MATEMATIK

NOTAT 20

CHI^2-TEST

AF:

CAND. POLYT.

MICHEL MANDIX

SIDSTE REVISION: APRIL 2024

Oversigt over græske bogstaver:

Kapitaler	Minuskler	Navn
Α	α	Alfa
Γ	γ	Gamma
Ε	ε	Epsilon
Η	η	Eta
Ι	ι	Jota
Λ	λ	Lambda
Ν	ν	Ny
Ο	o	Omikron
Ρ	ρ	Rho
Τ	τ	Tau
Φ	φ	Phi
Ψ	ψ	Psi

Kapitaler	Minuskler	Navn
Β	β	Beta
Δ	δ	Delta
Ζ	ζ	Zeta
Θ	θ	Theta
Κ	κ	Kappa
Μ	μ	My
Ξ	ξ	Xi
Π	π	Pi
Σ	σ	Sigma
Υ	υ	Ypsilon
Χ	χ	Chi
Ω	ω	Omega

Indholdsfortegnelse:

Chi²-test

Introduktion:

Overordnet ide

Chi²-fordelingen er et underemne i den matematiske disciplin: "Statistik".
Som altid, benyttes statistik til at give en kvalificeret beskrivelse af en population. Da det for det meste ikke er muligt at undersøge hele populationen, nøjes man med at undersøge en stikprøve – dvs. – et udsnit af hele populationen.
Når der udtages en stikprøve, og der efterfølgende testes på den stikprøve, er det for at undersøge om variationerne i stikprøven skyldes tilfældige udsving, eller om der er en signifikant (betydningsfuld) sammenhæng.
Man kan sige det på en anden måde, hvor man antager, at man vil lave fuldstændig det samme eksperiment en gang til. Det, der undersøges er sandsynligheden for, at eksperiment nummer to vil falde ud på samme måde som det første eksperiment, eller om det er for usandsynligt.
Hvis det ikke er sandsynligt, at eksperiment nummer to er rimeligt identisk med eksperiment nummer et, så er der ingen målbar sammenhæng mellem de observerede data, men hvis de to eksperimenter falder nogenlunde ens ud, så er der belæg for at sige, at der er en sammenhæng mellem data i stikprøven.

Den grundlæggende idé er: Man har lavet en stikprøveundersøgelse hvor deltagerne har besvaret to spørgsmål. Man vil undersøge om der er nogen sammenhæng mellem deltagernes svar på de to spørgsmål, eller om de er uafhængige.

I den offentlige debat høres ofte en række påstande om, hvorledes ting hænger sammen. Ofte følger der ikke datamateriale eller tests med, når påstande præsenteres, og derfor er der ikke særlig gode muligheder for at undersøge og dermed be- eller afkræfte sammenhænge. Det kan derfor være svært at gennemskue, om andres undersøgelser eller argumenter er til at stole på.

Chi²-test bruges til at teste, om der er ***uafhængighed*** mellem to variable, derfor også kaldet en "Test for uafhængighed".
Resultatet af en test for uafhængighed er en sandsynlighed. Testsandsynligheden angiver muligheden for at opleve endnu større forskelle, selv om variablene er uafhængige.

Hvad er en chi²-fordeling?:

Chi²-fordelingen er lidt beslægtet med normalfordelingen, men der er dog visse forskelle:

Som vist i et andet notat, er normalfordelingen afhængig af to variable, nemlig μ (my, middelværdien) og σ (sigma, spredningen).
Dette skrives som: $X \sim N(\mu;\sigma)$.

Chi²-fordelingen er derimod kun afhængig af én variabel, og det er antallet af frihedsgrader, benævnt som *df* (Degree of Freedom).
Dette skrives som: $X \sim \chi^2(df)$.

En chi²-fordeling kan have et vilkårligt antal frihedsgrader, og antallet er afhængigt af datatabellens størrelse. (Se afsnittet om frihedsgrader).

Begreber, forkortelser og symboler:

Hypotese
En påstand eller et udsagn, som antyder en sammenhæng mellem **to** variable.
Specielt benævnes nulhypotesen med H_0 og den alternative hypotese med H_1.

Sandsynlighed
Chancen (eller risikoen) for, at en hændelse indtræffer. Hvis et eksperiment udføres tilpas mange gange er sandsynligheden den hyppighed, hvormed en bestemt hændelse vil indtræffe.

Signifikansniveau
Signifikansniveauet er den nøjagtighed, hvormed testen udføres. Eller med andre ord: I hvor mange % af forsøgene, er det ok at tage fejl? Typisk er denne værdi lig med 5 %. I det virkelige liv, tages der hensyn til, hvad det koster at lave en undersøgelse. Jo større præcision der ønskes, desto mere vil det koste at lave undersøgelsen – typisk fordi der skal spørges mange flere personer, hvis præcisionen skal øges.

Pivottabel (eller antalstabel)
Data ankommer oftest som et stort tilfældigt rod af kombinationen af svar på to eller flere forskellige spørgsmål. Typisk vil svarene være ordnet sådan at en linje repræsenterer svarene fra en person (eller hændelse). Desuden vil linjerne stå i den rækkefølge, som de er indkommet – dvs. den først adspurgte står øverst og den senest adspurgte står nederst.
I Excel (eller et andet program), kan svarene optælles og fordeles. I en chi²-test vurderes der altid kun på to variable, så hvis der er stillet f.eks. 5 spørgsmål, kan man i pivottabellen udvælge de variable, som man ønsker at analysere sammenhængen for.

Forventede værdier
Når der er opstillet en pivottabel med optællingen af værdierne, skal de sammenlignes med de forventede værdier. Det er ikke altid nemt at gennemskue på forhånd, om der er en sammenhæng mellem de to variable eller ej. Det er jo ikke sikkert, at fordelingen af de adspurgte er jævnt fordelt. Måden at kompensere for dette er at finde de forventede værdier. De forventede værdier udregnes for hver celle i pivottabellen. Hvis de indsamlede værdier i pivottabellen er relativt tæt på de forventede værdier, forventes det, at nulhypotesen accepteres, men lad være med at gætte på udfaldet. Det er – som tidligere nævnt – svært at gennemskue uden at lave beregningerne.

Teststørrelsen, χ^2
Begrebet teststørrelse behandles af to omgange. Først diskuteres begrebet "**bidrag** til teststørrelse". Bidraget til teststørrelsen udregnes – ligesom de forventede værdier – for hver enkelt celle i pivottabellen.
Som det ses i et senere afsnit, vil den primære udregning for de enkelte bidrag være: "Observeret værdi – Forventet værdi)², og heraf kan det udledes, at hvis de forventede værdier ligger meget tæt på de observerede værdier, så vil dette bidrag blive forholdsvist lille.

Selve teststørrelsen, χ^2, udregnes som summen af alle bidragene.

Chi²-test

Kritisk værdi

Den kritiske værdi i en chi²-fordeling afhænger af to parametre: 1) Signifikansniveauet og 2) Antallet af frihedsgrader, som er givet ud fra pivottabellen.
Den kritiske værdi er et sammenligningsgrundlag for teststørrelsen, χ^2.

