsynlighederne viser sig dog at være særdeles vanskelig. Til gengæld er en simpel tilnærmet beregning af sandsynlighederne overraskende nok mulig. I den matematiske statistik viser man nemlig at sandsynligheden for en given krydstabel med stor tilnærmelse er knyttet til χ^2-fordelingen med 1 frihedsgrad. Det skal forstås sådan at sandsynligheden for at finde en teststørrelse, der er mindst lige så stor som den observerede med stor tilnærmelse er givet ved arealet under den graf, der er knyttet til χ^2-fordelingen med 1 frihedsgrad[14]:

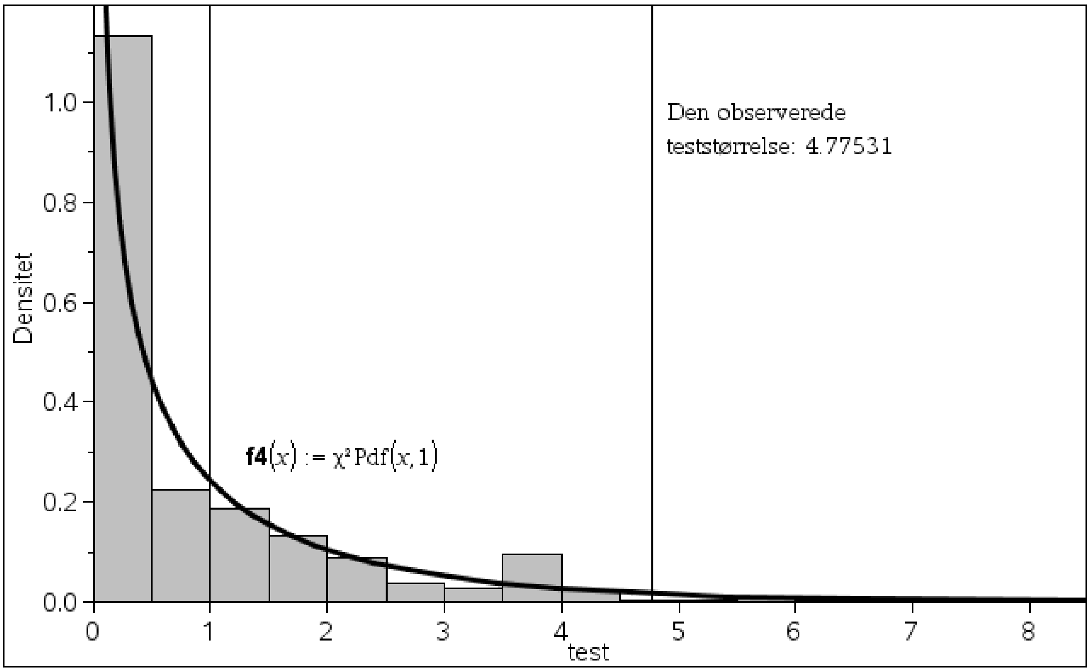


Som omtalt i teksten kunne vi i princippet også finde sandsynligheden ved at gentage spørgeskemaundersøgelsen 'uendeligt mange gange'. I praksis er vi selvfølgelig nødt til at stoppe efter et stort antal gange. Her ses resultatet af 1000 gentagne simuleringer af udregningen af teststørrelsen under antagelsen af H_0, og med fastholdte marginaler (jfr. note 13), hvor hver prik svarer til teststørrelsen for den fundne krydstabel:



I dette tilfælde fandt vi altså 27 krydstabeller med en teststørrelse, der er mindst lige så stor som den observerede. Det svarer til en p-værdi på 27/1000 = 2,7%, hvilket ligger rimeligt tæt på den teoretiske p-værdi 2,9%.

Vi kan også afbilde fordelingen af teststørrelserne i et histogram og derved sammenligne direkte med den teoretiske χ^2-fordeling med 1 frihedsgrad (idet histogrammet først er justeret, så det samlede areal er 1):



I appendiks kan man læse mere om en eksperimentel tilgang til χ^2-testet og den tilhørende fordeling af teststørrelsen.

Opgaver til uafhængighedstest (2×2-tabeller)

Opgave 1. I dagens politik spiller skattenedsættelser en stor rolle. Et stort antal lønarbejdere udspørges via et telefoninterview om deres holdning til en lønnedgang, der kompenseres med en skattenedsættelse:
"Vil du acceptere en lønnedsættelse, hvis den kompenseres med en skattenedsættelse, således at reallønnen (købekraften) er uændret?"
Svarene fordeler sig således på køn:

Køn\Holdning	Ja	Nej
Kvinder	187	105
Mænd	159	56

Fremstil tabellen grafisk i et passende søjlediagram. Kommenter hvad du ser, med brug af begrebet stikprøve.

Færdigudfyld den følgende tabel:

Køn\Holdning	Ja	Nej	I alt
Kvinder	187	105	
Mænd	159	56	
I alt			

Udregn de forventede antal:

Køn\Holdning	Ja	Nej	I alt
Kvinder			
Mænd			
I alt			

Udregn χ^2-teststørrelsen og find p-værdien.

Er forskellen mellem de observerede og forventede svar signifikant på 5% niveau? På 1% niveau?

Opgave 2:
I en Gallupundersøgelse blev 234 kvinder og 257 mænd spurgt om de var tilfredse med deres udseende, i den forstand at de anså sig selv som fysisk tiltrækkende for det modsatte køn. Hertil svarede (i afrundede procenttal) 71% af kvinderne, at de var tilfredse med deres udseende, mens 81% af mændene svarede, at de var tilfredse med deres udseende.

Færdigudfyld det følgende skema

Køn\Selvopfattelse	Tilfreds	Utilfreds	I alt
Kvinder			234
Mænd			257
I alt			

Fremstil tabellen grafisk i et passende søjlediagram. Kommenter hvad du ser, med brug af begrebet stikprøve.

Hvor stor en andel af samtlige adspurgte er tilfredse? Utilfredse?

Udregn χ^2-teststørrelsen og *p*-værdien. Er forskellen mellem de observerede og forventede svar signifikant på 5% niveau? På 1% niveau?

Opgave 3:
Det er en velkendt opfattelse at psykologiske og sociale forhold har indflydelse på overlevelseschancen efter en alvorlig sygdom. En undersøgelse fra 1980 handlede om en eventuel sammenhæng mellem kæledyr og chancen for at overleve i mindst et år efter en hjerteoperation. I undersøgelsen, der omfattede 53 patienter med kæledyr og 39 patienter uden kæledyr, var 78 af patienterne i live efter et år. Af de 78 patienter, der overlevede havde de 50 kæledyr.

Er der belæg i undersøgelsen for en statistisk sammenhæng mellem at overleve i mindst et år og at have kæledyr?

Opgave 4:
I en berømt klagesag fra 1973 blev University of California i Berkeley i USA beskyldt for kønsdiskrimination i sine optagelsesprocedurer. Dette er en meget alvorlig anklage i USA, da en række af offentlige tilskud er afhængige af at Universitet opfylder en række kriterier for 'god opførsel', herunder at Universitet ikke diskriminerer mod køn, race, religion osv. i sine optagelsesprocedurer.
Ud af 2691 mandlige ansøgere blev de 1198 optaget. Ud af 1835 kvindelige ansøgere blev 557 optaget.

Opstil en krydstabel for koblingen mellem køn og optagelse på University of California i Berkeley.

Undersøg om der er belæg i krydstabellen for at anklage universitet for kønsdiskrimination, idet du inddrager begreber som udgangshypotesen H_0, antallet af frihedsgrader, χ^2-teststørrelsen, *p*-værdien og signifikansniveauet.

Opgave 5 (ikke pensum til skriftlig eksamen):
I en undersøgelse af danske mænds og kvinders tv forbrug svarer 60% af kvinderne i stikprøven, at de ser mere end 20 timer tv om ugen, mens det tilsvarende tal for mændene er 53%. Stikprøven indeholder lige mange personer af hver køn, og hypotesen "lige mange mænd og kvinder ser mere end 20 timer tv om ugen" forkastes på 5% signifikansniveau, men accepteres på 1% signifikansniveau. Bestem en mulig størrelse af stikprøven.

