(kolofonside)

Forsideillustrationer: Ulla Korgaard, Designeriet

Projektmateriale 1 i tilknytning til Susanne Ditlevsens film: *Statistiske metoder i hverdagsliv og i neurovidenskab*
Emner i indledende statistik
faglig redaktion: Bjørn Grøn

Film og tilhørende materialer er produceret med støtte fra bla. Undervisningsministeriets udlodningsmidler, IKV, SDU og Cryptomathic

Videoerne hostes af L&R Uddannelse A/S • Vognmagergade 11 • DK-1148 • København K • Tlf: 43503030 • Email: info@lru.dk.

Projektmateriale 1 i tilknytning til Susanne Ditlevsens video: Emner i indledende statistik

Indhold

0. Indledning

Disse projektmaterialer er skrevet i tilknytning til filmen *Statistiske metoder i hverdagsliv og neurovidenskab*, der indgår i serien om matematisk forskning: *10 danske matematikere – 10 matematiske fortællinger*.

I filmen fortæller professor ved Københavns Universitet Susanne Ditlevsen dels om nogle af de centrale problemstillinger – og faldgruber – i den indledende statistik, og dels om sin forskning i hvordan neuroner kommunikerer. Til den sidste del af filmen foreligger *projektmaterialer 2*, mens det foreliggende materiale knytter sig til den indledende statistik, og nogle af de eksempler, der fortælles om i filmen.

Projektmaterialet er opdelt i en række kapitler, der kan gennemgås hver for sig. Det enkelte kapitel bygger således ikke på de foregående. Det fælles tema for hele projektmaterialet er: Udtagning af stikprøver. Hvorfor gør man det og hvordan gør man det.
Gennem øvelserne i de enkelte kapitler sætter vi – som i filmen – fokus på, hvor let det er at lave fejl, for derved at skærpe opmærksomheden om, at forudsætningen for at kunne lave god statistik er, at man forstår sine data, og at man har indsamlet disse så korrekt som forholdene nu engang tillader.

Materialet i de enkelte kapitler kan både indgå i statistikundervisningen på C, B og A. På C-niveau kan aktiviteterne indgå som et led i en introduktion til elementær statistik og sandsynlighedsregning. På B- og A-niveau kan nogle af aktiviteterne anvendes til en perspektivering af sandsynlighedsteori og statistik. Der er samtidig potentiale til studieretningsprojekter i flere af emnerne, hvor der ligger ekstra materiale på hjemmesiden.

De første to aktiviteter handler om repræsentativitet af stikprøver, og et af målene med øvelserne er at skærpe opmærksomheden på begrebet skjulte variable.

De næste to aktiviteter – henh. om soldyrkning og hudkræft, og om nedskydning af amerikanske kampfly under 2. verdenskrig – handler om selektionsbias. I begge aktiviteter er der unikke kildematerialer, som kan anvendes fx i en srp. For sagen om soldyrkning og hudkræft drejer det sig om de originale artikler i Journal of Epidemiology, hvor Niels Keiding og Theis Lange fra BioStat på KU gik i rette med den indsendte artikel, og hvor det endte med at tidskriftet publicerede både en artikel og en længere redaktionel kommentar, hvor de måtte give Keiding og Lange ret i deres kritik.

Aktivitet 5, 6 og 7 – henh. om rygning og om racefordomme – viser, hvorledes skjulte variable kan føre til Simpsons paradoks. Her er hypotesetest i spil og vi er på B og A-niveau.

Aktivitet 8 og 9 – henh. om fejlscreening af sygdomme og om overvågning af potentielle terrorister – inddrager betingede sammenhænge og kan fx inddrages i forløb om betingede sandsynligheder. Der perspektiveres til sidst til Bayesiansk statistik, som den eksempelvis anvendes i retssager.

1. Stikprøver og population (C, B og A)

Vi ønsker at få svar på et eller flere spørgsmål om en bestemt population. Det kunne eksempelvis være, om der er en sammenhæng mellem levealder og det job man har. Eller om der er en sammenhæng mellem politisk holdning til et bestemt spørgsmål og uddannelsesniveau. Det kunne også dreje sig om kvalitetskontrol: Er der virkelig 1 liter i letmælkskartonerne fra et bestemt mejeri? Eller: Virker fyrværkeriraketter efter hensigten eller eksploderer nogle af dem ved affyringen?