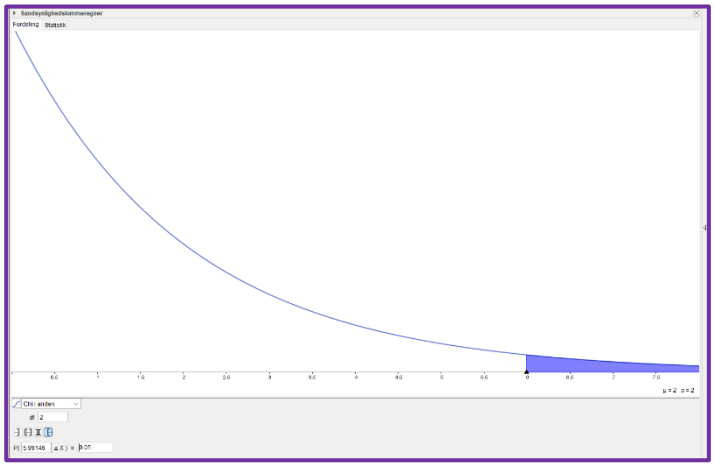
Testsandsynlighed, *p*

Testsandsynligheden er et tal for, hvor "skæv" undersøgelsen er – eller med andre ord: "Hvor langt fra den forventede virkelighed er de foreliggende data?"
Hvis der accepteres, at der tages fejl i 5 % af forsøgene, betyder det, at man kan tage fejl i et ud af tyve forsøg.
Hvis den beregnede testsandsynlighed er på f.eks. 1 %, betyder det, at sandsynligheden for at få et resultat som det der er givet, er 1 %. Altså at det vil forekomme i 1 ud af 100 tilfælde. Dette betyder så, at sandsynligheden for netop dette forsøg vil forekomme i endnu færre tilfælde, end de "tilladte" 5 % (som er givet af signifikansniveauet, og som tidligere nævnt kan ændre sig, hvis man accepterer en anden fejl- eller succesrate).

Ser man på den grafiske afbildning af en chi²-fordeling, er der tale om en kurve. Udseendet af denne kurve varierer, afhængigt af antallet af frihedsgrader *df*.

Ser man på det areal, som ligger under kurven, kan man sige, at dette areal samlet er lig med 1.

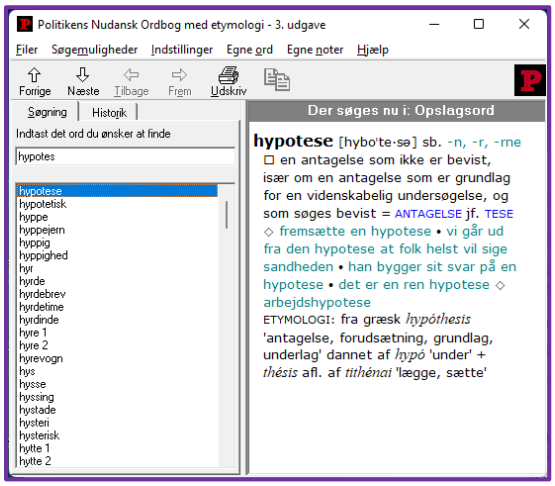
Når man bestemmer signifikansniveauet til 5 %, betyder det således, at i 95 % (19 ud af 20) tilfælde, så er undersøgelsen ok, men i de sidste 5 %, så ligger data for langt væk fra de forventede værdier. Dette kan vises på grafen, som det blå område.



Den udregnede testsandsynlighed kan anskues på samme måde. Hvis den udregnede sandsynlighed er mindre end f.eks. 5 %, så betyder det at det markerede areal er endnu mindre og den nederste grænse for arealet skal derfor rykkes til højre. Deraf følger at data er endnu længere væk fra signifikansniveauet end de tilladte øverste 5 %, og dermed kan nulhypotesen ikke accepteres.

Hvis f.eks. den udregnede testsandsynlighed er meget tæt på 0, betyder det altså at data er så langt væk fra de forventede værdier, at der er 0 % sandsynlighed for at få et datasæt magen til dette, hvis eksperimentet gentages.

Hypoteser:



Som det ses af skærmbilledet af opslaget i Politikens Nudansk Ordbog, kan en hypotese løseligt oversættes med: ”forudsætning” eller ”antagelse”.
Man kan også anvende ordet: ”påstand”

I forbindelse med chi²-testen er den bedste oversættelse: ”antagelse”, idet nogle af udregningerne kun kan gennemføres under den antagelse, at der ikke er nogen sammenhæng mellem de to spørgsmål.

Man antager indledningsvist, at der IKKE er nogen sammenhæng mellem de to spørgsmål (variable), der måles på i testen.

Notatets eksempel omhandler spørgsmålet om, hvorvidt mænd eller kvinder foretrækker at gå i et bestemt fitnesscenter. De to spørgsmål bliver således:

1) ”Er du mand eller kvinde?”
2) ”Foretrækker du at træne i København, Køge eller Solrød?”

Til brug i en chi²-test opstiller man altid den negative (ikke-bekræftende) hypotese først og kalder den for Nulhypotesen (H_0). Så i dette tilfælde bliver Nulhypotesen: ”Der er INGEN sammenhæng mellem de adspurgtes køn og deres foretrukne træningscenter”.

Hvis der ikke er sammenhæng, må det modsatte være gældende. I så fald er det givet, at der ER sammenhæng mellem de adspurgtes køn og deres foretrukne træningscenter.
Dette alternative scenarie kaldes for: ”Den alternative hypotese (H_1)”.

Det er vigtigt at opgaven begyndes med at definere de to hypoteser, men da de altid er det ”samme” – dvs. at nulhypotesen altid er den benægtende påstand, hvor der ikke er nogen sammenhæng, og den alternative hypotese, som er ”alt det andet” – dvs. at der er en eller anden sammenhæng mellem de to variable – burde det ikke give noget større problem.

Signifikansniveau:

Hele ideen med denne type opgave er at bestemme hvorvidt man kan acceptere nulhypotesen, eller om man i stedet skal vælge den alternative hypotese. Et spørgsmål der rejser sig i den forbindelse er: ”Hvad er succeskriteriet? Hvornår skal man acceptere nulhypotesen?”

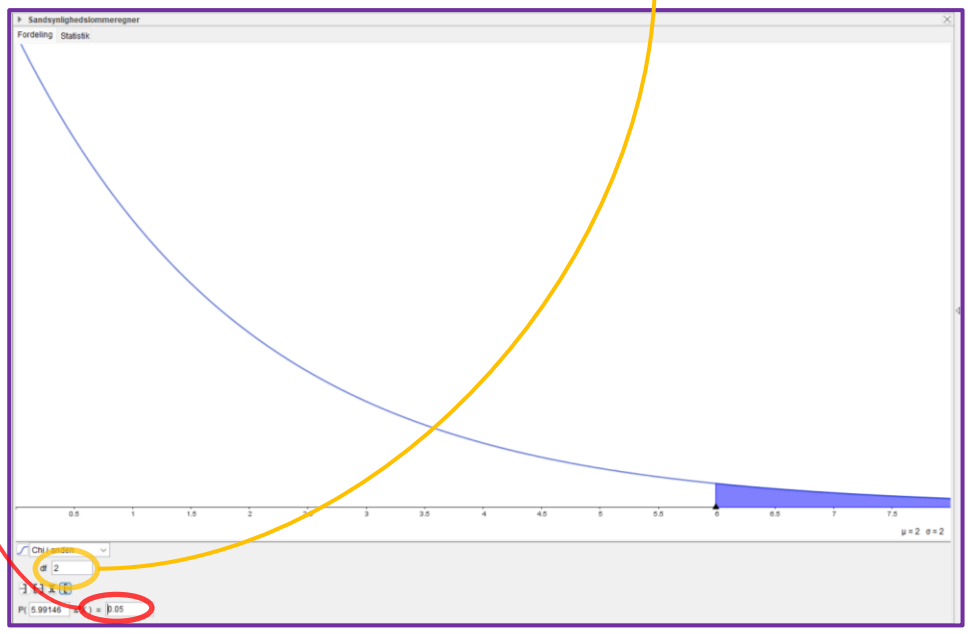
Svaret afhænger af, hvor meget man kan leve med at tage fejl. Den almindelige opfattelse – eller tradition – i samfunds- og naturvidenskaberne er, at det er ok hvis man tager fejl i ét ud af tyve tilfælde – eller med andre ord – i 5 % af tilfældene.
I matematikken siger man, at der vælges et signifikansniveau på 5 %.