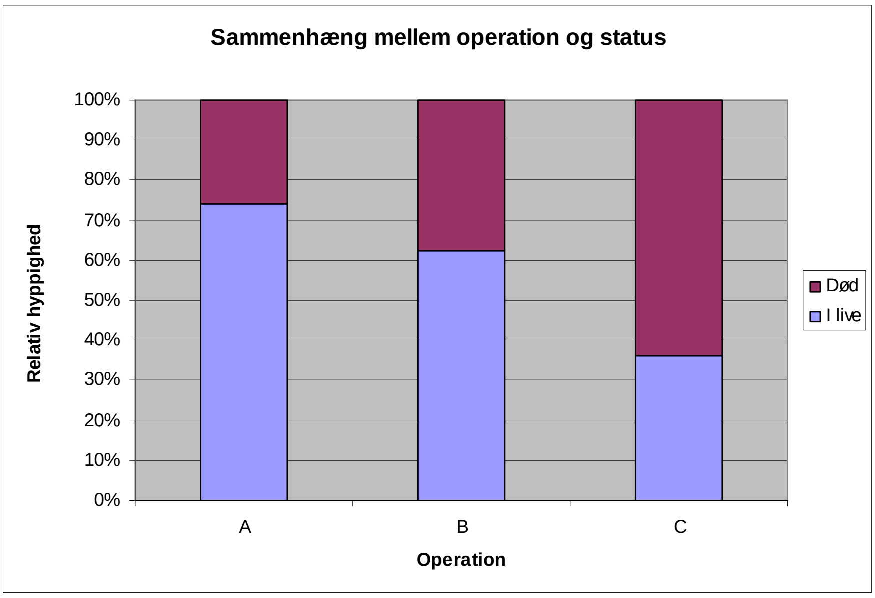
2. Hvad gør vi, hvis vi har flere niveauer på hvert af inddelings-kriterierne?

Ofte er der flere end to kriterier, hvorefter vi inddeler. Betragt følgende eksempel: En hjerneforsker forsker i forskellige måder at operere for en bestemt type 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 der fjernes en større del af det omkringliggende væv.

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):

Operation\Status	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

Afbildes tabellen grafisk som stablede søjlediagrammer fås:



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 forskellige operationer.

Vi lader p_A være sandsynligheden for at en patient er 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	$0{,}6211 \cdot 54 = 33{,}54$	$(1-0{,}6211)\cdot 54 = 20{,}46$	54
Operation B	$0{,}6211 \cdot 16 = 9{,}94$	$(1-0{,}6211)\cdot 16 = 6{,}06$	16
Operation C	$0{,}6211 \cdot 25 = 15{,}53$	$(1-0{,}6211)\cdot 25 = 9{,}47$	25
i alt	59	36	95

Alle de forventede størrelser er mindst 5, så vi kan bruge χ^2-testet igen, nu bare med nogle lidt andre tal.

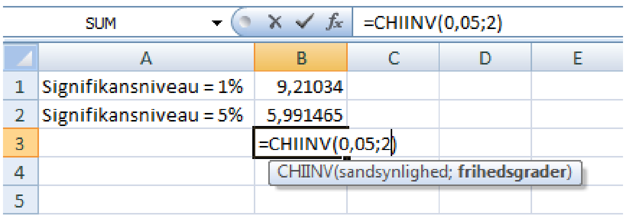
Teststørrelsen udregnes som før ved at udregne:

$$\chi^2 = \sum \frac{(observeret\ antal - forventet\ antal)^2}{forventet\ antal}$$
,

hvor der summeres over alle celler. Dvs. vi får her:

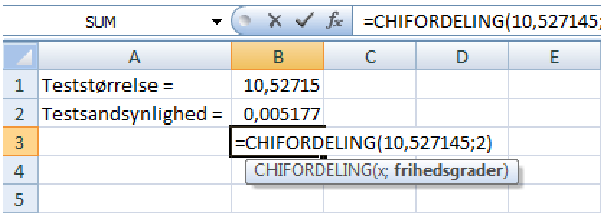
$$\frac{(40-33{,}54)^2}{33{,}54} + \frac{(14-20{,}46)^2}{20{,}46} + \frac{(10-9{,}94)^2}{9{,}94} + \frac{(6-6{,}06)^2}{6{,}06} + \frac{(9-15{,}53)^2}{15{,}53} + \frac{(16-9{,}47)^2}{9{,}47} = 10{,}53$$

Når vi skal vurdere teststørrelsen kan vi som før anvende lommeregner, regneark eller et statistisk værktøjsprogram. I dette tilfælde skal vi benytte en χ^2-fordeling med 2 frihedsgrader (for en nærmere uddybning af antallet af frihedsgrader se evt. det følgende afsnit). Det giver os mulighed for at udregne såvel de kritiske teststørrelser som *p*-værdien:



Ved opslag får vi denne gang, at hvis dette forsøg blev gentaget et meget stort antal gange ("uendeligt mange gange") ville vi i 5% af tilfældene få en teststørrelse, der er større end 5,99 og i 1% af tilfældene få en teststørrelse større end 9,21, alt sammen under antagelse af at udgangshypotesen H_0 om uafhængigheden af operationstypen og patientens status er sand.

Som før kan vi også bestemme en *p*-værdi for teststørrelsen 10,53, dvs. bestemme den præcise sandsynlighed for at få en teststørrelse på 10,53 eller derover; igen selvfølgelig under forudsætning af, at H_0 er sand. I fx Excel finder vi da en p-værdi på 0,0052 = 0,52%:



SUM | =CHIFORDELING(10,527145;

	A	B	C	D	E
1	Teststørrelse =	10,52715			
2	Testsandsynlighed =	0,005177			
3		=CHIFORDELING(10,527145;2)			
4		CHIFORDELING(x; **frihedsgrader**)			
5					

Så i dette tilfælde er vores valgmuligheder:

- ENTEN fastholder vi 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 lige store chancer for overlevelse og siger:

"Der er fundet en sammenhæng mellem overlevelseschance og operationstype, der er statistisk signifikant på 1% niveau."

Ved at udregne de enkelte bidrag til teststørrelsen kan man tydeliggøre, hvilke celler der giver de store bidrag til teststørrelsen, når man må forkaste nulhypotesen. Man ser tydeligt at det først og fremmest er bidragene fra operation C, der er problematiske.

Observation:

Operation\Status	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

Forventet:

Operation\Status	i live	død	i alt
Operation A	33,53684	20,46316	54
Operation B	9,936842	6,063158	16
Operation C	15,52632	9,473684	25
i alt	59	36	95

Bidrag til chi2:

Operation\Status	i live	død
Operation A	1,245568	2,041347
Operation B	0,000401	0,000658
Operation C	2,743265	4,495906

Har vi ikke yderligere viden, må vi derfor foreslå, at operationstype C stilles i bero. 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 faktorerne i betragtning.

Vi kigger lidt nærmere på tallene fra før:

Hyppigheder:

Operation\Status	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

Procenter (overlevelseschance):

Operation\Status	i live	død	i alt
Operation A	74,1%	25,9%	100%
Operation B	62,5%	37,5%	100%
Operation C	36,0%	64,0%	100%
I alt	62,1%	37,9%	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:

Hyppigheder:

Operation\Status	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

Procenter (overlevelseschance):

Operation\Status	i live	død	i alt
Operation A	96,4%	3,6%	100%
Operation B	100,0%	0,0%	100%
Operation C	100,0%	0,0%	100%
I alt	96,8%	3,2%	100%

Over 50 år:

Hyppigheder:

Operation\Status	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

Procenter (overlevelseschance):

Operation\Status	i live	død	i alt
Operation A	50,0%	50,0%	100%
Operation B	57,1%	42,9%	100%
Operation C	33,3%	66,7%	100%
I alt	45,3%	54,7%	100%

Når man tager alderen med i betragtning er operation C pludselig ikke længere så problematisk (ligesom operation A ikke længere er så overbevisende).
Her sker der det, at operationstypen og alderen ikke er uafhængige af hinanden eller af overlevelseschancen. Alderen var i den første undersøgelse en skjult variabel, som viste sig at have stor indflydelse. Når der er flere vigtige faktorer der spiller ind på en gang, så bør de så vidt muligt alle tages med i analysen, der så bliver noget mere kompliceret.

Om teststørrelsen fordeling under H_0

Antallet af frihedsgrader findes ved samme argument som for en 2×2-tabel. Vi har i alt undersøgt 95 patienter. Fordelt på operationstype har 54 patienter fået operation A, 16 patienter operation B og 25 patienter operation C. Fordelt på status (efter et år) er 59 patienter i live, mens 36 er døde. Disse tal fremgår af randværdierne for krydstabellen.