Hvis vi kunne undersøge eller udspørge hele populationen, var der ingen grund til at tage stikprøver. Men de sidste eksempler illustrerer, at det ofte er umuligt. Og de første eksempler illustrerer, at var det principielt muligt, ville det både være overordentlig dyrt og meget besværligt. Så vi tager i stedet stikprøver:

Definition: Stikprøve og population

- *Populationen* angiver den mængde af personer, dyr, produkter, hændelser osv, som vi gerne vil vide noget om. Populationen består af *individer*.
- En *stikprøve* er den delmængde af individer, som vi undersøger nærmere for at kunne sige noget om hele populationen.

En stikprøve skal udtages, så den er *repræsentativ*, ellers kan vi ikke konkludere noget om hele populationen. En 1.g klasse kan eksempelvis ikke være repræsentativ i alle mulige forhold for alle skolens elever, og heller ikke for alle 1.g'erne i hele Danmark.

Det er imidlertid lettere sagt end gjort at opnå repræsentativitet af en stikprøve, fordi der normalt er rigtig mange variable i spil. Og hvilke af de variable har en afgørende indflydelse på det vi netop nu har i fokus? Tænk fx på opinionsmålinger: politiske holdninger kan variere med køn, indkomstforhold, uddannelsesforhold, geografi, alder, job, boligforhold, familieforhold … . Bare det at fastlægge hvilke variable, der kan tænkes at have indflydelse på den politiske holdning – eller på levealder eller hvad man nu lægger an til at undersøge – kan være svært nok.

Man kan sige, at alt dette bør de statistiske bureauer have styr på – men igen er det lettere sagt end gjort: Lad os antage, vi har et projekt, hvor vi undersøger levealderen, og at vi har fastlagt, hvilke variable, *der kan tænkes at have indflydelse* på denne. Når vi dernæst skal finde ud af hvilke af disse variable, *der faktisk har indflydelse* på levealderen, er det vigtigt, at vi kun ændrer på én forklarende variabel ad gangen.

Repræsentativitet af stikprøver er i fokus i flere af projekterne i de følgende kapitler. Når man skal svare på sådanne spørgsmål er det vigtigt at dele problemet op i to forskellige:
1. *Er den faktiske stikprøve*, vi har taget, repræsentativ?
2. *Er den metode*, der er anvendt til at indsamle stikprøven, en acceptabel metode til at sikre repræsentativitet?

Der er ingen metode, der kan garantere, at vi med sikkerhed kan svare ja på det første. Det ligger simpelthen i sagens natur: Når stikprøver indsamles tilfældigt, så vil vi af og til få en stikprøve, der er meget anderledes end populationen. Det svarer til, at vi af og til vil opleve at der slås 5 seksere med 5 terninger.

Øvelse 1.1
Hvor tit vil det ske, at man i et terningespil slår 5 seksere ud af 5 mulige?
Hvornår vil du reagere, og sige at du tror der er snyd med i spillet?

Selv om vi kan svare ja på spørgsmål 2, kan vi godt opleve, at svaret af og til vil være et nej til spørgsmål 1. Men svarer vi nej til spørgsmål 2 kan vi være rimeligt sikre på, at så er svaret også altid nej til spørgsmål 1. Derfor er det centrale spørgsmål nr 2: Hvilken metode har vi anvendt i indsamlingen af stikprøven.

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

Der er mange forskellige typer af bias, og vi vil møde nogle af disse i de følgende øvelser og i projekterne i de følgende kapitler.

Selv om man har gennemført sine statistiske beregninger korrekt og anvendt korrekte grafiske værktøjer, så skal man alligevel altid være varsom med at drage slutninger om *årsagssammenhænge* ud fra en statistisk sammenhæng. 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. Hvis ens undersøgelse eksempelvis resulterer i et meget overraskende resultat, så kan forklaringen som omtalt være, at der er problemer med stikprøven, men det kan også skyldes, at der er skjulte variable på spil

Definition: Skjulte variable
En skjult variabel er en forklarende variabel med signifikant betydning for det spørgsmål, man undersøger, men som vi ikke har afdækket eller måske slet ikke har kendskab til (endnu).