Hvis – efter at testen er udført – at man finder at testsandsynligheden er større end signifikansniveauet, konkluderer man at nulhypotesen er korrekt og accepterer den.

Signifikansniveauet opdeler egentlig blot χ^2-fordelingen i to dele. Med det menes, at man betragter grafen for χ^2-fordelingen og lader arealet under grafen være lig med f.eks. 1.

Da vil det være sådan, at 95 % af dette areal ligger til højre for den kritiske værdi, og de sidste 5 % af arealet ligger til højre for det kritiske niveau.

På en graf for en χ^2-fordeling med 2 frihedsgrader ($df = 2$), vil det se således ud:



Værdien på x-aksen, som adskiller det hvide og det blå område, kaldes for ”den kritiske værdi”. I dette tilfælde er den kritiske værdi 5,99, men for en χ^2-fordeling med et andet antal frihedsgrader og/eller et andet signifikansniveau, ville den kritiske værdi ligge et andet sted.

Data:

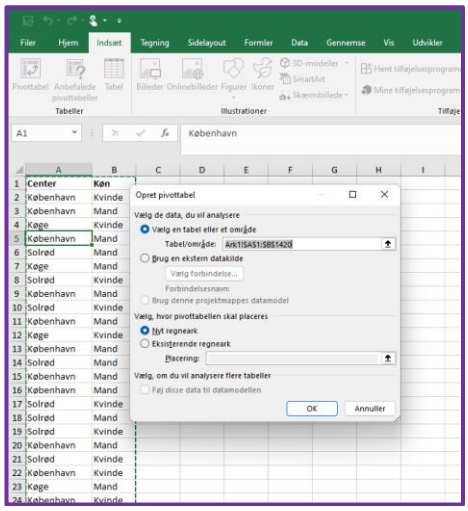
Typisk vil data komme i en Excel-fil, hvor data er skrevet i to kolonner.
Kolonne et indeholder svarmulighederne til det første spørgsmål – altså: ”København”, ”Køge” eller ”Solrød”. Den anden kolonne indeholder svarmulighederne på det andet af de to spørgsmål. I dette tilfælde, kan der således stå ”Mand” eller ”Kvinde” i anden kolonne.

I enkelte tilfælde, gives der ikke en datafil, men blot en pivottabel, som er resultatet af de manipulerede rådata. I så fald springes dette punkt bare over.

Pivottabel:

Også kaldet en antalstabel.

Data KAN markeres, men det er nemmere at klikke på én af cellerne i tabellen.
Når dette er gjort, bruges Excel-kommandoen <Indsæt> → <Pivottabel>.

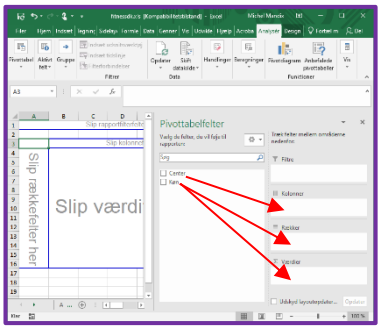


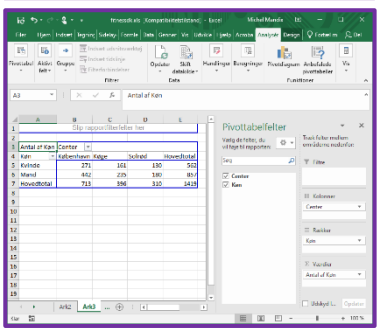
Der fremkommer en dialogboks, som den vist til venstre.

I langt de fleste tilfælde, skal indstillingerne bare accepteres som de er. Tryk på ”Ok” for at lukke dialogboksen og fortsætte.

Her vælges data til pivottabellen. Excel er rimelig intelligent, så den vælger automatisk hele tabellen. Hvis man ikke ønsker hele tabellen, skal data markeres manuelt.

Pas på med at markere. Hvis man markerer forkert, kan man risikere at data bliver redigeret utilsigtet og bliver ugyldige.





Følgende dialogboks vises.

Her er det muligt at vælge de data, som ønskes vist på hhv. rækker og kolonner.
Træk f.eks. ”Køn” ned i ”Række”-boksen og ”Center” ned i ”Kolonne”-boksen. Herved defineres række- og kolonneoverskrifterne.
For at få beregnet (talt) værdierne trækkes f.eks. ”Køn” ned i ”Værdier”-boksen. Det er faktisk ligegyldigt, om man her benytter ”Køn” eller ”Center”. Bemærk, at man IKKE kan bruge en af de etiketter, som allerede er i enten række- eller kolonneboksen. Den skal igen hentes fra oversigten.

Allerede nu, er der dannet en tabel (pivottabel) over hvor mange kvinder hhv. mænd, der træner i de forskellige byer.

Skriv evt. disse værdier ned, for de skal overføres til GeoGebra for at lave resten.

Det er også muligt at lave i Excel, hvis man kan huske de korrekte funktioner.

GeoGebra:

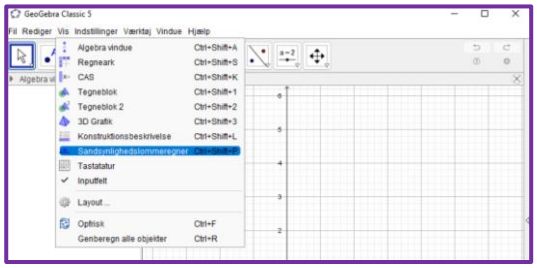
Indledningsvist gennemgås det hurtigt, hvordan man bruger GeoGebra til at lave en chi²-analyse. De enkelte elementer beskrives efterfølgende.

Der er tre steder i GeoGebra, som skal bruges til at lave en komplet chi²-analyse. Det ene beskriver statistikken (testen), det andet beskriver chi²-fordelingen og det sidste bruges til udregningen af den kritiske værdi.

Statistik:

Bruges for at finde antallet af frihedsgrader (*df*), Teststørrelsen (χ^2) og testsandsynligheden (*p*).

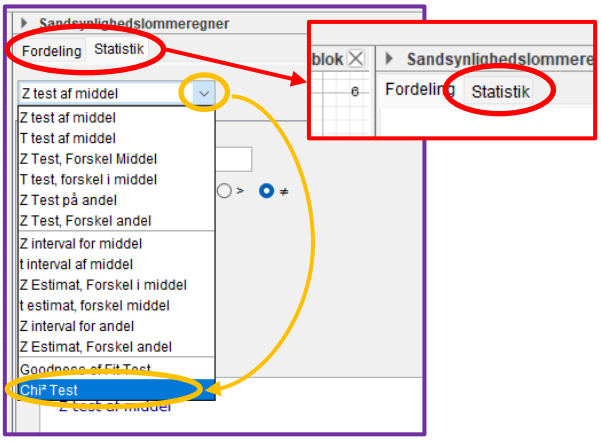
Når man har lavet (eller har fået givet) tabellen med observationerne, kan GeoGebra forholdsvist nemt lave en statistik over data.



I GeoGebra vælges først fra menuen:
<Vis> → <Sandsynlighedslommeregner>

Typisk, vil GeoGebra vise Sandsynlighedslommeregneren i et lille smalt vindue i højre side. Træk i dette vindues venstre kant for at få det vist i en ordentlig størrelse.