Operation\Status	i live	død	i alt
Operation A			54
Operation B			16
Operation C			25
i alt	59	36	95

Vi interesserer os nu for koblingen mellem operationstypen og status. Det svarer til cellerne inde i krydstabellen. Disse tal kan ikke variere vilkårligt, idet summen af de enkelte rækker, såvel som summen af de enkelte søjler er givet ved randværdierne[15]. Så snart vi fx har fastlagt hvor mange patienter, der er i live efter operation A, følger det automatisk, hvor mange patienter, der er døde efter operation A, idet der i alt er 54 patienter, der har fået operation A osv. Når vi ser på de *mulige krydstabeller* kan vi derfor kun vælge antallet i to af cellerne frit (indenfor visse bånd). I dette tilfælde kan vi fx vælge antallet af patienter der er i live efter at have fået enten operation A eller operation B. Tallet i cellen for 'operation A/i live' skal da ligge mellem 0 og 54 og tilsvarende skal tallet i cellen for 'operation B/i live' ligge mellem 0 og 16 – alt sammen under hensyntagen til at tallene i de øvrige celler ikke være negative, hvilket når vi inddrager operation C kan vises at være det samme som at summen af de to tal skal ligge mellem 34 og 59). Fx kan antallet af patienter, der har fået operation A og er i live vælges til at være 32, mens antallet af patienter, der har fået operation B og er i live kan vælges til at være 9. Resten af krydstabellen må da have følgende opbygning:

Operation\Status	i live	død	i alt
Operation A	**32**	22	54
Operation B	**9**	7	16
Operation C	18	7	25
i alt	59	36	95

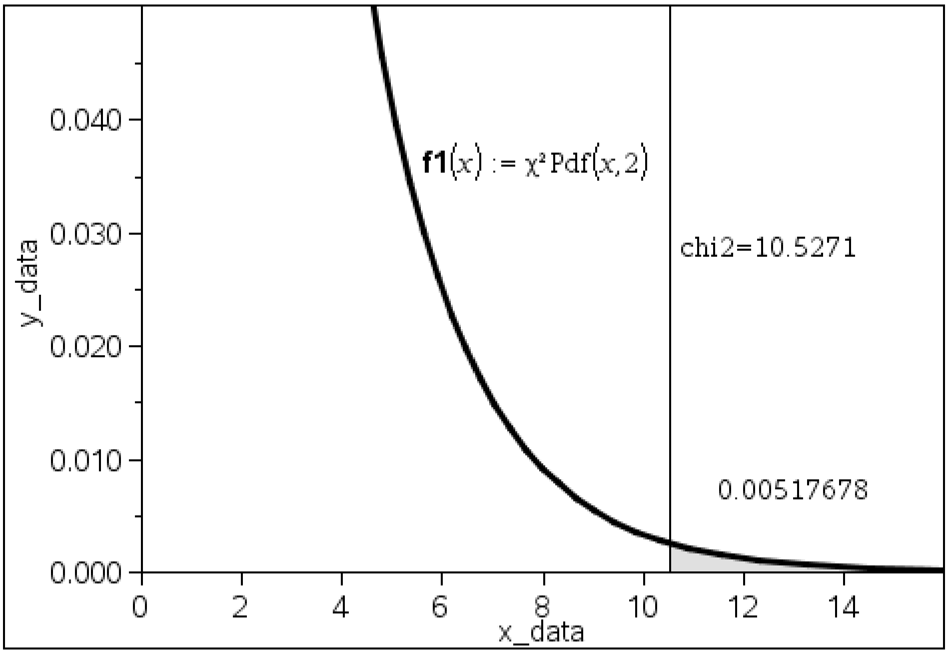
Da vi kun kan vælge to af cellerne frit (indenfor visse bånd) siger vi, at krydstabellen har 2 frihedsgrader[16]. Generelt finder vi på samme måde at en n×m-krydstabel har (n-1)·(m-1) frihedsgrader.

Til hver af de mulige krydstabeller kan vi nu knytte en χ^2-teststørrelse som beskrevet i det foregående. I det ovenstående eksempel fås fx

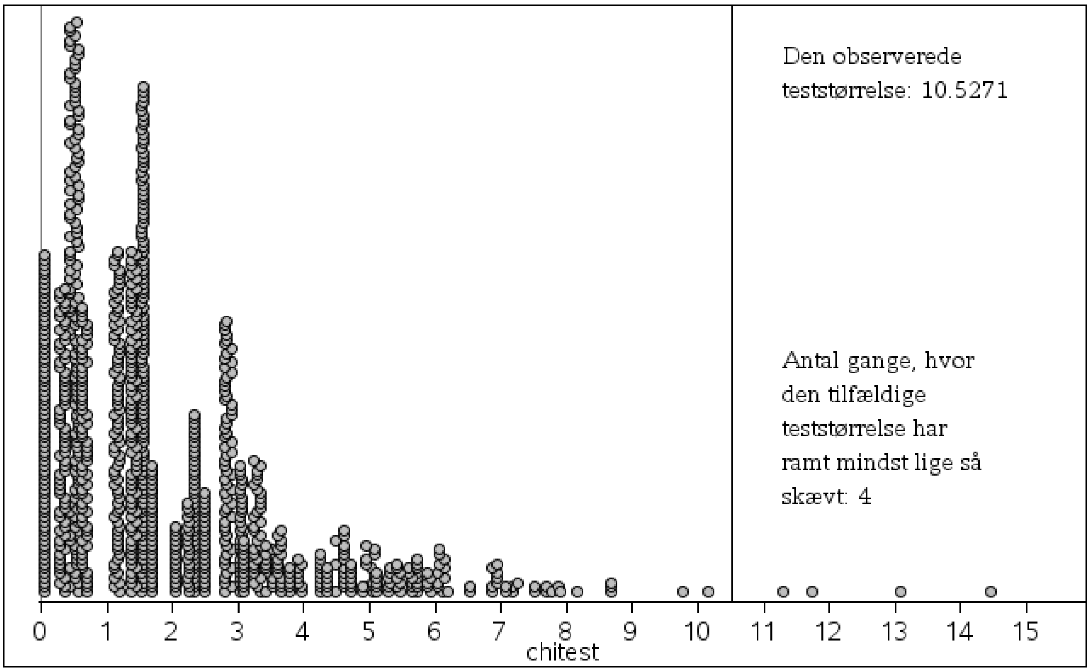
$$\frac{(32-33{,}54)^2}{33{,}54}+\frac{(22-20{,}46)^2}{20{,}46}+\frac{(9-9{,}94)^2}{9{,}94}+\frac{(7-6{,}06)^2}{6{,}06}+\frac{(18-15{,}53)^2}{15{,}53}+\frac{(7-9{,}47)^2}{9{,}47}=1{,}46$$

I χ^2-testet antager vi nu, at operationstypen og overlevelseschancen er uafhængige, og dermed at krydstabeller, der ligger tæt på de forventede værdier (og dermed har en lille teststørrelse) har høj sandsynlighed, mens krydstabeller, der ligger langt fra de forventede værdier (og dermed har en stor teststørrelse) har lav sandsynlighed.

En præcis beregning af sandsynlighederne viser sig at være endog særdeles vanskelig. Til gengæld er en simpel tilnærmet beregning mulig. I den matematiske statistik viser man nemlig at sandsynligheden for en given krydstabel med stor tilnærmelse er knyttet til χ^2-fordelingen med 2 frihedsgrader. Det skal forstås sådan at sandsynligheden for at finde en teststørrelse, der er mindst lige så stor som den observerede med stor tilnærmelse er givet ved arealet under den graf der er knyttet til χ^2-fordelingen med 2 frihedsgrader:[17]

