Projekt 9.14 Er rygning usundt – den første store undersøgelse i Whickham, England

(*Whickham undersøgelsen var et led i en række større undersøgelser af rygningens sundhedsskadelige effekt. Men undersøgelsen kom til at spille en helt anden rolle både i debatten og i diskussionen blandt statistikere. Som det vil fremgå, var der noget helt galt i rapporten og de konklusioner man kunne drage. Undersøgelsen fik den modsatte effekt af det, man havde håbet, idet den gav rygerne nye argumenter. Men den kom samtidig til at sætte projektørlyset på et indtil da ret overset problem, nemlig Simpsons paradoks, som allerede var fremlagt af Simpson i 1951. Simpsons paradoks undersøges nærmere i HEM2, projekt 9.15, men forklares også her.*
*Den statistiske undersøgelse bygger på kendskab til Chi-i-anden fordelingen, som er behandlet i HEM2, projekt 9.13. Hovedkilden til nedenstående kan hentes **her**.*

1. Er det usundt at ryge?

Har rygning indflydelse på helbredet? Det forsøgte en berømt undersøgelse af 1314 kvinder fra Whickham at svare på. Whickham er et blandet land- og bydistrikt tæt ved NewCastle upon Tyne i England. I årene 1972-74 blev de spurgt, om de var rygere, og tyve år senere registrerede man, hvor mange af de adspurgte, der stadigvæk var i live. Man fandt da følgende resultater, som vi har samlet i en krydstabel med deres r*ygevaner* som forklarende variabel og deres *helbredstilstand* som responsvariabel. Som i grafer har vi sat den uafhængige forklarende variabel vandret og den afhængige responsvariabel lodret

		Rygevaner		
	Observerede værdier	ja	nej	I alt
Helbreds-til-stand	død	139	230	369
	i live	443	502	945
	I alt	582	732	1314

Kilde: Appleton, David R., French, Joyce M. and Vanderpump, Mark P. J. (1996). Ignoring aÜovariate: An example of Simpson's paradox. The American Statistician, 50, 340-341

Spørgsmålet er nu, om der er belæg i tabellen for en sammenhæng mellem rygevaner og helbredstilstand? Har rygere en anden helbredstilstand end ikke-rygere?

For at kunne belyse denne problematik med et statistisk test, bør vi først gøre os klart, i hvilket omfang det er rimeligt at betragte den pågældende gruppe af kvinder som en repræsentativ stikprøve for en langt større population, fx alle indbyggerne i England? Kan vi reelt slutte noget om englændernes helbredstilstand ud fra en enkelt gruppes opførsel?

Normalt sikrer man sig repræsentativitet ved at vælge deltagerne i stikprøven tilfældigt. Men disse kvinder er valgt alt andet end tilfældigt: De er fx alle sammen fra et stærkt afgrænset område af England. Der er også mange andre variable, som der ikke taget højde for.

Øvelse 1
Nævn tre andre variable, der kunne have indflydelse på undersøgelsens resultat.

Hvis nogle af de variable, der er kommet frem i øvelsen, faktisk har indflydelse på helbredstilstanden, er det selvfølgelig afgørende, at disse variable er tilfældigt fordelt på de to grupper af rygere og ikke-rygere, så det reelt er effekten af rygning, vi ser, og ikke effekten af en sådan skjult variabel. I første omgang vil dog ignorere dette aspekt.

Første skridt i den statistiske undersøgelse er at fastlægge *nulhypotesen*:
Der er ingen sammenhæng mellem helbredstilstand og rygevaner.

Nulhypotesen kan også formuleres således: *De to variable er uafhængige*.

Når vi skal teste nulhypotesen, starter vi med at fastlægge et signifikansniveau på 5%.

Dernæst udregner vi χ^2 – teststørrelsen for afvigelsen mellem de observerede værdier og de forventede værdier. Nulhypotesens antagelse om uafhængighed betyder, at de *forventede værdier* har samme procentfordeling for rygere og ikke-rygere. Vi får derfor følgende tabel over de forventede værdier:

		Rygevaner		
	Observerede værdier	ja	nej	I alt
Helbreds-tilstand	død	163,43	205,55	369
	i live	418,57	526,45	945
	I alt	582	732	1314

De forventede værdier fremkommer således: 28,08 % (dvs. 369/1314) af de 582 rygere giver, at 163,43 forventes at være døde, mens 71,92% af de 582 rygere eller i alt 418,57 forventes at være i live. Tilsvarende for ikke-rygerne.