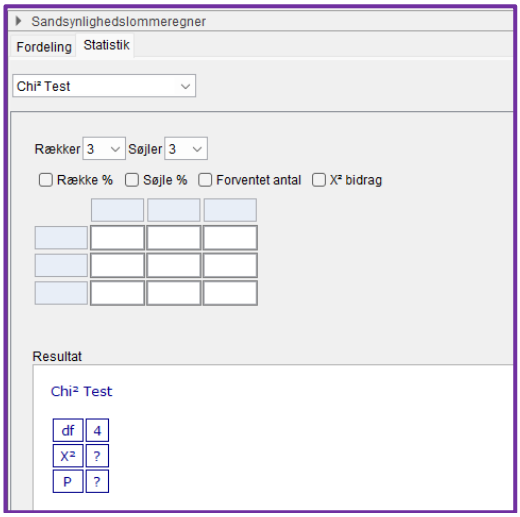
(Spoiler: Der kan skrives et helt notat om at vinduet har den størrelse til at begynde med …
Forestiller man (dem, som har lavet GeoGebra) sig virkelig, at man vælger den funktion, fordi man IKKE skal arbejde med den?)



Det ses, at der er to faneblade. For at lave chi²-testen vælges fanebladet "Statistik" (hvis det ikke allerede er valgt).

Derefter vælges fordelingstype. Hvis man lige har åbnet GeoGebra, står den som standard på "Z test af middel". Klik på den lille pil (vist i den lille gule cirkel) og vælg "Chi² Test" fra menuen

Nu er GeoGebra næsten klar til at modtage data fra tabellen.

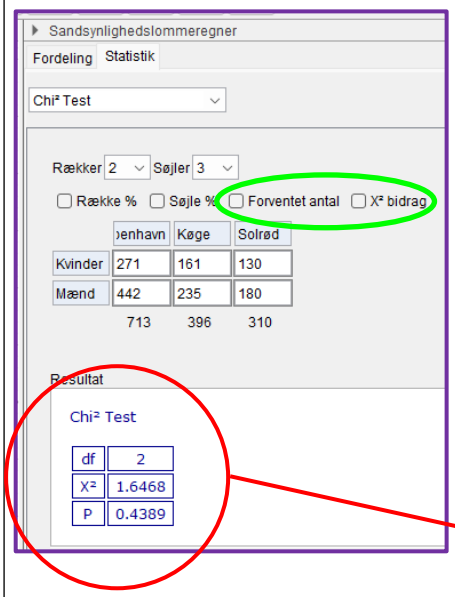


Øverst i feltet, vælger man hvor mange rækker og kolonner der er i tabellen. Her er det vigtigt, at man vælger rigtigt, for hvis man kommer i tanker om undervejs, at man har skrevet forkert, så slettes alle data (også selv om man ikke er politiker), og man kan begynde forfra.

I eksemplet er der 2 rækker og 3 kolonner. Husk, at rækketotaler og kolonnetotaler IKKE skal medregnes. Der er kun de rene observationer i tabellen, som skal indtastes.

Herefter udfyldes tabellen. Det er også en god ide at udfylde række- og kolonneoverskrifter.

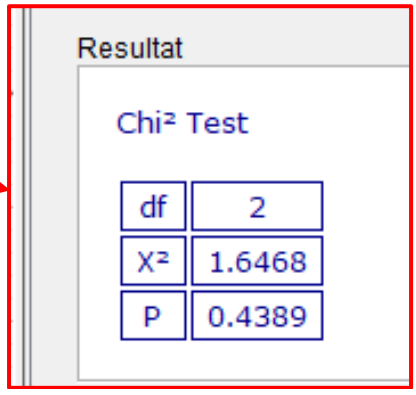
Så vidt vides, er der ikke nogen smart måde at indtaste data i tabellen. Det er ikke muligt at kopiere cellerne i Excel og derefter indsætte dem i GeoGebra.



Data indtastes, og (på godt og ondt) opdateres resultaterne løbende i det lille skema nedenunder.

Kolonnetotalerne udregnes også, men det er vigtigt at bemærke at rækketotalerne IKKE udregnes. Så man må stole på kolonnetotalerne.

Når alle data er indtastet, kan de vigtigste værdier aflæses i boksen: "Resultat".



Disse tre værdier er sådan set produktet af **denne del** af arbejdet med GeoGebra.

Den forventede værdi er: $df = 2$

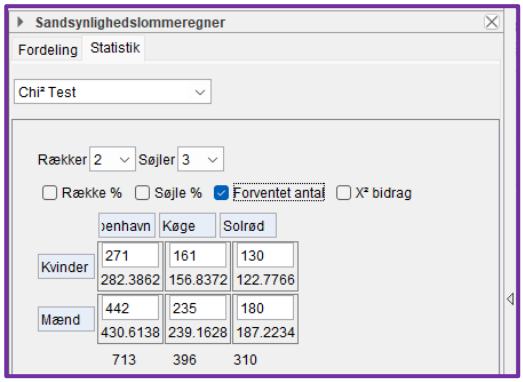
Teststørrelsen er: $\chi^2 = 1,6468$ og

Testsandsynligheden er: $p = 0,4389$

Betydningen af disse værdier og de konklusioner, som de medvirker til beskrives nærmere senere, men det skal allerede her nævnes, at der er mere information at hente. I ovenstående figur, ses det (i en grøn ellipse), at der er to afkrydsningsfelter, som kan aktiveres.

Chi²-test

Afkrydser man feltet: ”Forventet antal”, udfoldes der en ekstra række under hver observation, hvor de forventede værdier kommer til at stå.

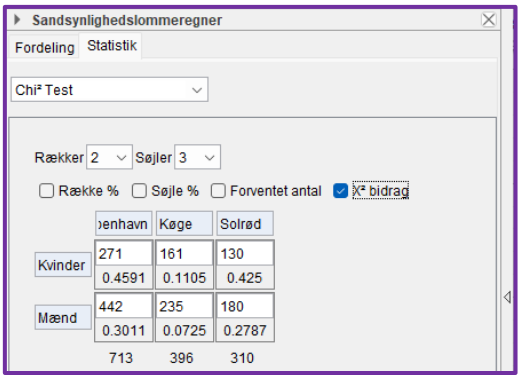


Allerede her, kan man begynde at gætte lidt på, at der ikke er nogen sammenhæng mellem de adspurgtes køn og foretrukne træningscenter.

Den største forskel mellem observation og forventet værdi, findes hos de mænd, som træner i København. Den er på ca. 11,4. Det er ikke voldsomt, set i forhold til det observerede antal, som er 442.

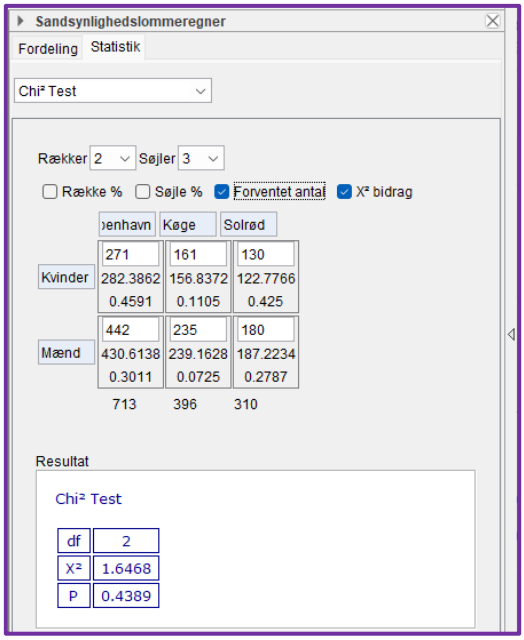
Resten af forskellene er mindre, og det ses også i testsandsynligheden, som er oppe på ca. 44 %.

Ligeledes kan man afkrydse feltet: ”X² bidrag”. GeoGebra benytter ikke det græske bogstav chi, men skriver bare et stort ”X”.
På samme måde som ved de forventede værdier, udfoldes der en ekstra række under observationerne, hvor χ^2-bidragene er blevet udregnet.