Det kan forekomme lidt underligt at definere og give navn til noget vi *ikke* kender, og man skal også passe på, at begrebet *skjult variabel* ikke kommer til at træde ind på scenen, hver gang, man ikke kan finde et svar, en begrundelse eller en løsning på en opgave: Anvendes begrebet *skjult variabel* skal man kunne sandsynliggøre, at der må være noget ekstra på spil, at der er noget vi ikke har afdækket. Og sandsynliggørelsen kan netop ske ud fra, at man har fundet nogle meget mærkelige resultater.

Da William Thomsen (senere bedre kendt som lord Kelvin) i 1862 beregnede sig frem til at Jorden var mellem 20 og 40 millioner år gammel, skete det på grundlag af en matematisk model for afkøling. I sine oprindelige beregninger vedgår han, at der kan være skjulte variable på spil (*a source now unknown to us*). Og det er der – Jorden holdes varm af det radioaktive materiale i undergrunden. Den historie har vi fortalt i C-bogens kapitel 4, og det er et klassisk eksempel på, at forklarende variable kan være helt ukendte for os – radioaktivitet blev først opdaget nogle årtier senere. I kapitlet om Simpsons paradoks ser vi, hvordan skjulte variable kan vende en forklaring helt på hovedet.

1.1 Øvelser om stikprøver og population (C, B og A)

I svarene på de følgende øvelser skal du anvende de begreber som stikprøve, population, repræsentativitet, bias og skjulte variable, der er omtalt i det foregående afsnit.

Øvelse 1.2
På en skole med 700 elever ønsker en af de politiske ungdomsorganisationer at få mulighed for at stille et bord op, hvor eleverne i spisefrikvarteret kan hente materialer og få information. Da skolens ledelse siger nej, opfordrer organisationen alle elever til at tilkendegive om de er for eller imod dette. 127 afgav deres stemme og heraf støttede 92 forslaget.

- Hvad er populationen og hvad er stikprøven?
- Hvor stor en andel stemte ja til forslaget?
- Kommenter undersøgelsen.
- Rådgiv organisationen om, hvorledes de kunne foretage en mere kvalificeret undersøgelse.

Øvelse 1.3
En amerikansk politiker udtaler til en Tv-station, at han er interesseret i at høre vælgernes holdning til en lov om våbenkontrol. Hans sekretær opgør efter en uges tid, at de har modtaget breve om spørgsmålet fra 361 vælgere. 323 var imod loven.

- Hvad er populationen og hvad er stikprøven?
- Kommenter undersøgelsen.
- Rådgiv politikeren om, hvorledes der kunne foretages en mere kvalificeret undersøgelse.

Øvelse 1.4
På et site på nettet kan man læse følgende:

STEM NU

Ville du have ret til at gøre en ende på dit liv, evt. med lægelig bistand hvis du fik at vide, at du var uhelbredelig syg.

Hvis ja: klik her Hvis nej: klik her

Giv en vurdering af metoden i denne statistiske undersøgelse.

Øvelse 1.5
Et sundhedsmagasin ønsker at undersøge om store doser vitamintilskud forbedrer sundhedstilstanden. Bladet anmoder de af læserne, som gennem længere tid har taget store doser vitamintilskud om at skrive ind og fortælle om positive og negative erfaringer med dette. 2754 læsere skriver ind. 93% fortæller, at de kan spore en vis forbedring af helbredet.

- Hvad er population og hvad er stikprøve.
- Kommenter undersøgelsen.
- Giv en vurdering af, om andelen af hele befolkningen, der vil få forbedret sundhedstilstanden ved at indtage store doser vitamintilskud, er større, den samme eller mindre end 93%. Begrund dit svar.

Øvelse 1.6
En kvindelig redaktør af et stort amerikansk kvindemagasin spurgte engang sine læsere, om de ville stille sig tilfreds med mænd, der gav dem kærlighed og hengivenhed, men ingen sex.
90.000 kvinder skrev ind og 72.000 svarede ja.