Praxis: Udregning af forventede værdier for en krydstabel (uafhængighedstest)
Man finder de forventede værdier for de enkelte kategorier ud fra procentfordelingen i den *samlede* population. Det kræver, at man har række- og søjletotaler til rådighed.

		Rygevaner		
	Observerede værdier	ja	nej	I alt
Helbreds-tilstand	død	$\frac{369}{1314}\cdot 582 = 163{,}43$		369
	i live			945
	I alt	582	732	1314

Øvelse: 2
Gennemfør udregningen af de forventede værdier i den ovenstående tabel i detaljer,

χ^2 – teststørrelsen udregnes som en sum af alle bidrag af formen:

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

Her får vi:

$$\chi^2 = \frac{(139-163{,}43)^2}{163{,}43} + \frac{(443-418{,}57)^2}{418{,}57} + \frac{(230-205{,}55)^2}{205{,}55} + \frac{(502-526{,}55)^2}{526{,}55}$$

$$\chi^2 = 3{,}65 + 1{,}43 + 2{,}91 + 1{,}14 = 9{,}12$$

Antallet af frihedsgrader i en 2 x 2 tabel er 1. Dette betyder teoretisk betyder, at forventningen i middel til teststørrelsen er omkring 1, hvis nulhypotesen holder. Så meget tyder på, den ikke kan holde. Det kan vi nu undersøge nærmere på 2 måder.

Eksperimentel metode:
Vi antager, at nulhypotesen holder, dvs. at der er uafhængighed. Lad os forestille os alle 1314 kvinder havde et kartotekskort med de to oplysninger skrevet på hver sin halvdel af kortet: Rygevaner nederst og død / i live skrevet øverst.

Vi klipper nu disse kort midt over, samler i to bunker, og blander kortene med rygevaner vilkårligt rundt. Så lægger vi de to kortbunker ved siden af hinanden og limer dem sammen igen, så vi nu får nye kort, men stadig med rygevaner nederst og død / i live øverst. Med de nye kort er der stadig 582 kvinder som ryger og 732 der ikke gør, og der er stadig 369 kvinder, der er døde, og 945 der er i live. Det er alene kombinationerne af rygning og helbred, der er ændret. Men i de sammenblandede kort er helbredstilstanden nødvendigvis uafhængig af rygevaner. Derfor er der nogenlunde samme fordeling af helbredstilstanden for rygere og for ikke-rygere. Vi har altså på denne måde simuleret nulhypotesen, dvs. uafhængigheden af rygevaner og helbredstilstand.

Øvelse 3

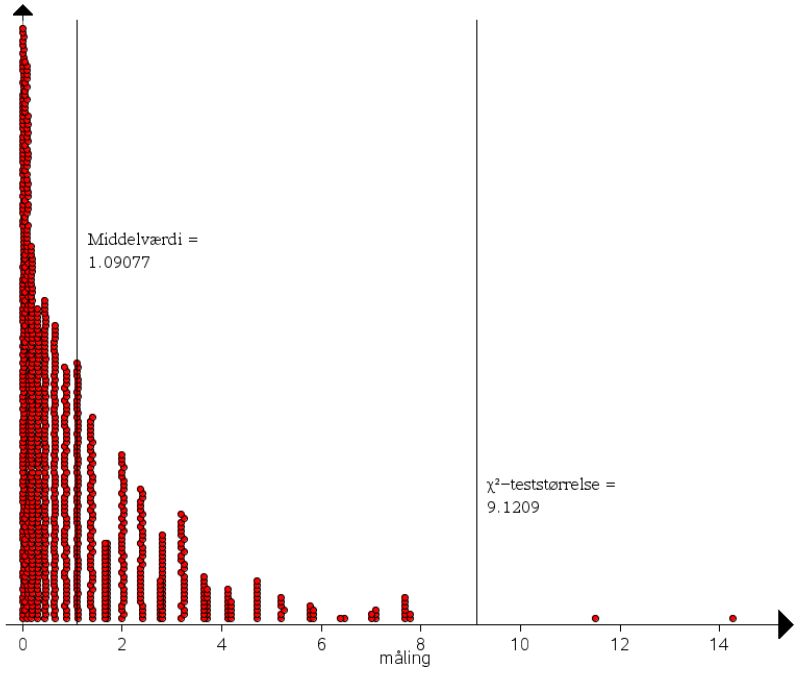
En sådan simulering (omrøring) kan gennemføres i et værktøjsprogram: Den ene variabel holdes fast, mens den anden blandes vilkårligt rundt, og resultatet samles i en ny antalstabel.

