

Projekt 10.2 Rygning, Whickham-undersøgelsen – og Simpsons paradoks

(projektet anvender χ^2 -fordelingen i den statistiske undersøgelse)

Har rygning indflydelse på helbredet? Et mærkeligt spørgsmål i vores tid. Men det var det ikke i det meste af det 20. århundrede, hvor eksempelvis en af de argeste modstandere var datidens mest berømte statistiker, Ronald Fisher.

En nu berømt undersøgelse af 1314 kvinder fra Whickham skulle give det endegyldige svar på dette.

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.

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

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

For at kunne belyse denne problematik med en 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 opførelse?

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 bestemt afgrænset område af England. Der er også mange andre variable, der ikke er 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 øvelse 1, 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 vi 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, begynder 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:

		Forventede værdier		
		Rygevaner		
Helbreds-tilstand		ja	nej	I alt
	død	163,44	205,56	369
	i live	418,56	526,44	945
	I alt	582	732	1314

De forventede værdier fremkommer således: Først omregnes kolonnen I alt til procentandele: 369 udgør 28,08% af 1314 og 945 udgør 71,9% af 1314. Antagelsen om samme fordeling for rygere og ikke-rygere gør, at vi udregner disse to procentdele af henholdsvis 582 og 732. Eksempelvis er 28,08% af 582 lig med 163,44.

Øvelse 2

Gennemfør udregningen af de forventede værdier i den ovenstående tabel i detaljer.

χ^2 -teststørrelsen udregnes nu 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, at middeltallet for teststørrelsen er omkring 1, hvis nulhypotesen holder. Så meget tyder på, den ikke kan holde. Det kan vi nu undersøge nærmere på to måder.

Ekspérimentel metode:

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

Vi klipper nu disse kort midt over, samler dem 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

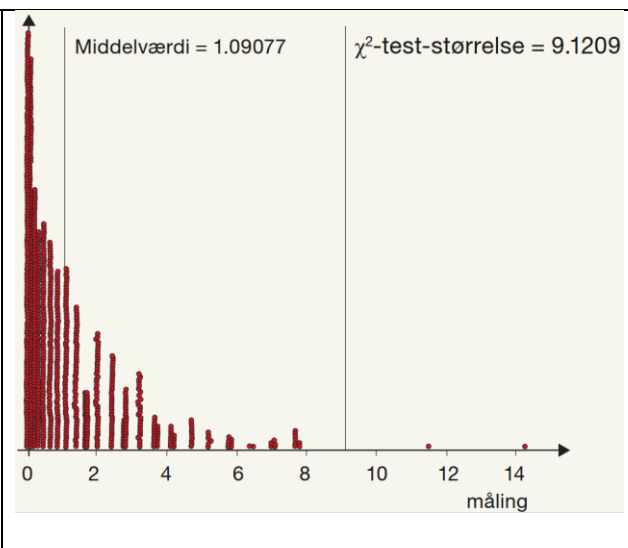
En sådan simulering (der populært kaldes for en *omrøring*) kan gennemføres i et værktøjsprogram: Den ene variabel holdes fast, mens den anden blandes vilkårligt rundt, og resultatet samles i en ny antalstabel.

a) Gennemfør en sådan omrøring.

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

c) Gennemfør et mindre antal simuleringer, fx 20. 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 registreres, og præsenter fordelingen af teststørrelsen i et prikdiagram (som vist her) eller i et passende histogram. Plot også den observerede tekststørrelse.



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 mellem rygevaner og helbredstilstand.

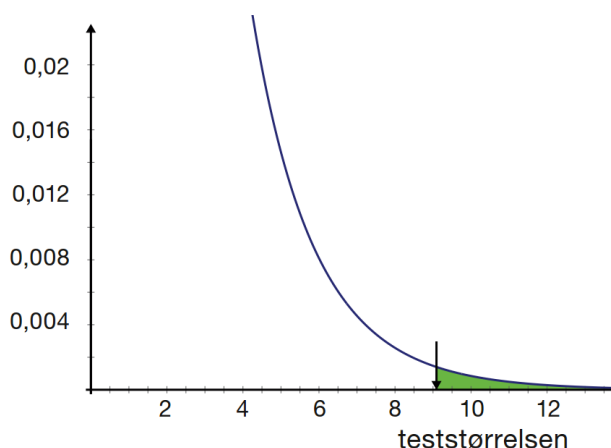
Formelbaseret metode:

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

Grafen viser tæthedsfunktionen for χ^2 -fordelingen frihedsgrad.

Værktøjsprogrammet giver:

$$\chi^2 Cdf(9.1209, \infty, 1) = 0.002527$$



med 1

Øvelse 4

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

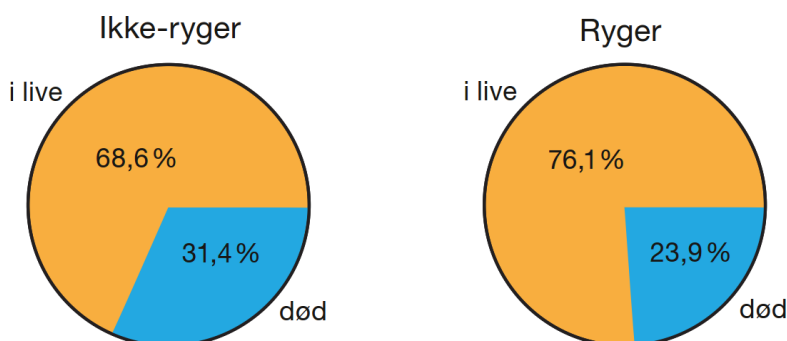
b) Hvor lille skal teststørrelsen være, for at vi ikke længere kan forkaste nulhypotesen?

- c) Udnyt den 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 helbredstilstand.

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 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), Gammel (mindst 55 år), så finder vi ind bag de sære tal, her opstillet i krydstabeller:

	Ung		Midaldrende		Gammel	
	Ja	Nej	Ja	Nej	Ja	Nej
Død	5	6	41	19	93	205
I live	174	213	198	180	71	109

Øvelse 5

- Kopier tabellerne ind i et værktøjsprogram, og udregn række- og søjlesummerne.
- Udregn overlevelsesprocenterne for rygere og ikke-rygere i de tre aldersgrupper.
- Illustrer resultatet grafisk.
- Hvordan ser sammenhængen nu ud mellem rygevaner og helbred?

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.

I HEM1, kapitel 9 kan du dels finde et større projekt med uddybende materialer om Simpsons paradoks med mange eksempler. Der findes også et mindre projekt 9.6, der ved hjælp af Simpsons paradoks afdækker et eksempel på den systemiske racisme i USA.

Når vi skal finde ud af, hvilke faktorer der har indflydelse på levealderen, er det vigtigt, at vi kun ændrer på en 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.

*I Projekt 10.3 *The Mortality of Doctors* kan du finde en grundig behandling af den undersøgelse, der endeligt påviste rygningens forfærdeligt skadende virkning*