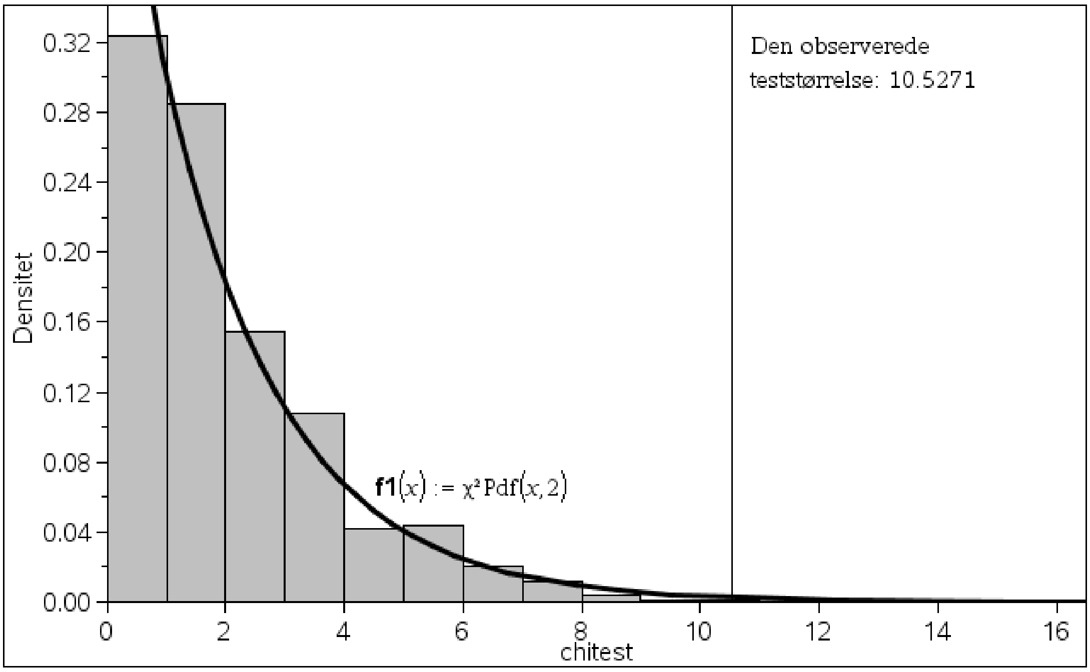


Som omtalt i teksten kunne vi i princippet også finde sandsynligheden ved at simulere en gentagelse af operationerne 'uendeligt mange gange' under antagelse af at operationens udfald er uafhængig af dens type. I praksis er vi selvfølgelig nødt til at stoppe efter et stort antal gange. På næste side ses resultatet af 1000 gentagne simuleringer af udregningen af teststørrelsen under antagelsen af H_0 og med fastholdte marginaler (jfr. note 13), hvor hver prik svarer til teststørrelsen for den fundne krydstabel (I appendiks kan man læse mere om en eksperimentel tilgang til χ^2-testet og den tilhørende fordeling af teststørrelsen):



I dette tilfælde fandt vi altså 4 krydstabeller med en teststørrelse, der er mindst lige så stor som den observerede. Det svarer til en p-værdi på 4/1000 = 0,4%, hvilket ligger rimeligt tæt på den teoretiske p-værdi 0,52%.

Endelig kan vi afbilde fordelingen af teststørrelserne i et histogram og derved sammenligne direkte med den teoretiske χ^2-fordeling med 2 frihedsgrader (idet histogrammet først er justeret, så det samlede areal er 1):



Opgaver til uafhængighedstest (n×m-tabeller)

Opgave 1. I eksemplet med kobling af operationens type og overlevelseschancen så vi at også den skjulte variabel alderen spillede en afgørende rolle. For patienter med en alder på over 50 år (gamle patienter) fandt vi de følgende (fiktive) hyppigheder:

Operation\Status	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

Fremstil tabellen grafisk i et passende søjlediagram. Kommenter hvad du ser, med brug af begrebet stikprøve.

Udregn de forventede antal:

Operation\Status	i live	Død	i alt
Operation A			26
Operation B			14
Operation C			24
i alt	29	35	64

Udregn χ^2-teststørrelsen og find *p*-værdien.

Er forskellen mellem de observerede og forventede resultater signifikant på 5% niveau? På 1% niveau?

Konklusion: Hvad vil du nu anbefale vedrørende de fremtidige operationstyper?

Opgave 2: 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 1: Opstil den relevante nulhypotese og den alternative hypotese for at undersøge, om der er uafhængighed mellem køn og brug af sikkerhedssele?

Spørgsmål 2: Udregn tabellen med de forventede værdier og $\chi 2$ teststørrelsen

Spørgsmål 3: Vil det være rimeligt at bruge en $\chi 2$-fordeling til at vurdere teststørrelsen her?

Spørgsmål 4: Giver stikprøven grundlag for at sige, at der er forskel på de to køns brug af sikkerhedsseler?

Opgave 3: 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.

Opgave 4. I en stor skotsk undersøgelse fra 1981 har man kigget på en eventuel sammenhæng mellem øjenfarve og hårfarve:

Øjenfarve\Hårfarve	Lys	Rød	Middel	Mørk	Sort
Blå	326	38	241	110	3
Lys	688	116	584	188	4
Middel	343	84	909	412	26
Mørk	98	48	403	681	85

Fremstil tabellen grafisk i et passende søjlediagram. Kommenter hvad du ser, med brug af begrebet stikprøve.

Udregn de forventede værdier og teststørrelsen.

Find antallet af frihedsgrader og beregn p-værdien.

Virker det rimeligt at antage at øjenfarven er uafhængig af hårfarven?

Opgave 5: En dyrlæge har fået en mistanke om, at hunde af racen labrador har en større tendens til at udvikle allergi end andre hunderacer. (Bemærk! Fiktivt eksempel)

Gennem det sidste år er der i klinikken blevet registret følgende undersøgelser og resultater af allergi hos hunde:

Race \ allergi	Ingen allergi	Mild allergi	Allergi
Labrador	25	2	10
Schæfer	32	0	7
Puddel	28	3	4
Andet	64	7	8

Spørgsmål 1: Opstil den relevante nulhypotese og den alternative hypotese for at undersøge, om der er uafhængighed mellem hunderace og udvikling af allergi?

Spørgsmål 2: Udregn tabellen med de forventede værdier.

Spørgsmål 3: Vil det være rimeligt at bruge en $\chi 2$-fordeling til at vurdere teststørrelsen her? Hvis ikke, hvordan kan man så komme videre med undersøgelsen?

Spørgsmål 4: Giver stikprøven grundlag for at sige, at der er forskel på risiko for at udvikle allergi for forskellige hunderacer?

Opgave 6: I Titanics tragiske skibsforlis omkom 817 passager, mens 499 overlevede. Blandt de omkomne rejste 122 på første klasse og 528 på tredje klasse. I alt rejste 325 på første klasse og 706 på tredje klasse.

Udfyld det følgende skema:

Klasse\Status	overlevede	døde	i alt
Første klasse			
Anden klasse			
Tredje klasse			
i alt			

Udregn overlevelseschancen i procent for de forskellige klasser.

Er der belæg i tabellen for en statistisk sammenhæng mellem chancen for at overleve og så den klasse man rejste på?

Opgave 7: **Simpsons paradoks**
Dette er en fortsættelse af opgave 4 fra afsnit 1 om påstået kønsdiskrimination i optagelsesproceduren for Berkeley Universitetet i 1973. Vi inkluderer nu en *skjult variabel* i form af de forskelle optagelsesområder (fakulteter). Optagelsestallene fra de seks hovedområder ser nu således:

Hovedområde 1:

Status **Køn**	optaget	afvist
Kvinder	89	19
Mænd	512	313

Hovedområde 2:

Status **Køn**	optaget	afvist
Kvinder	17	8
Mænd	353	207

Hovedområde 3:

Status **Køn**	optaget	afvist
Kvinder	202	391
Mænd	120	205

Hovedområde 4:

Status **Køn**	optaget	afvist
Kvinder	131	244
Mænd	138	279

Hovedområde 5:

Status **Køn**	optaget	afvist
Kvinder	94	299
Mænd	53	138

Hovedområde 6:

Status **Køn**	optaget	Afvist
Kvinder	24	317
Mænd	22	351

Undersøg nu igen sammenhængen mellem køn og optagelsesstatus indenfor hvert af de seks hovedområder.

Er der stadigvæk belæg for påstanden om kønsdiskrimination i optagelsesproceduren?

3. Statistisk test for fordeling af en stikprøve (χ^2–test for goodness of fit)

Når man udtager en stikprøve er man ofte interesseret i at kunne afgøre, hvor godt denne svarer til populationen, set ud fra bestemte inddelingskriterier. Er der en overrepræsentation af bestemte indtægtsgrupper, af bestemte faggrupper eller aldersgrupper ? χ^2–testet for goodness of fit kan hjælpe med til at afgøre sådanne spørgsmål. De inddelingskriterier, som stikprøven testes på er fastlagt af nulhypotesen. Dette illustreres ved følgende eksempel:

Danmarks statistiks opgørelse af indkomstfordelingen for personer over 15 år i Danmark år 2007 viser følgende billede:

I = Indkomst i 1000 kr.	I<50	50≤I I<100	100≤I I<150	150≤I I<200	200≤I I<300	300≤I I<400	400≤I I<500	500≤I
% af befolkning	6,4	9,3	17,8	12,3	24,3	18,0	6,6	5,3

En markedsanalytiker har foretaget en undersøgelse af 1000 personers kendskab til et særdeles kostbart fladskærmsprodukt, men efterfølgende er der opstået tvivl om udvælgelsen af stikprøven, der er forgået som interviewundersøgelse over et par dage i et lokalt supermarked. Det frygtes, at stikprøven har fået for mange respondenter med i de lavere indkomstklasser. Heldigvis er der blevet spurgt om folks indkomst, så man kan lave et test for, om indkomstfordelingen i stikprøven synes at komme fra et specielt segment af befolkningen og altså dermed ikke have den samme fordeling som indkomstfordelingen i Danmark. Hvis det er tilfældet, kan man nemlig ikke generalisere undersøgelsens resultat til hele befolkningen.