- Hvad er populationen og hvad er stikprøven.
- Giv en vurdering af metoderne i denne statistiske undersøgelse og af hvilke konklusioner, der kan drages.

Øvelse 1.7
Antallet af studenter på et bestemt institut er vokset betydeligt, uden der er blevet mere plads eller flere undervisningslokaler. Instituttet vil undersøge studenternes syn på holdstørrelse og andre spørgsmål vedr. de fysiske rammer. De sætter ressourcer af til at interviewe 25 ud af i alt 450 studenter.
Rådgiv dem om, hvorledes du synes de skal udvælge de 25, så der kan drages bedst mulige konklusioner ud af materialet.

Øvelse 1.8
En bestemt sygdom påvirker de røde blodlegemer og forårsager stor smerte. Et medicinsk præparat til behandling af sygdommen er udviklet, og kvaliteten af præparatet ønskes afprøvet på en population på 300 patienter, der har haft særligt mange smerteanfald.

- Forklar hvorfor det ville være en dårlig strategi at lade alle 300 få den nye medicin.
- Beskriv et forsøg, der kunne give information om pågældende præparats virkning over for smerteanfald.

Øvelse 1.9
En gruppe matematiklærere tilrettelægger undervisningen i deres klasser således, at eleverne kan vælge mellem at deltage i en lærerstyret klasseundervisning, eller at arbejde selv i eget tempo, med samme stof, men ud fra lærebogen, arbejdsark og interaktive programmer.
Efter ét år ønsker de at sammenligne de to måder at lære matematik på. De giver derfor alle eleverne den samme prøve og sammenligner så resultaterne for at se, om den ene af grupperne scorer klart mere end de andre.

- Kommenter den valgte metode til at sammenligne
- Antag du har 30 elever, der er villige til at følge begge typer af undervisning. Hvordan ville du sammenligne de to læringsmetoder, og afgøre hvilken, der er mest effektiv?

Øvelse 1.10

En sproglærer mener, at studiet af fremmedsprog også forbedrer de studerendes beherskelse af modersmålet dansk. Han laver test på forskellige årgange, og det viser sig faktisk, at personer, der studerer eller har studeret et fremmedsprog er bedre til dansk.
Giv en vurdering af undersøgelsen.

Øvelse 1.11

Vil indtagelse af urtete styrke helbredet hos de ældre? Dette ønsker en gruppe studerende at undersøge. Over en periode på 6 måneder besøger de nogle tilfældigt udvalgte beboere på et plejehjem og serverer urtete for dem. Efter 6 måneder viser det sig, at de beboere, der fik serveret urtete faktisk har færre sygedage, end de som ikke fik serveret noget.
Giv en vurdering af undersøgelsen og af troværdigheden af resultatet

Øvelse 1.12

En artikel i Politiken 21-02-2004 omhandler fænomenet "Mænd der får tæsk", og giver en række overraskende oplysninger om voldelige mænd og voldelige kvinder. Bl.a. fortælles det, at "Australian Institute of Health and Welfare" har opgjort, at 40% af børnemishandlinger i landet står de enlige mødre for. De enlige fædre står kun for 5%.
Kommenter disse oplysninger. Du må gerne selv indføre taleksempler i din argumentation.

Øvelse 1.13

Et konsulentfirma bliver bedt om at sammenligne kvaliteten af behandlingen på to hospitaler A og B. En lille arbejdsgruppe indhenter et materiale, som de opstiller i følgende tabel:

Antal der overlevede operative indgreb		
	Hospital A	Hospital B
Døde	63	16
Overlevede	2037	784
I alt	2100	800

Tabellen dannede grundlag for arbejdsgruppens vurdering af de to hospitaler. Gruppen fremhævede B som det bedste.
En af arbejdsgruppens medlemmer er imidlertid ikke tilfreds med materialet og indhenter supplerende oplysninger, som stilles op i følgende tabel:

Antal der overlevede operative indgreb				
	God helbredstilstand før operation		Dårlig helbredstilstand før operation	
	Hospital A	Hospital B	Hospital A	Hospital B
Døde	6	8	57	8
Overlevede	594	592	1443	192
I alt	600	600	1500	200

Skriv din egen konklusion om kvaliteten af behandlingen på hospital A og B.
Kommenter samtidig undersøgelsen og materialet i de to forskellige tabeller.