Hvis alle disse bidrag adderes, kommer man frem til summen: 1,6468, hvilket man også fik at vide i statistikoversigten.

Naturligvis kan begge felterne afkrydses på samme tid. Vær her opmærksom på, hvilke værdier, som betyder hvad, så der ikke byttes rundt på dem.



Her ses eksemplet, hvor begge felter er krydset af.

De to sidste afkrydsningsfelter, benyttes ikke til disse øvelser.

I boksen ”Resultat”, ses – som allerede nævnt – følgende værdier:

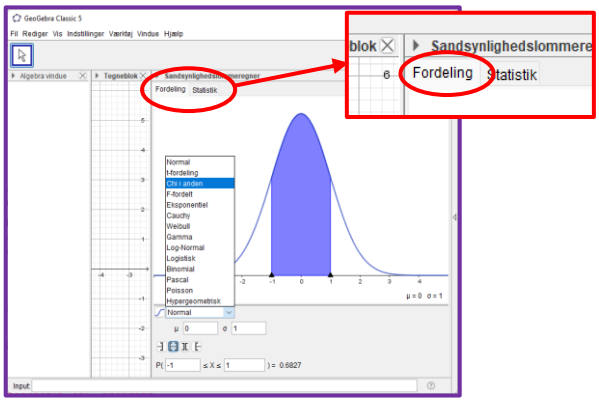
Antal frihedsgrader: $df = 2$

Teststørrelsen: $\chi^2 = 2$

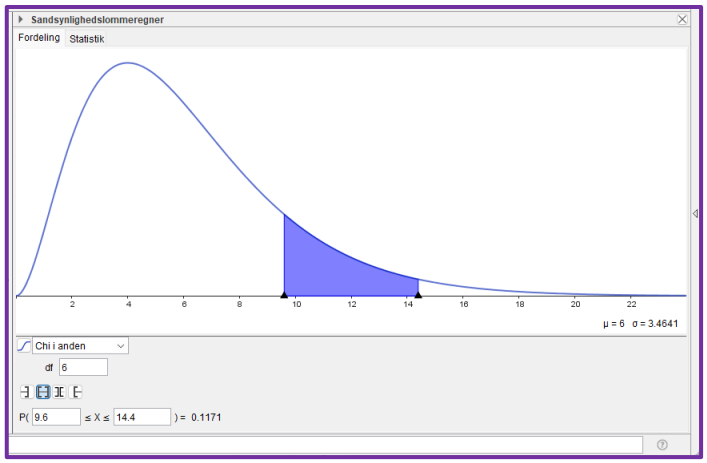
Testsandsynligheden: $p = 0,4389 \left(\approx 43,89\%\right)$

GeoGebra bruger notationen ”P” (Stort P) for testsandsynligheden. I undervisningen bruges her normalt et lille ”*p*”,

Fordeling:
Bruges for at finde testsandsynligheden, *p*.

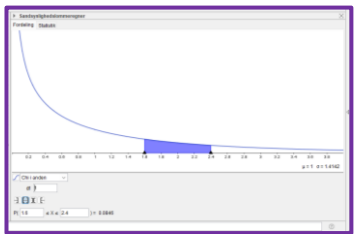


En anden ting, som kan være værdifuld i forbindelse med fortolkningen af chi²-testen er grafen for fordelingsfunktionen. Fordelingsfunktionen er "maskinrummet" i chi²-testen. De resultater der blev fundet i forbindelse med statistikken er sådan set nok til at lave vurderingen, men forståelsen af det, der bør indgå i vurderingen findes i fordelingsfunktionen. Her er det godt, at have GeoGebra som værktøj. For selve det matematiske udtryk for fordelingsfunktionen for en chi²-fordeling er ret kompliceret.

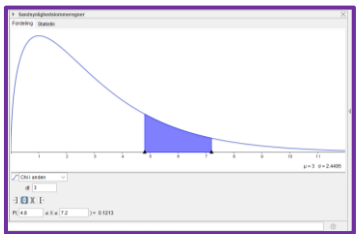


Til at begynde med, ser skærmen nogenlunde således ud, som set i figuren til venstre. Funktionens form kan variere, afhængigt af, hvad man tidligere har arbejdet med i GeoGebra.

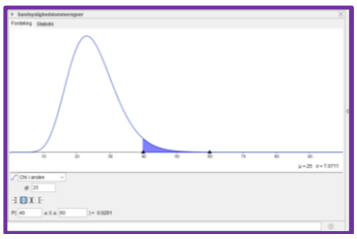
Fordelingsfunktionens form afhænger udelukkende af antallet af frihedsgrader, *df* .



df = 1



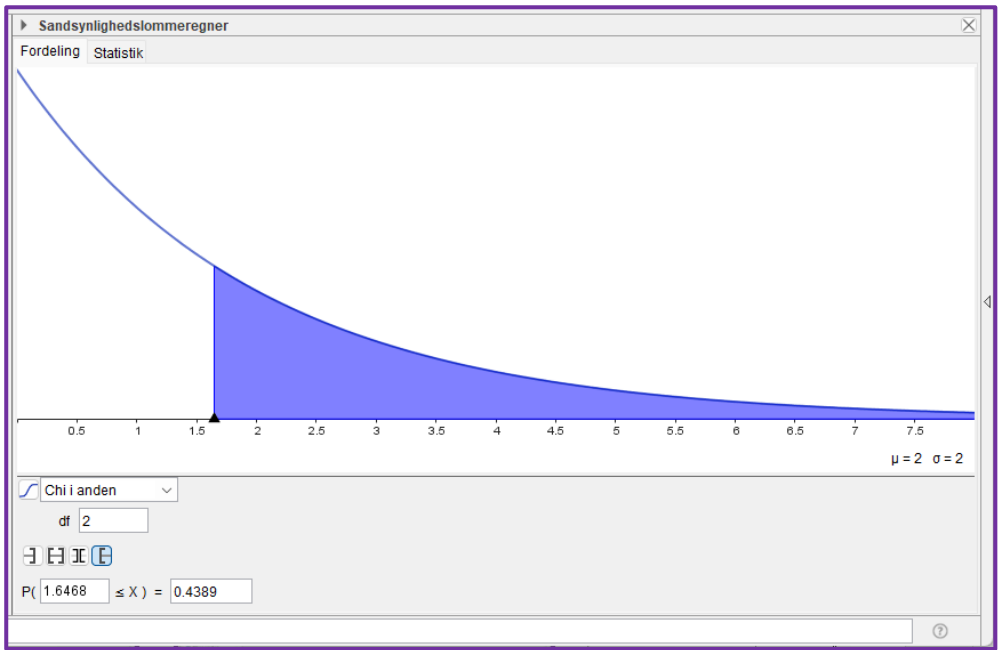
df = 3



df = 25

Som det ses af ovenstående tre figurer, vil grafen for fordelingsfunktionen for chi² mere og mere tage form af en normalfordeling, desto højere antallet af frihedsgrader *df* er.

Det erindres, at i statistik-fanebladet,blev teststørrelsen, χ^2 udregnet til 1,6468.
Dette indsættes i GeoGebra. Resultatet giver sandsynligheden, *p*.