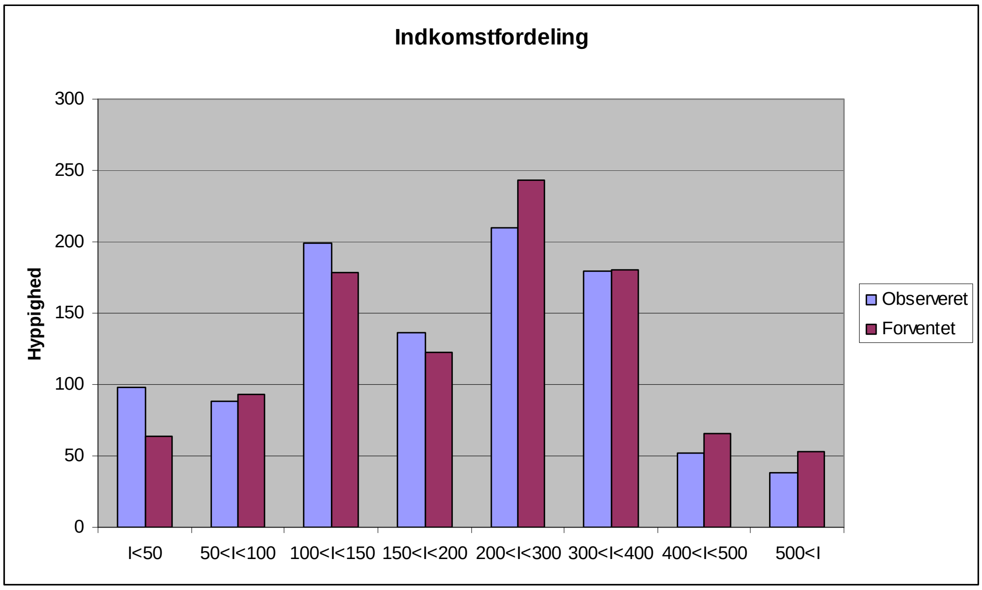
Indkomstfordelingen i stikprøven var:

I = Indkomst i 1000 kr.	I<50	50≤I I<100	100≤I I<150	150≤I I<200	200≤I I<300	300≤I I<400	400≤I I<500	500≤I
Observeret antal	98	88	199	136	210	179	52	38

Den forventede fordeling i stikprøven baseret på de ovenstående procenter er tilsvarende givet ved:

I = Indkomst i 1000 kr.	I<50	50≤I I<100	100≤I I<150	150≤I I<200	200≤I I<300	300≤I I<400	400≤I I<500	500≤I
Forventet antal	64	93	178	123	243	180	66	53

Sammenholder vi de observerede hyppigheder med de forventede følges de så nogenlunde ad. Men man kunne måske være bekymret for, om de laveste indkomster er overrepræsenteret i stikprøven. Her ligger den observerede hyppighed et godt stykke over den forventede.



Modellen er følgende: Sandsynligheden for at en tilfældig udvalgt person over 15 år tilhører en given indkomstklasse, er givet ved den procentdel af befolkningen, der tilhører denne klasse ifølge Danmarks statistiks indkomstfordeling. Sandsynligheden for at en tilfældig udvalgt person har en indkomst på mindre end 50 000 kr. om året er fx 6,4%.

Udgangshypotesen H_0 er, at vores stikprøve har en indkomstfordeling, der er den samme som den danske befolknings. Hvis vores stikprøve skal repræsentere en tilfældig stikprøve på den danske befolkning, ville vi derfor forvente at 1000·0,064 = 64 personer har en indkomst på mindre end 50 000 kr. På den måde kan vi som vist ovenfor beregne, hvor mange vi ville forvente i hver af de 8 indkomstkategorier.

Hypoteserne kan skrives:
H_0 : Indkomstfordelingen i stikprøven er den samme som indkomstfordelingen i populationen.
H_1 : Indkomstfordelingen i stikprøven er forskellig fra indkomstfordelingen i populationen.

Teststørrelsen udregnes på samme vis som før ved at summere over alle indkomstgrupper:

$$\chi^2 = \sum \frac{(observeret\ antal - forventet\ antal)^2}{forventet\ antal}$$

.

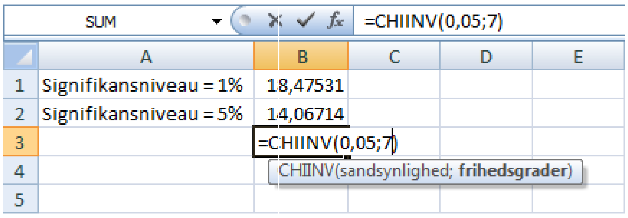
Dvs. vi finder denne gang:

$$\frac{(98-64)^2}{64}+\frac{(88-93)^2}{93}+\frac{(199-178)^2}{178}+\frac{(136-123)^2}{123}+\frac{(210-243)^2}{243}+\frac{(179-180)^2}{180}+\frac{(52-66)^2}{66}+\frac{(38-53)^2}{53} = 33{,}88$$

$$\frac{(98-64)^2}{64}+\frac{(88-93)^2}{93}+\frac{(199-178)^2}{178}+\frac{(136-123)^2}{123}+\frac{(210-243)^2}{243}+\frac{(179-180)^2}{180}+\frac{(52-66)^2}{66}+\frac{(38-53)^2}{53} = 33.88485$$ dvs.

$$\chi^2 = 33.88$$

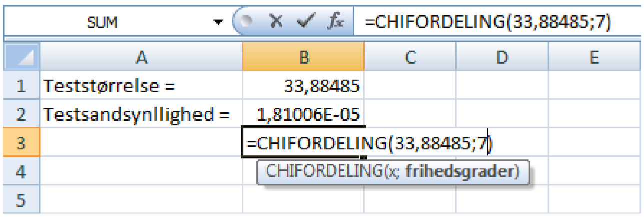
Teststørrelsen er denne gang χ^2-fordelt med 7 frihedsgrader (antallet af frihedsgrader uddybes i det følgende afsnit). Når vi skal vurdere teststørrelsen kan vi som før anvende lommeregner, regneark eller et statistisk værktøjsprogram. I Excel ser det fx således ud (se appendiks for andre værktøjer):



Ved opslag får vi her, at hvis dette forsøg blev gentaget et meget stort antal gange ("uendeligt mange gange") ville vi i 5% af tilfældene få en teststørrelse, der er større end 14,07 og i 1% af tilfældene få en teststørrelse større end 18,48, under den antagelse at udgangshypotesen H_0 er sand. En teststørrelse på 33,88 giver derfor følgende konklusion:

Udgangshypotesen H_0 forkastes på 1% signifikansniveau.

Som før kan vi også bestemme en *p*-værdi for teststørrelsen 33,88, dvs. bestemme den præcise sandsynlighed for at få en teststørrelse på 33,88… eller derover, igen under forudsætning af, at H_0 er sand. I Excel får vi fx

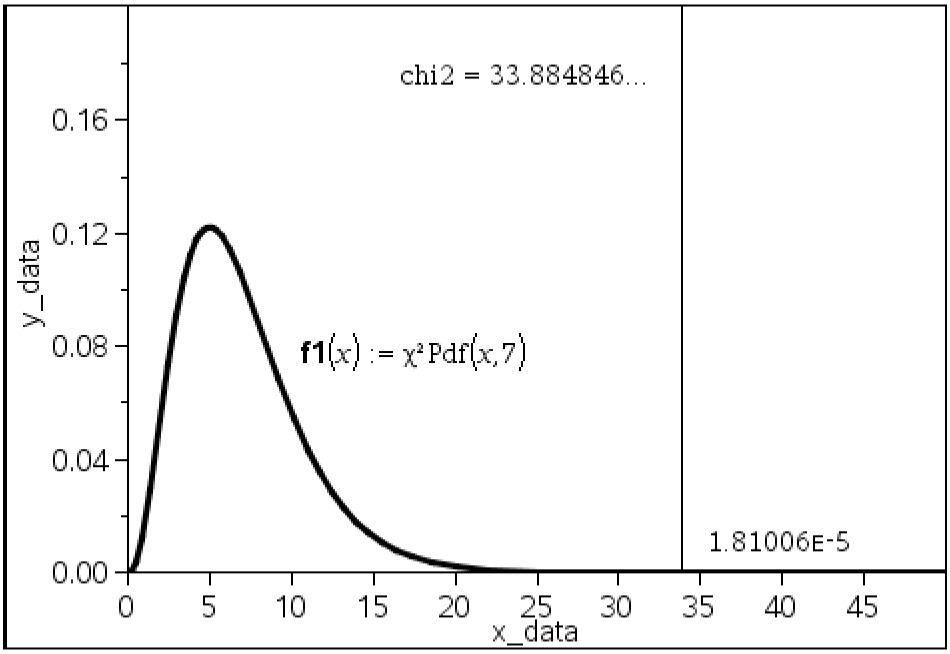


Vi altså en *p*-værdi på 0,0000181, eller praktisk taget 0. Dermed kan vi igen konkludere, at indkomstfordelingen i stikprøven afviger signifikant fra den generelle indkomstfordeling i Danmark, og at markedsanalytikeren skulle have fulgt reglerne for indsamling af repræsentative stikprøver.