2. Stikprøver, skjulte variable og selektionsbias (C, B og A)

Denne aktivitet bygger på et materiale, som Inge Henningsen udarbejdede til den indledende undervisning i statistik på Matematisk institut, KU.

Lærervejledning.

Formålet med aktiviteten er bl.a. at introducere/formalisere/diskutere statistiske begreber, så eleverne får nogle sproglige og matematiske værktøjer, der kan sætte dem i stand til at kritisere og forklare misbrug/forkert brug af sammenhæng i tabeller. Samt specielt at klargøre forholdet mellem statistisk og kausal afhængighed.

Udgangspunktet for arbejdet er en række opdigtede avisnotitser. De er selvfølgelig opdigtede, men realistiske om end på en overdreven måde. Den stokastiske variation er taget væk for at gøre pointerne tydeligere. Ved at beskæftige sig med situationer, som de har erfaring med og derfor intuition omkring ledes eleverne frem til at operationalisere en række af de statistiske begreber, population, stikprøver, repræsentativitet, skjulte variable, bias, samt se sammenhængen med de teoretiske begreber statistisk og kausal afhængighed.

Eleverne arbejder i grupper med en række af de små cases, og de skal med diagrammer (fx Venn-diagrammer) eller med tabeller, hvor de selv indfører nogle taleksempler, gennemføre analyser af situationerne og være i stand til at præsentere et ræsonnement, der afdækker fejlene i "avisnotitserne". Endelig skal de selv formulere et antal tilsvarende notitser – eller endnu bedre: finde eksempler fra aviserne.

2.1 Opdigtede avisnotitser

Øvelse 2.1. Humanister går i små sko.
Ved en sammenligning af en gruppe naturvidenskabelige og en gruppe humanistiske studenter opdagede man en overraskende forskel. De naturvidenskabelige studenter var i gennemsnit næsten 3 cm højere end de humanistiske og brugte 11/2 nummer større i sko.

Øvelse 2.2. Hoftebrud medfører forhøjet kræftrisiko.
Ved en undersøgelse på Rigshospitalet har det vist sig, at patienter indlagt for brud på lårhalsknoglen har en 10 gange så høj risiko for at få kræft inden for en periode på 5 år, som patienter indlagt med henblik på en meniskoperation. Lægerne undersøger nu, hvorfor en forskellig traumelokalisering giver en så markant forskel i risikoen for få kræft. På baggrund af undersøgelsen påpeger bandagist NN, at systematisk brug af firmaets nyudviklede underbukser med stødabsorberende indlæg til beskyttelse af hoftepartiet, vil kunne reducere den alt for høje danske kræftrisiko.

Øvelse 2.3. Jørn Hjorting skyld i ekstra kræfttilfælde.
En undersøgelse foretaget af Amtsrådsforeningen har vist, at personer der lytter til "De ringer, vi spiller" har en 3 gange så høj risiko for at få kræft inden for en periode på 5 år, end dem der jævnligt ser udsendelsen "Beat". Kræftrisikoen ved at høre Jørn Hjorting er forhøjet med 2000% i forhold til at høre børneradio. Disse fund bør få konsekvenser for programlægningen.

Øvelse 2.4 Mascara beskytter mod testikelkræft.
En undersøgelse på Frederikssund Sygehus, der tidligere har haft en overhyppighed af testikelkræfttilfælde, har vist, at brug af mascara reducerer forekomsten af testikelkræft drastisk. I en 5-års periode har man i hele Frederiksborg Amt kun haft 1 tilfælde af testikelkræft blandt mascarabrugere, hvor man i en aldersmæssigt tilsvarende gruppe ville have forventet 37 tilfælde.

Øvelse 2.5. P-piller disponerer for rygning. Beskytter mod hoftebrud.
En undersøgelse i Glostrup af alle kvinder over 20 år har vist, at brug af p-piller medfører øget tendens til rygning. Til gengæld synes p-pillerne at give en vis beskyttelse mod hoftebrud.