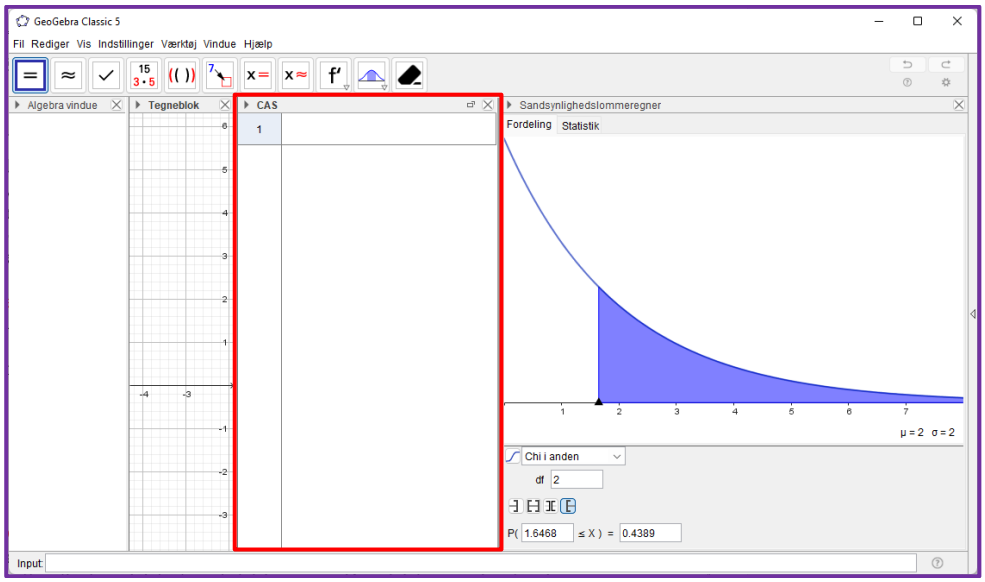
Som det ses af ovenstående figur, er sandsynligheden 43,89 %. (Skrevet som "0.4389" i det nederste felt. Dette er sandsynligheden for, at hvis der igen foretages det samme eksperiment (dvs. at spørge 1419 personer (fra det samme lokalområde), hvor de helst vil træne), så er sandsynligheden omkring 44 % for at svarene vil fordeles på samme måde. Bemærk, at det er det samme tal, som et af de tre, som blev beregnet i statistik-delen.

GeoGebra (CAS):

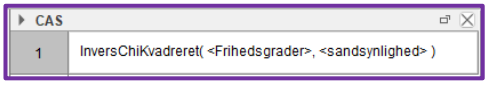
Bruges for at finde den "kritiske værdi"

Den sidste del der skal bruges fra GeoGebra er udregningen af den kritiske værdi. Den udregnes ikke i Sandsynlighedslommeregneren, ligesom de sidste par afsnit, men derimod i CAS-vinduet.

I GeoGebra vælges: Vis→CAS for at åbne CAS-vinduet, som er markeret i en rød ramme i nedenstående figur.



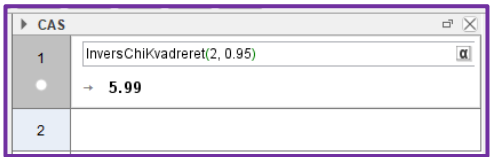
I linje 1 (eller hvor man nu er nået til), skrives kommandoen:



Naturligvis skal man ikke skrive: ”<Frihedsgrader>” eller ”<Sandsynlighed>”.

I stedet indtastes (uden skarpe parenteser) det aktuelle antal frihedsgrader og den sandsynlighed, man accepterer som signifikansniveau – eller med andre ord: ”Den garanti man vil have for at sandsynligheden taler sandt – målt i procent.” Se nedenstående figur.

I dette tilfælde ønskes et signifikansniveau på 95 %, så de to parametre bliver hhv. 2 og 0,95. (Husk at indtaste decimaltal med punktum i stedet for komma i GeoGebra – men dog stadig som komma i en besvarelse).
Svaret kommer i GeoGebra øjeblikkeligt:



Den kritiske værdi for teststørrelsen, χ^2, – altså der, hvor resultatet skifter mellem at acceptere nulhypotesen eller at forkaste den er, når teststørrelsen, χ^2, er 5,99.

Disse tre procedurer bør laves, hvis der skal laves en komplet χ^2-test.

Række- og kolonnetotaler og total:

Hvis ikke GeoGebra eller Excel selv har udregnet alle totalerne – dvs. rækketotaler, kolonnetotaler og hovedtotalen (herefter bare: totalen), så skal det gøres nu. Disse totaler er vigtige for at kunne udregne de forventede værdier.

Forventede værdier:

For hver celle i pivottabellen er udregnet (sammentalt) de antal observationer, som er kombinationen af svarmuligheden i hhv. rækken og kolonnen. Disse værdier er baseret på det virkelige eksperiment. Men kan man nu regne med, at dette er de værdier, som fås igen, hvis eksperimentet gentages? Det kan man naturligvis ikke regne med. Så derfor udregnes de forventede værdier (fv). De forventede værdier viser faktisk lige netop de værdier, man kunne forvente hvis der var fuldstændig uafhængighed. I dette tilfælde altså at hverken mænd eller kvinder ville have et favorittræningscenter, men i stedet ville vælge helt tilfældigt.
De forventede værdier skal udregnes for hver enkelt celle i tabellen, men det vil næsten altid klares med CAS – altså GeoGebra eller Excel.
Dog skal man altid lige udregne en enkelt værdi for at demonstrere det matematiske overblik. Således vil alle udregninger baseres på denne grundformel:

$$fv_{rk} = \frac{rækketotal_r \cdot kolonnetotal_k}{total}$$

For eksemplets skyld, udregnes den forventede værdi for mænd, som helst vil træne i København:

$$fv_{21} = \frac{rækketotal_2 \cdot kolonnetotal_1}{total} = \frac{857 \cdot 713}{1419} = 430,61$$

Så delkonklusionen er, at **hvis** der var tale om komplet uafhængighed, så ville der være ca. 431 mænd, som ville foretrække at træne i København.

Som allerede nævnt, er det ikke meningen at man skal sidde og udregne alle de forventede værdier i hånden. Udregn en enkelt og lad CAS klare resten. Det er til gengæld vigtigt at vise, at man KAN udregne den – særligt i en eksamenssituation.

Hvis nogen af de forventede værdier er under 5, kan der være belæg for at standse udregningen. Hvis en forventet værdi er under 5, kan det skyldes, at der ikke er en tilstrækkelig stor mængde data.
Det kan også i enkelte tilfælde betyde, at svarkombinationen i en celle ikke er særlig "populær". Hvis denne svarkombination f.eks. indeholder "ved ikke" eller en anden mulighed for at undgå at svare, må det vurderes, om en fortsat beregning er forsvarlig.

Frihedsgrader, *df* :

Frihedsgrader er et udtryk for, hvor stor tabellen er. Eller med andre ord, hvor mange parametre, der kan "skrues på", og det stadig giver det samme resultat.
Bemærk, at chi²-fordelingen udelukkende er afhængig af antallet af frihedsgrader (df), i modsætning til f.eks. normalfordelingen, som er afhængig af både middelværdien, μ, og spredningen, σ.

Frihedsgrader udregnes som: $df = (r-1)\cdot(k-1)$, hvor *r* er antallet af rækker og *k* er antallet af kolonner.

Hvorfor er frihedsgraderne konstrueret således? Egentlig består tabellen i dette eksempel af to rækker og tre kolonner. Men da totalerne er givet på forhånd (og at disse er faste og ufravigelige), så betyder det jo egentlig, at når man har udfyldt af cellerne i en kolonne, så er der jo reelt kun en celle, som man frit kan indsætte data i – så vil den anden celle i kolonnen jo være givet, da de to celler tilsammen skal give kolonnetotalen.
På samme måde skal rækkerne også give en bestemt sum (rækketotalen). Det betyder, at når en celle i en række er udfyldt, så er der frit slag i den næste celle også. Men værdien i den sidste celle er givet, da de tre celler tilsammen skal give rækketotalen.