Om teststørrelsen fordeling under H_0

Antallet af frihedsgrader findes ved samme argument som tidligere. Vi har i alt udspurgt 1000 kunder i supermarkedet. De kan fordele sig på mange måder på de forskellige indkomstgrupper. Det afgørende er bare at vi tilsammen har 1000 adspurgte (hvilket repræsenterer randværdien i modellen). Vi kan derfor indenfor det overordnede bånd på 1000 adspurgte vælge antallet af udspurgte med indkomst under 50 000 kr. osv. lige til vi kommer til den sidste indkomstgruppe, hvor antallet følger af randværdien.
På samme måde viser man generelt at antallet af frihedsgrader for en kendt fordeling med n inddelingskriterier er givet ved n-1.

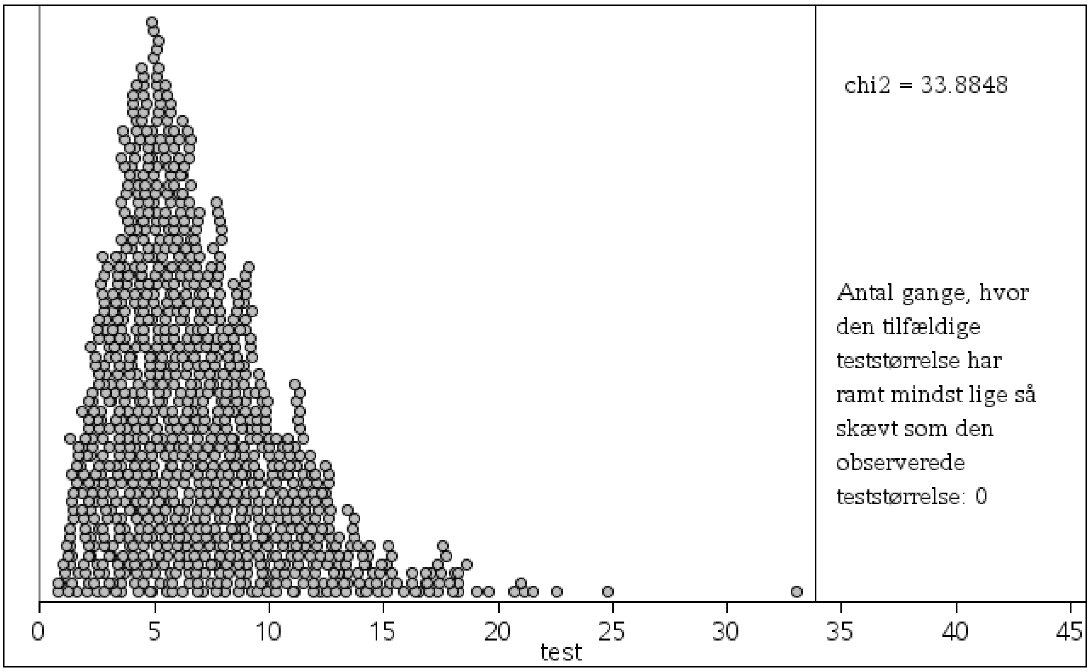
Til hver af de mulige antalstabeller kan vi nu knytte en χ^2-teststørrelse på sædvanlig vis. Denne gang er det faktisk ikke så uoverkommeligt at opstille en formel for sandsynlighederne, men det er nemmere igen at udnytte en tilnærmet beregning af sandsynlighederne knyttet til antalstabellerne. I den matematiske statistik viser man nemlig at sandsynligheden for at finde teststørrelse, der er mindst lige så stor som den observerede med god tilnærmelse er givet ved arealet under den graf, der er knyttet til χ^2-fordelingen med 7 frihedsgrader[18]:

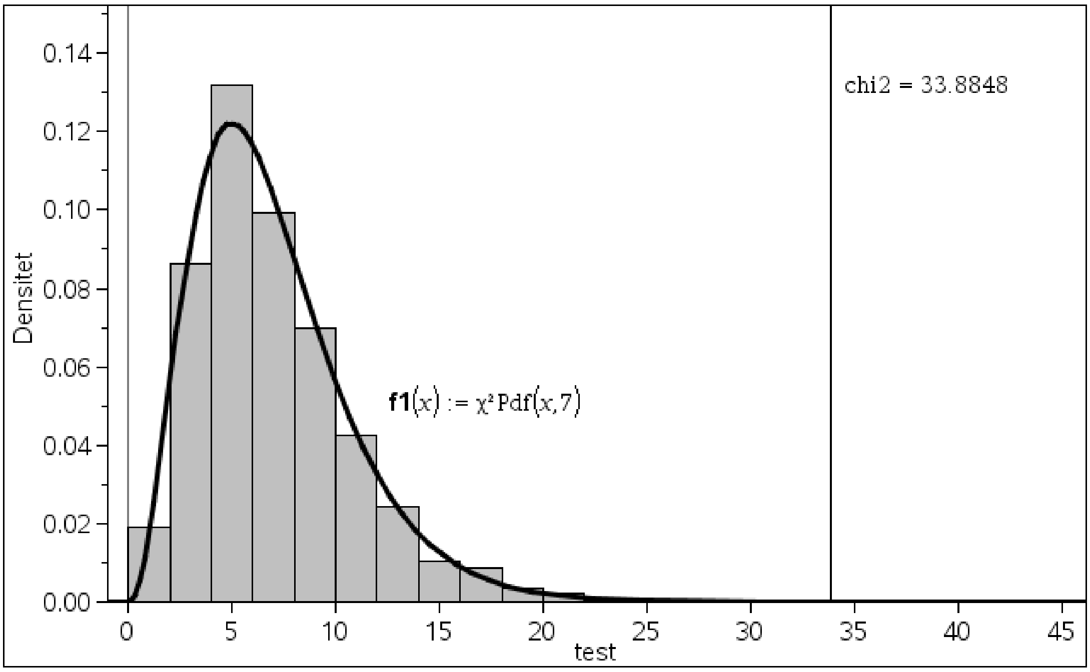


Hvis vi simulerer interviewundersøgelsen et stort antal gange fx 1000 gange under antagelse af H_0, fås den følgende fordeling af teststørrelsen for de 1000 simulerede antalstabeller, hvor hver prik svarer til teststørrelsen for den fundne antalstabel (i appendiks kan man læse mere om en eksperimentel tilgang til χ^2-testet og den tilhørende fordeling af teststørrelsen).

I dette tilfælde fandt vi ingen antalstabeller med en teststørrelse, der er mindst lige så stor som den observerede. Det svarer til en *p*-værdi under 1 ‰.

Endelig kan vi afbilde fordelingen af teststørrelserne i et histogram og derved sammenligne direkte med den teoretiske χ^2-fordeling med 7 frihedsgrader (idet histogrammet først er justeret, så det samlede areal er 1):


chi2 = 33.8848
Antal gange, hvor den tilfældige teststørrelse har ramt mindst lige så skævt som den observerede teststørrelse: 0
0
5
10
15
20
25
30
35
40
45
test


chi2 = 33.8848
f1(x) := χ²Pdf(x,7)
Densitet
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
5
10
15
20
25
30
35
40
45
test

Opgaver til goodness-of-fit test

Opgave 1: Du har en mistanke om, at en af dine venner har en "falsk" terning.

Derfor har du i al hemmelighed noteret udfaldet af alle vedkommendes kast med terningen gennem en hel aftens spil. Dine optegnelser viser, at terningen er endt på "1" i alt 5 gange, "2" i alt 4 gange, "3" i alt 5 gange, "4" i alt 6 gange, "5" i alt 5 gange og "6" i alt 13 gange.