Øvelse 2.6. Nedlægger intensiv afdeling.
Amtssygehuset i XX har besluttet at nedlægge den intensive afdeling, idet en undersøgelse har vist, at dødeligheden på denne afdeling ligger langt over hospitalets gennemsnit. I fremtiden vil alle alvorligt syge patienter blive indlagt på ortopædkirurgisk afdeling. Hospitalsdirektøren tror, at den nye organisering vil blive en væsentlig sundhedsmæssig gevinst for amtets borgere.

3. Soldyrkere lever længere (B og A)

Materialerne, der arbejdes med i dette afsnit, knytter sig direkte til fortællingen om, at soldyrkere og folk med hudkræft lever længere end resten af befolkningen. Artiklerne fra Politiken og fra tidsskriftet International Journal of Epidemiology kan hentes fra en mappe på hjemmesiden.

I oktober 2013 kunne man i Politiken læse nedenstående artikel. Den har et ret sensationelt indhold og blev da også bragt på forsiden.

Øvelse 3.1
Læs artiklen og præsenter informationerne heri med anvendelse af de statistiske begreber, du har lært:

a) Kan du indkredse, hvad populationen er? Angiv fx nogle individer, der er med i populationen og nogle der ikke er med.
b) Kan du lokalisere en stikprøve? Hvad er denne gruppe en stikprøve af?
c) I hvilken forstand er den repræsentativ for populationen
d) I artiklen tales om gennemsnitlig levealder. Hvem er det man tager gennemsnittet af, og hvordan udregnes et sådant?

1. ARTIKEL OM SOL OG HUDKRÆFT

15. OKT. 2013

Soldyrkere lever meget længere

Ny forskning blandt 4,4 millioner danskere viser, at soldyrkere i gennemsnit lever seks år længere. Kræftens Bekæmpelse finder tallene spændende.

Henrik Larsen

Et hold danske forskere er på vej med en videnskabelig artikel, som rejser spørgsmålet:
Er der særlige livsforlængende 'sager' i solens stråler?

Artiklen står foran offentliggørelse i videnskabstidsskriftet International Journal of Epidemiology og viser, at mennesker, som har været ivrige soldyrkere – og har fået såkaldt almindelig hudkræft, den ikkedødelige form for hudkræft – i gennemsnit lever seks år længere end befolkningen som helhed.

Overdreven solforskrækkelse
Gennemsnitsdanskeren – kvinder og mænd under ét – bliver i dag 80 år. Men når det gælder denne gruppe soldyrkere, kan vi altså se, at de i snit når at fejre 86-års fødselsdagen. Og at de i øvrigt har en lavere forekomst af både blodpropper i hjertet og knogleskørhed end resten af befolkningen, siger en af forskerne bag undersøgelsen, professor Børge Nordestgaard, Herlev Hospital.

Forskerne kan ikke påpege den direkte årsagssammenhæng mellem soldyrkning og fundene: *Men tallene som sådan lyver ikke. Det vil være oplagt gennem en række forskningsprojekter at prøve at blive klogere på, om solen har nogle helbredsbefordrende egenskaber, vi hidtil ikke har kendt til*, siger han og tilføjer: *Man skal bestemt have respekt for solen – og undgå solskoldninger, ikke mindst*

hos børn og unge, for på den måde især at mindske forekomsten af den slemme form for hudcancer, malignt melanom. Men vi har de seneste år set en vis grad af solforskrækkelse, og det er efter min mening at gå i den anden grøft.

Ingen egentlig årsagssammenhæng

Hos Kræftens Bekæmpelse – der står bag de såkaldte solråd, som blandt andet opfordrer danskerne til at omgås solen med større forsigtighed – finder Inge Haunstrup Clemmensen, overlæge i Kræftens Bekæmpelse, undersøgelsen spændende: *Det vil være oplagt gennem en række forskningsprojekter at prøve at blive klogere på, om solen har nogle helbredsbefordrende egenskaber, vi hidtil ikke har kendt til.*

Børge Nordestgaard, professor, Herlev Hospital fortæller om undersøgelsen: *Den rejser en mistanke om, at der kan være en sammenhæng mellem solens stråler og de positive helbredsfund, der påvises hos gruppen af mennesker, som har haft almindelig hudkræft. Men det er bare vigtigt at holde fast i, at den ikke viser en egentlig årsagssammenhæng*. Børge Nordestgaard har sammen med tre kolleger – læge Peter Brøndum-Jacobsen og seniorforsker Sune Nielsen, begge Herlev Hospital, og overlæge Marianne Benn, Gentofte Hospital – analyseret helbredsrelaterede data fra 4,4 millioner danskere i alderen 40-100 år.