Bidrag til χ^2 (teststørrelsen):

Det man skal vide om (bidrag til) teststørrelser er, at de er altid positive (eller 0), og jo større et bidrag er, des længere ligger de forventede og observerede værdier fra hinanden. Det vil sige at hvis bidragene er små, så ligner undersøgelsen de forventede værdier, men hvis de er store, så gør den ikke. Med andre ord, så er $\chi^2 = 0$, hvis de observerede værdier og de forventede værdier er identiske, og hvis χ^2 er et (relativt) stort tal, så ligger de observerede værdier og de forventede værdier langt fra hinanden.

Bidragene til teststørrelsen er som regel kommatal, så man bliver nødt til at runde af. Når man gør det, skal man altid have mindst to decimaler med, ellers bliver resultatet for upræcist.

Som med de forventede værdier regnes bidraget til teststørrelsen ud for hver celle. Igen forventes det, at man viser et enkelt eksempel, og overlader resten til CAS.

Igen for eksemplets skyld udregnes bidraget til teststørrelsen for den celle, som beskriver mænd, der helst vil træne i København:

$$Bidrag\_\chi^2_{21} = \frac{(Obs_{21} - fv_{21})^2}{fv_{21}} = \frac{(442 - 430,61)^2}{430,61} = 0,3011$$

Samlet teststørrelse: χ^2:

Den samlede teststørrelse findes ved at addere bidragene fra alle cellerne.

$$\chi^2 = \sum Bidrag\_\chi^2_{rk} = \sum \frac{(Obs_{rk} - fv_{rk})^2}{fv_{rk}}$$

$$\Updownarrow$$

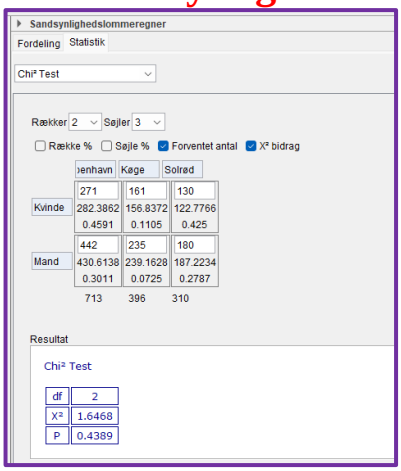
$$\chi^2 = \frac{(Obs_{11} - fv_{11})^2}{fv_{11}} + \frac{(Obs_{12} - fv_{12})^2}{fv_{12}} + \cdots + \frac{(Obs_{23} - fv_{23})^2}{fv_{23}}$$

$$\Updownarrow$$

$$\underline{\chi^2 = 1,6468}$$

Den samlede teststørrelse er et mål for forskellen mellem de forventede værdier og de observerede værdier, og bruges til at konkludere hvorvidt der er tale om uafhængighed eller ej. Dvs. om forskellen mellem de forventede værdier og de observerede værdier skyldes den almindelige tilfældige fluktuation (udsving) eller om der er tale om en signifikant forskel. Denne konklusion omtales senere i dette notat.

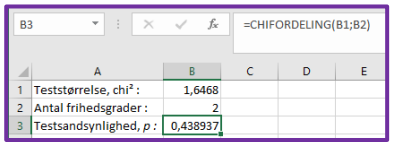
Testsandsynlighed:



Testsandsynligheden, *p*, er netop den værdi, som ligger til grund for selve konklusionen. Testsandsynligheden findes nemmest ved at aflæse værdien i GeoGebra. Selve beregningen er kompliceret, men kan udledes på to forskellige måder i GeoGebra.

Den nemmeste måde er at aflæse resultatet i chi²-testen i Resultat-boksen.

Chi²-test



Testsandsynligheden kan også udregnes vha. Excel. Brug funktionen "CHIFORDELING".
Her kræves to værdier: Teststørrelsen, χ^2, og antallet af frihedsgrader, *df*.

Uanset beregningsmetoden vil det sige, at hvis nulhypotesen er sand, så er der knap 44 % chance for at de data, som er brugt findes. Da de 44 % er højere end signifikansniveauet på 5 %, accepteres nulhypotesen.

Fortolkning:

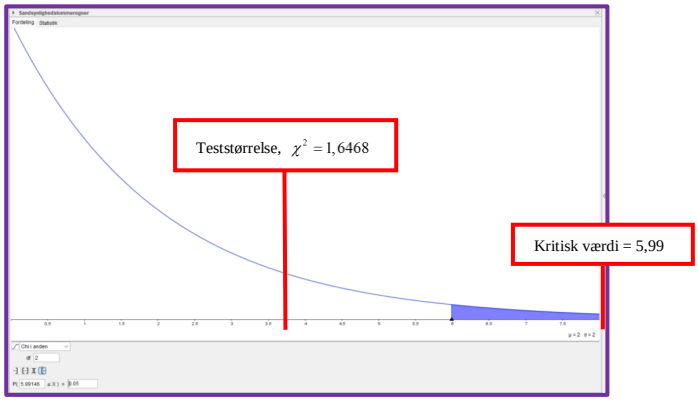
Alt, hvad der er sket indtil nu har været udregninger af diverse værdier og bestemmelse af fordelingens grafiske udseende.

Nu skal der tolkes på dette, og det kan gøres på flere måder.

Før selve fortolkningen beskrives, opsummeres der for, hvilke værdier, der nu kendes:

Inden de konkrete værdier analyseres, kontrolleres der for, om nogle af de forventede værdier er 5 eller derunder.
Det viser sig, at den mindste af de udregnede forventede værdier er lig med: 122,78 så der er ingen fare for, at data bliver for "grovkornede".

Grænseværdier: Eller med andre ord de værdier, som skiller de to situationer: Nulhypotesen accepteres eller nulhypotesen forkastes.		**De fundne værdier:** De værdier, som er fremkommet ved at beregne på de data, som er givet i spørgeundersøgelsen.		**Konklusion:**
Signifikansniveau:	5 %	Beregnet testsandsynlighed:	43,89 %	**Da testsandsynligheden er større end signifikansniveauet, accepteres nulhypotesen.**
Kritisk værdi:	5,99	Teststørrelse, χ^2 :	1,6468	**Teststørrelsen er mindre end den kritiske værdi, og derfor accepteres nulhypotesen.**



Da nulhypotesen accepteres, betyder det altså, at der er statistisk sandsynlighed for, at der ikke er nogen sammenhæng mellem de adspurgtes køn og deres foretrukne træningscenter.

Algoritme:

Det hele kan sammenfattes til en algoritme – eller sagt på en anden måde: en opskrift i en kogebog, som blot skal følges til punkt og prikke. Jo mere man kan standardisere og genbruge tekst fra denne "opskrift" eller skabelon, desto nemmere er det at sammenfatte en opgave, som omhandler en chi²-test.

Gør følgende:

1) De observerede data skal indsamles.
Ofte kommer data i en Excel-fil med mange rækker, hvor hver række repræsenterer en persons svar på de to spørgsmål.
Disse data samles og optælles nemmest i Excel, ved brug af en pivottabel.
Er data givet som en sådan tabel, kan dette punkt springes over.

2) Opstil en nul-hypotese, H_0 og efterfølgende en alternativ hypotese H_1.
Her er det vigtigt at:
H_0: (nul-hypotesen) er en påstand, som siger, at der er uafhængighed (dvs. at der IKKE er en sammenhæng) mellem de to spørgsmål (variable).
H_1: (den alternative hypotese), er den komplementære (modsatte eller inverse) påstand, som siger, at der IKKE er uafhængighed (dvs. at der ER en sammenhæng) mellem de to spørgsmål.

3) De forventede værdier, fv_{rk}, udregnes under antagelse af, at der er uafhængighed.
De udregnes automatisk i GeoGebra, men vedlæg mindst en manuel udregning.