Giver dine observationer anledning til at din mistanke bestyrkes? Du forventes at argumentere ud fra statistiske hypoteser og test, med berøring af begreber som signifikans og/eller p-værdi. Desuden forventes du at kommentere, hvilke konsekvenser du vil lade den statistiske analyse få i den konkrete problemstilling.

Opgave 2: En mindre restaurant med et menukort bestående af 5 forskellige, men faste menuer plejer at have følgende ordrefordeling på disse:
menu 1: 30 %, menu 2: 25 %, menu 3: 20 %, menu 4: 15 % og menu 5: 10 %.

Restauranten foretager sine indkøb for at imødegå en efterspørgsel, der følger dette mønster. Imidlertid er man flere gange i den seneste tid løbet tør for menu 5, og man ønsker at afgøre, om det er en tilfældighed, eller om man skal revidere indkøbsplanerne.

I den seneste uge har man haft 543 gæster. Af disse bestilte 152 menu 1, 101 bestilte menu 2, 110 bestilte menu 3, 91 bestilte menu 4 og 89 bestilte menu 5.
Skal man revidere indkøbsplanerne?

Opgave 3 Når man fremstiller almindelige M&M-slikpiller, blander man de forskellige farver tilfældigt efter det følgende skema:

Almindelige M&M:

Farve	Rød	gul	grøn	orange	brun	blå
Generel andel	20%	20%	10%	10%	30%	10%

For at undersøge om den samme fordeling holder for peanut-M&M-piller blev der udtaget en tilfældig stikprøve på 75 peanut-piller, der fordelte sig således efter farve:

Peanut-M&M:

Farve	Rød	gul	grøn	orange	brun	blå
Observeret antal	11	16	8	5	17	18

Fremstil en passende graf, der sammenligner frekvenserne for almindeligt M&M og peanut-M&M. Kommenter grafen med brug af begrebet stikprøve.

Formuler hypoteserne for et statistisk test hørende til den ovenstående undersøgelse.

(opgaven fortsættes)

Udregn de forventede antal under antagelse af nulhypotesen H_0:

Farve	Rød	gul	Grøn	orange	brun	blå
Forventet antal						

Udregn χ^2-teststørrelsen og find *p*-værdien.

Er forskellen mellem de observerede og forventede antal signifikant på 5%-niveau? På 1%-niveau?

Konklusion?

Opgave 4 I en interviewundersøgelse fra 1978 om danskernes levevilkår udvalgte man tilfældigt 5960 personer. Det lykkedes imidlertid kun at opnå interview med 5166 personer. Der kunne derfor rejses tvivl om hvorvidt den opnåede stikprøve aldersmæssigt var repræsentativ for hele befolkningen.

Aldersfordelingen af stikprøven er givet ved den følgende tabel:

Aldersgruppe	20-29	30-39	40-49	50-59	60-69
Stikprøve-antal	1257	1256	905	914	834

Aldersfordelingen for hele befolkningen fremgår af den følgende tabel:

Aldersgruppe	20-29	30-39	40-49	50-59	60-69
Populations-antal (tusinder)	769	719	556	575	506

Er der belæg for bekymringen over at stikprøven måske ikke er repræsentativ for aldersfordelingen i den danske befolkning?

Større opgaver, der kan indgå i projekter sammen med biologi.

Opgave 1. Mendels egne data
De klassiske arvelighedslove stammer fra Mendel (fra før man vidste noget om gener). I et forsøg med krydsning af ærter hæftede Mendel sig ved om de var glatte/rynkede henholdsvis gule/grønne. Her er glat og gul *dominerende egenskaber*, dvs. ifølge Mendels arvelighedslove er der 3 gange så stor sandsynlighed for at afkommet bliver glat, som at det bliver rynket osv. Kombineres en dominerede egenskab med den tilsvarende vigende egenskab fås nemlig fire mulige kombinationer. De tre af dem fremstår nu som dominerende.

A = glat a = rynket	A	a
A	A A	A a
a	a A	a a

B = gul b = grøn	B	b
B	B B	B b
b	b B	b b

I et af Mendels forsøg med 556 ærteplanter fandt Mendel nu den følgende fordeling:

Type	glat gul	rynket gul	glat grøn	rynket grøn
Antal	315	101	108	32

Argumentér for at fordelingen af de fire karakteristika, hvor vi kombinerer glathed med farve, ifølge Mendels arvelighedslove bør være 9:3:3:1.

Undersøg nu om Mendels egne data stemmer overens med den foreslåede fordeling, idet du inddrager begreber som nulhypotese, forventede værdier, frihedsgrader, teststørrelse, testsandsynlighed og signifikansniveau.

Der har været fremsat beskyldninger mod Mendel for at have lidt for pæne data: Er der belæg for denne påstand i denne undersøgelse?

Opgave 2: **Mutantmus**
Vi ser nu på data fra et forsøg med mutantmus, der involverer to gener: tau og map1b. Data er hentet fra artiklen: Defects in Axonal Elongation and Neuronal Migration in Mice with Disrupted *tau* and *map1b* Genes.

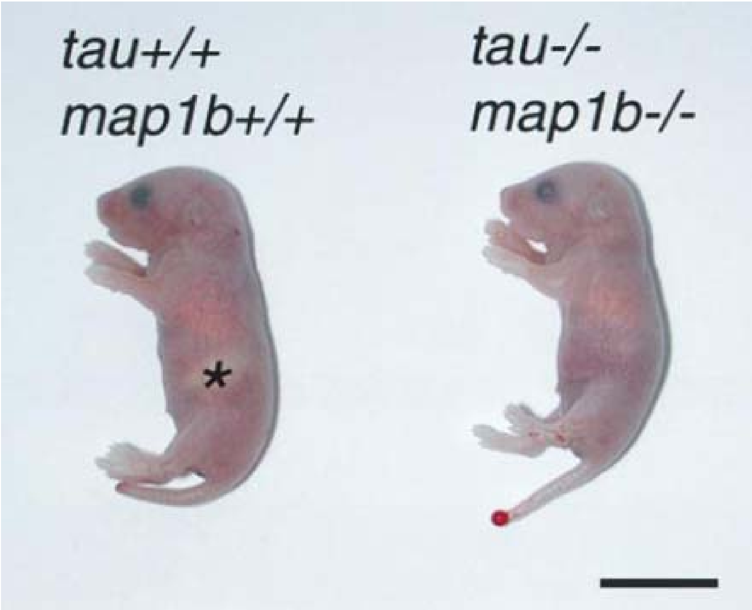


Figure 1. Gross appearance of *tau*1/1*map1b*1/1 and *tau*2/2
*map1b*2/2 mice (P0.5). Note the lack of milk in the stomach of the *tau*2/2*map1b*2/2 animal. The area indicating stomach filled with milk in the *tau*1/1*map1b*1/1 mouse is marked with an asterisk.
Bar, 10 mm.

Her ses tabellen over de udspaltede data. Hvert af generne forekommer i to varianter + og -. De kan derfor kombineres på 16 forskellige måder, men rækkefølgen er ligegyldig, så mange af kombinationerne er ækvivalente:

Table I. Genotype Distribution of Offspring after Mutual tau+/−map1b+/− Mating

	Number of progeny			
Age	E16.5	P0.5	4 wk*	Mendelian distribution
tau+/+map1b+/+	11 (8.7%)	24 (5.8%)	54 (9.7%)	6.25%
tau+/−map1b+/+	17 (13.4%)	47 (11.3%)	80 (14.4%)	12.5%
tau−/−map1b+/+	9 (7.1%)	29 (7.0%)	40 (7.2%)	6.25%
tau+/+map1b+/−	17 (13.4%)	54 (12.9%)	86 (15.5%)	12.5%
tau+/−map1b+/−	34 (26.8%)	102 (24.5%)	162 (29.2%)	25%
tau−/−map1b+/−	11 (8.7%)	66 (15.8%)	64 (11.6%)	12.5%
tau+/+map1b−/−	8 (6.3%)	24 (5.8%)	20 (3.6%)	6.25%
tau+/−map1b−/−	10 (7.9%)	44 (10.6%)	42 (7.6%)	12.5%
tau−/−map1b−/−	10 (7.9%)	27 (6.5%)	6 (1.1%)	6.25%
Total	127	417	554	

Numbers of pups of each genotype are shown.

Første søjle er hyppigheder for de yngste fostre (E16.5), anden søjle er hyppigheder for de lidt ældre fostre og tredje søjle er data for 4 uger gamle museunger (4wk). Til sidst kommer den forventede udspaltning i følge Mendels love.