a) Gennemfør en sådan omrøring. Det er gjort her i TI Nspire CAS, men kan gennemføres på smme måde i andre programmer..

b) Opstil formlen for χ^2 – teststørrelsen for en simulering efter samme princip som ovenfor.

c) Gennemfør et mindre antal simuleringer, fx 20 gange. Ser det ud til at være nemt at finde en simulering, der er lige så skæv som den observerede?

d) Gennemfør nu 1000 simuleringer, hvor teststørrelsen registrerer og præsenter fordelingen af teststørrelsen i et prikdiagram (som vist nedenfor) eller i et passende histogram. Plot også den observerede teststørrelse.



Teststørrelsen er så usædvanlig, at kun to simuleringer ud af 1000 giver en større værdi. De to skæve udfald svarer til et skøn over p-værdien på 0,2%

Konklusion: Nulhypotesen forkastes. Vi slutter derfor at der er en sammenhæng mellem rygevaner og helbredsstilstand.

Formelbaseret metode:

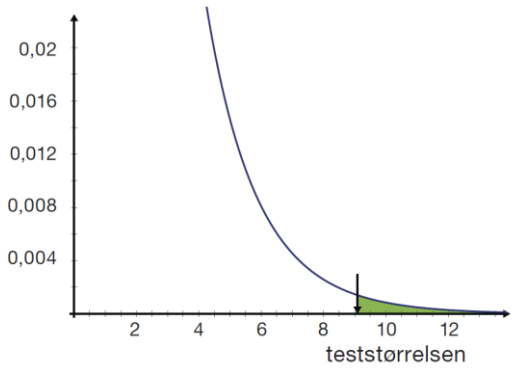
Vi har beregnet teststørrelsen til at være 9,12. En 2 x 2 antalstabel har 1 frihedsgrad. Vi finder derfor den følgende *p*-værdi ud fra den teoretiske χ^2-fordeling og med brug af den såkaldt kumulerede χ^2-fordelingsfunktion

Øvelse 4

a) Benyt dit værktøjsprogram til at finde *p*-værdien såvel grafisk som ved beregning ud fra den kumulerede χ^2-fordeling.

b) Hvor lille skal teststørrelsen være, for at vi ikke længere kan forkaste nulhypotesen?

c) Udnyt det indbyggede uafhængighedstest i et værktøjsprogram til automatisk at udføre testen og derigennem få udregnet fx testværdien og *p*-værdien

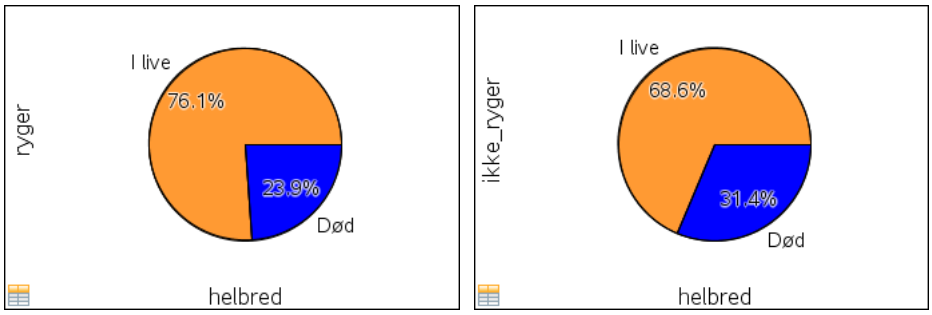


.

p-værdien er altså 0,0025 svarende til 0,25 % og ligger derfor klart under signifikansniveauet på 5 %.
Konklusion: Nulhypotesen forkastes. Vi slutter derfor at der er en sammenhæng mellem rygevaner og helbredsstilstand.