4) Chi²-teststørrensen udregnes. Den skal bruges i konklusionen.
Teststørrelsen udregnes automatisk i GeoGebra, men vedlæg mindst en manuel udregning af et af bidragene til teststørrelsen.

5) Antallet af frihedsgrader, *df* bestemmes. De skal ligeledes bruges i konklusionen.
df udregnes automatisk i GeoGebra, men da det er så simpel en udregning, bør den også udregnes manuelt.

6) Testsandsynligheden, *p*, udregnes. Den skal også bruges i konklusionen.
Hvis testsandsynligheden er meget tæt på 0, kan det være en fordel at udregne den præcise værdi i Excel, da GeoGebra vil afrunde til 0 efter få decimaler.

7) Konklusionen skrives. Husk dette! Det er selve målet med opgaven. Beregnede værdier er ALDRIG nok!

That's all folks!

Et hurtigt eksempel:

Givet en **antalstabel** som viser svarene på, om de adspurgte er hhv. "Kvinde" eller "Mand" og om de inden et forestående folketingsvalg i 2020 ville stemme på "Rød blok" eller"Blå blok".

Der skal laves en chi²-test for at vurdere om der er en matematisk sammenhæng mellem de adspurgtes køn og deres politiske ståsted.

"Danskerne i 2020"	Rød blok	Blå blok	Total
Kvinde	1735	1193	2928
Mand	1183	1186	2369
Total	2918	2379	5297

Hypoteser:

H_0 – Der er INGEN sammenhæng mellem de adspurgtes køn og politiske tilhørsforhold.

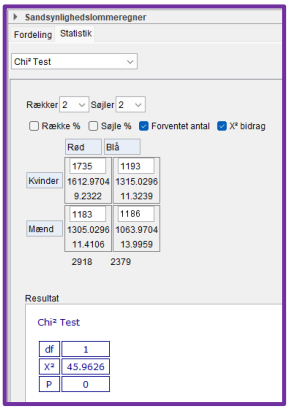
H_1 – Der ER en sammenhæng mellem de adspurgtes køn og politiske tilhørsforhold.

Signifikansniveau:
Der testes på et 95 % signifikansniveau.

Forventede værdier:
Bidrag til teststørrelsen:
Teststørrelsen, χ^2:
Antal frihedsgrader, *df*:

Disse fire ovenstående trin udregnes her vha. GeoGebra. Dog vises et eksempel på udregning af en forventet værdi, bidrag til teststørrelsen (Kvinder, der stemmer på Rød blok) og antallet af frihedsgrader.

Det noteres også, at ingen af de udregnede forventede værdier er lig med 5 eller derunder, hvilket betyder, at testen kan fortsættes uden videre.



$$fv_{21} = \frac{rækketotal_2 \cdot kolonnetotal_1}{total} = \frac{2928 \cdot 2918}{5297} \Leftrightarrow \underline{fv_{21} = 1612{,}97}$$

$$Bidrag\_\chi^2{}_{21} = \frac{\left(Obs_{21} - fv_{21}\right)^2}{fv_{21}} = \frac{\left(1735 - 1612{,}97\right)^2}{1612{,}97}$$

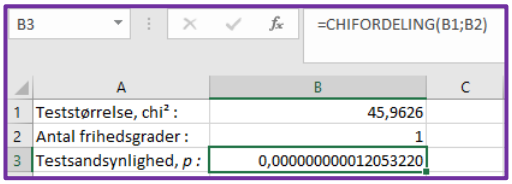
$$\Updownarrow$$

$$\underline{Bidrag\_\chi^2{}_{21} = 9{,}2322}$$

$$df = (r-1)\cdot(k-1) = (2-1)\cdot(2-1) = 1\cdot 1 \Leftrightarrow \underline{df = 1}$$

Den kritiske værdi udregnes i GeoGebra:

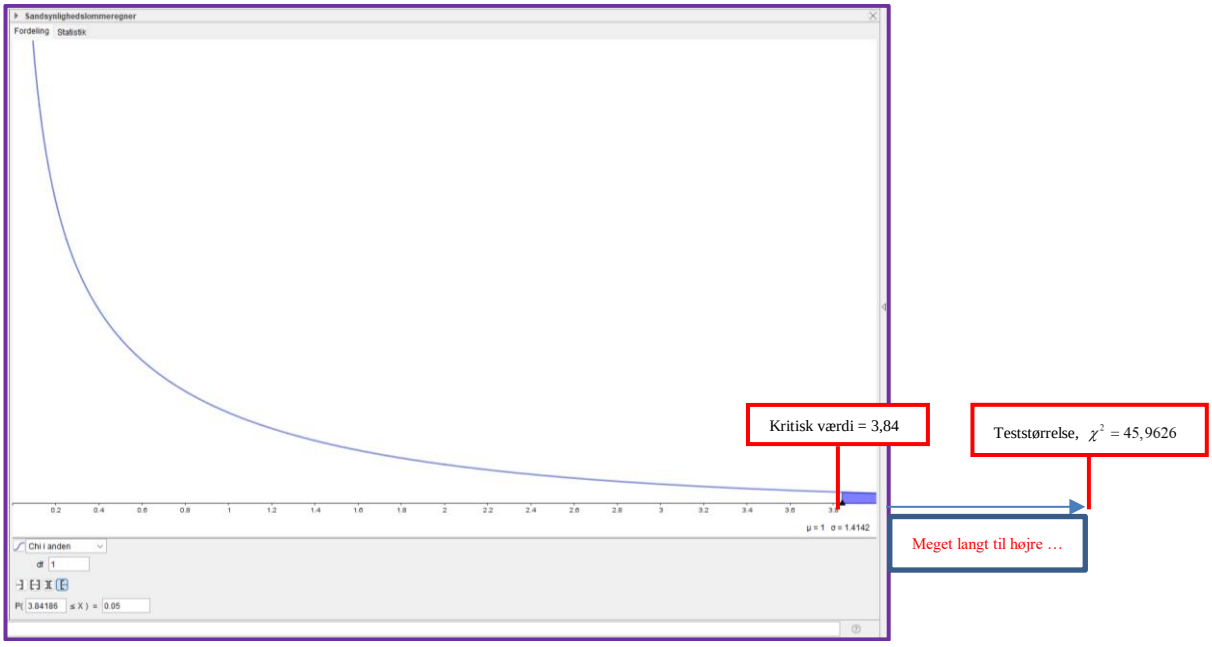
Det bemærkes, at den i GeoGebra udregnede testsandsynlighed er lig med 0.
Dette betyder, at sandsynligheden er meget lille, men mon ikke et bedre tal er til rådighed?
Derfor udregnes dette i Excel:



Der er altså tale om et meget lille tal, så det accepteres.

Nu kendes alle de værdier, der skal til for at lave en ordentlig konklusion:

Grænseværdier: Eller med andre ord de værdier, som skiller de to situationer: Nulhypotesen accepteres eller nulhypotesen forkastes.		**De fundne værdier:** De værdier, som er fremkommet ved at beregne på de data, som er givet i spørgeundersøgelsen.		**Konklusion:**
Signifikansniveau:	5 %	Beregnet testsandsynlighed:	≈0,00 %	**Da testsandsynligheden er mindre end signifikansniveauet, forkastes nulhypotesen og den alternative hypotese antages.**
Kritisk værdi:	3,84	Teststørrelse, χ^2 :	45,9626	**Teststørrelsen er større end den kritiske værdi, og derfor forkastes nulhypotesen og den alternative hypotese antages.**



Da nulhypotesen forkastes, betyder det altså, at der er statistisk sandsynlighed for, at der er en sammenhæng mellem de adspurgtes køn og deres foretrukne træningscenter. De to variable er altså afhængige. Dog kan der ikke siges noget om, hvori denne afhængighed består.