Start med at overveje hvordan Mendels arvelighedslove kan bruges til at forklare den forventede fordeling, som den fremgår af skemaet.

Udregne nu de forventede antal for de yngste musefostre hørende til de ni musetyper, idet du opstiller en antalstabel som den følgende:

De yngste fostre

Musetype	Observeret	Forventet
tau + + map + +	11	
tau + – map + +	17	
tau – – map + +	9	
tau + + map + –	17	
tau + – map + –	34	
tau – – map + –	11	
tau + + map – –	8	
tau + – map – –	10	
tau – – map – –	10	
i alt	127	

Udregn dernæst χ^2-teststørrelsen, antallet af frihedsgrader og testsandsynligheden *p*. Virker det som om den observerede fordeling følger Mendels love?

Prøv derefter at regne på de fire uger gamle musefostre. Virker det som om den nu observerede fordeling følger Mendels love?

Litteraturliste

Engelsk:
1: Ann E. Watkins, Richard L. Scheaffer and George W. Cobb. Statistics in Action – Understanding a World of Data. Key Curriculum Press.
[Moderne amerikansk lærebog på gymnasieniveau i statistik, der indeholder et fint kapitel om χ^2-test, der også lægger vægt på en eksperimentel tilgang]
2: J. Burt & G. Barber. Elementary Statistics for Geographers. The Guildford press.
3: P. Newbold. Statistics for Business and Economics. Prentice Hall International Editions.
[nr. 2 og 3 anbefales af Susanne Christensen som generel baggrundslitteratur om statistiks anvendelse i andre fag]

Dansk:
4: P. Mortensen. Repræsentative undersøgelser. Systime.
[Her uddybes de krav man bør stille til repræsentativitet af en undersøgelse]
5: Aksel Bertelsen. Statistik med Matematik, Systime.
[Indeholder et instruktivt kapitel om antalstabeller og χ^2-test, der også lægger vægt på en eksperimentel tilgang. Kapitlet inddrager også odds-ratio som teststørrelse. Bogen indeholder også en teoretisk redegørelse for sammenhængen mellem normalfordelingen og χ^2-fordelingen]
6: Helge Gram Christensen, Sandsynlighedsregning og statistik for gymnasiet, GAD
[Bogen er fra 1981 og findes i bogsamlingen på en del gymnasier. Indeholder ca. 20 sider om χ^2-test, med mange instruktive eksempler]
Flere af de gængse lærebogssystemer har afsnit om χ^2-fordelingen.

Værktøjsprogrammer:
7: Per Vejrup Hansen. Statistik med Excel. Forlaget Samfundslitteratur.
[Udførlig gennemgang af såvel den deskriptive statistik som hypotesetest med Excel]
8: Bjørn Felsager. Statistik med DataMeter. (Web-bog. Kan hentes på EMU – http://www.emu.dk/gym/fag/ma/undervisningsforloeb/statistik/index.html)
[Udførlig gennemgang af såvel den deskriptive statistik som hypotesetest med DataMeter]
9: Bjørn Felsager. Statisk med TI-Nspire. (Web-bog. Kan hentes på EMU – http://www.emu.dk/gym/fag/ma/undervisningsforloeb/statistik/index.html)
[Udførlig gennemgang af såvel den deskriptive statistik som hypotesetest med TI-Nspire]
10: Jan B Sørensen. Statistik med TI-Interactive. (Noten kan hentes på EMU – http://www.emu.dk/gym/fag/ma/undervisningsforloeb/statistik/index.html)

Noter

[1] *Population*: Samlingen af alle individer, der er i fokus i en undersøgelse, fx samtlige gymnasieelever i landet. *Stikprøve*: Et udsnit (delmængde) af populationen.
[2] Det grundlæggende princip er, at stikprøvens elementer eller individer udtages *simpelt tilfældigt*, så alle individer i populationen har samme sandsynlighed for at blive udtaget. Det kan fx ske på basis af personnumre eller lister over tilfældige tal. Har man en formodning om at svarene afhænger af køn, alder, indkomst eller andet, kan man spørge til sådanne kategorier og senere teste om stikprøven var *repræsentativ* fx i forhold til indkomst, jfr. afsnit 3. En stikprøve vil i sagens natur aldrig være repræsentativ ift. alle sider af en population. Der vil altid være stikprøvefejl.
[3] Hvis stikprøven er udtaget, så alle individer i populationen har samme sandsynlighed for at blive valgt, så er *sandsynligheden* for at vælge fx en kvinde med lavt forbrug lig med den andel af populationen, kvinder med lavt forbrug udgør.
[4] En anden statistisk undersøgelse kunne foretages med udgangspunkt i svarene om størrelsen af de brugte beløb og fx undersøge, om det gennemsnitlige forbrug for kvinder er større end det gennemsnitlige forbrug for mænd. I så fald kunne man bruge den statistiske metode, der hedder *sammenligning af to middelværdier*. Det kan man fx læse mere om i den anbefalede litteratur 1-3.
[5] p_k kaldes også *den empiriske sandsynlighed*.
[6] At to svar eller to hændelser er *uafhængige* af hinanden har en præcis matematisk definition, som kan findes i alle bøger om sandsynlighedsregning. I de fleste tilfælde svarer den matematiske definition af uafhængighed til dagligsprogets opfattelse, som er den der er givet i teksten.
[7] I den ideelle verden formuleres først hypotesen, dernæst udvælges stikprøven og man sikrer sig den er repræsentativ, og at den enkeltes svar ikke påvirkes af andres svar. I praksis kan man være nødt til at gå på kompromis med dette.
[8] *Signifikans* er et begreb i statistisk testteori. Man starter med at fastlægge et *signifikansniveau*, der udtrykker grænsen for at hypotesen accepteres eller forkastes. Dernæst beregnes sandsynligheden for at det fundne resultat ville indtræffe *under forudsætning af vores nulhypotese*. Hvis denne sandsynlighed kommer under signifikansniveauet anses hypotesen for utroværdig. Grænsen beror en konvention: I praksis benyttes ofte 5%. Men det afhænger af problemstilling og ikke mindst af konsekvenserne af at tage fejl. Er disse konsekvenser alvorlige – fx ved test af lægemidler – anvendes signifikansniveauer på 1% eller måske 0,01%. Man konkluderer med udsagn af typen: Der er påvist en sammenhæng mellem …, der er signifikant på 5% niveau. Eller: *Der er signifikant forskel mellem … på 5% signifikansniveau.*
[9] χ er det græske bogstav chi. χ^2 udtales 'ki-i-anden'. Σ er det græske bogstav sigma. Det anvendes i matematik som symbol for summation. Det læses: 'summen over alle størrelserne … '
[10] Et af kravene for at kunne anvende et χ^2 test er, at *alle de forventede værdier er mindst 5*. Men grænsen for hvornår man siges at bruge mange penge på tøj var jo subjektivt fastsat. Havde vi sat denne grænse så højt elle lavt, at der i en af cellerne med de forventede værdier var kommet et tal mindre end 5, så kunne vi ikke umiddelbart bruge χ^2 testet. En løsning kan af og til være at lave grænserne om, så tallene grupperes anderledes.
[11] se note 8.
[12] Dette kaldes *fejl af første art*.
[13] I virkeligheden er det eneste tal, der ligger helt fast 360. Også marginaltallene (200kvinder, 160 mænd osv.) kunne varieres. Man kan imidlertid vise i den teoretiske statistik, at det i alle praktiske sammenhænge giver korrekte konklusioner at ræsonnere ud fra faste marginaler.
[14] Det er grafen for *tæthedsfunktionen* for χ^2 fordelingen med 1 frihedsgrad. χ^2 PDF er maskinsprog for denne tæthedsfunktion.
[15] Samme kommentar vedr. marginaler til dette afsnit som til det tilsvarende med 2×2 tabeller.
[16] *Frihedsgrader* kendes også fra andre emner. Fx er vinkelsummen i en trekant 180°. Vi kan derfor (indenfor visse grænser) vælge den første vinkel frit, ligeså den anden, men derefter er den tredje vinkel fastlagt. Vinklerne i en trekant har derfor to frihedsgrader.
[17] Det er grafen for tæthedsfunktionen. χ^2PDF er maskinsprog og angiver tæthedsfunktionen.
[18] Det er grafen for tæthedsfunktionen. $\chi^2\text{PDF}(x,7)$ er maskinsprog og angiver tæthedsfunktionen for χ^2 fordelingen med 7 frihedsgrader.