2. Der er noget galt - skjulte variable og Simpsons paradoks

Men der er et problem: Sammenhængen peger den forkerte vej! Kigger vi nærmere på de observerede procentfordelinger, ser vi nemlig, at rygerne har den største chance for at overleve. Det ser altså ud til at være sundt at ryge!



De 76% af rygerne er stadigvæk i live mod kun 69% af ikke-rygerne. Så hvad foregår der egentlig?

Problemet viste sig netop at være en *skjult variabel*. Gruppen af rygere og ikke rygere er ikke ens fordelt i forhold til alder. Hvis vi opdeler i tre aldersgrupper: Ung (fra 18-34 år), midaldrende (fra 35-54 år) og gammel (mindst 55 år) så finder vi de følgende krydstabeller:

Ung

	Ja	Nej
Død	5	6
I live	174	213

Midaldrende

	Ja	Nej
Død	41	19
I live	198	180

Gammel

	Ja	Nej
Død	93	205
I live	71	109

Øvelse 5

a) Opret tabellerne i et regneark og udregn række og søjle-summerne.
b) Udregn overlevelsesprocenterne for rygere og ikke-rygere i de tre aldersgrupper.
c) Illustrér resultatet grafisk.
d) Hvordan ser sammenhængen nu ud mellem rygevaner og helbred?

Øvelse 6
Du kan ***her*** hente en endnu mere detaljeret opdeling mht alder og rygevaner.
Hent denne og stil dig samme spørgsmål som i øvelse 5.

Den ovenstående situation, hvor en statistisk sammenhæng *vender*, når man inddrager en skjult variabel i analysen, kaldes *Simpsons paradoks*. Den understreger hvor forsigtig man skal være med at drage slutninger om årsagssammenhænge ud fra en statistisk sammenhæng.
Problemet ligger i den *manglende variabelkontrol*. Når vi skal finde ud af hvilke faktorer, der har indflydelse på levealderen, er det vigtigt, at vi kun ændrer på én forklarende variabel ad gangen. Når vi fokuserer på rygning, skal alle andre faktorer altså alt andet lige være ens fordelt i de to grupper: rygere og ikke-rygere. Det kan være svært i praksis at sikre sig dette. Bare det at fastlægge hvilke variable, der kan tænkes at have indflydelse på levealderen, kan være svært nok. I praksis vil man derfor ofte komme ud for, at stikprøverne er *skævt sammensat* med hensyn til andre variable, end dem man undersøger.

Definition: Bias
En stikprøve, der overrepræsenterer eller underrepræsenterer individer med bestemte karakteristika (variable), og hvor disse har indflydelse på det spørgsmål, man undersøger, siges at være præget af bias

Den eneste sikre strategi er, at alle andre variable er tilfældigt fordelt på de to grupper i stikprøven, såkaldt *statistisk variabel kontrol*, så en eventuel indflydelse fra skjulte variable udjævnes. Men også dette kan være svært at styre i praksis.

Hvis man er i samarbejde med et andet fag, kan der muligvis ud fra dette fags viden peges på en mekanisme, der kan forklare påvirkningen fra den ene variabel til den anden. Men også dette kan vise sig at være yderst vanskeligt. Havde vi fx ikke haft tabellerne med aldersfordelingen, kunne vi jo ikke have påvist, hvor problemet lå.

Øvelse 7.
Hent den fil, der er anvendt som kilde ***her***.
1. Læs den engelske tekst – den er kun 1½ side – og noter dig forfatternes pointer undervejs. Gennemfør en diskussion i gruppen eller i hele klassen, hvor I dykker ned i disse pointer og gør jer fortrolige med Simpsons paradoks, og med hvordan vi i praksis skal håndtere dette.
2. Prøv selv at konstruere et eksempel på fejlslutning, hvor Simpsons paradoks spiller os et puds.