Analysen dækker perioden 1980-2006, og forskerne havde adgang til en meget bred vifte af registre, hvorfra de i anonymiseret form kunne trække oplysninger om hver eneste: om blandt andet sygdomme, uddannelsesniveau, bopæl, erhverv, dødsårsag og dødstidspunkt.

Forskerne gik en form for bagvej

Selv om danske registre tager meget med om den enkelte, indeholder de ingen oplysninger om soldyrkningsvaner. Så her måtte forskerne gå en form for bagvej, fortæller Børge Nordestgaard: *Det er velkendt, at mennesker, der får almindelig hudkræft, som hovedregel har udviklet de celleforandringer, der er tale om, fordi de har opholdt sig ekstraordinært meget i solen. For at finde eventuelle helbredsbevarende sammenhænge, der muligvis kunne skyldes solen – vores egentlige mål – sammenlignede vi derfor 130.000 borgere, der havde haft almindelig hudkræft, med resten af befolkningen*.

Og de positive opdagelser, vi gjorde blandt disse 130.000 mennesker i form af længere gennemsnitslevetid og lavere forekomst af blodpropper i hjertet og af knogleskørhed, viste sig at være helt uafhængige af forhold som køn, alder, om man bor i byen eller på landet – ligesom uddannelse og erhverv heller ikke spiller nogen rolle. Derfor er det nærliggende at stille spørgsmålet: Kan dette i en eller anden udstrækning være knyttet til solens aktivitet, mener Børge Nordestgaard.

Hos Kræftens Bekæmpelse mener Inge Haunstrup Clemmensen dog også, at en mulig forklaring kan være, at mennesker, der har råd og tid til at være meget i solen, ofte vil dyrke udendørs sport eller motion – og at de på den måde styrker helbredet.

Artiklen blev dagen efter kommenteret af Kræftens Bekæmpelse, og gav i det hele taget anledning til en ganske ophedet debat. Og så 10 dage efter tager sagen en dramatisk vending, da to statistikere fra afdelingen for Biostat på Københavns Universitet retter en sønderlemmende kritik mod den statistiske metode, og ikke nok med det: De sender deres indvendinger til samme tidsskrift, hvor artiklen netop er blevet offentliggjort.

DANSK SOLARIE FORENING

Dansk forskning: Soldyrkere lever meget længere

Danske forskere vil offentliggøre ny undersøgelse der viser, at danskere der kan defineres som *soldyrkere* lever seks år længere.

Foruden den længere levetid viser undersøgelsen også, at soldyrkere har en lavere forekomst af både blodpropper i hjertet og knogleskørhed.

Stadig sol med omtanke

- Vi har de seneste år set en vis grad af solforskrækkelse, og det er efter min mening at gå i den anden grøft, udtaler professor Børge Nordestgaard, der er en af forskerne bag den nye undersøgelse til Politiken, og han understreger i den forbindelse, at man stadig skal solbade med omtanke og undgå solskoldninger, ikke mindst hos børn og unge.

Kræftens Bekæmpelse ikke overbevist

Selvom den nye forskning omfatter 4,4 millioner danskere og i øvrigt

Ikke overraskende blandede Dansk Solarieforening sig med nogle stærke udfald mod Kæftens Bekæmpelse >

Øvelse 3.2

Læs artiklen og præsenter informationerne heri med anvendelse af de statistiske begreber, du har lært:

a) Hvad er problemet i forholdet mellem population og stikprøve, sådan som den oprindelige forskergruppe har behandlet det?

b) Hvad går den centrale indvending mht. gennemsnitlig levealder for de to grupper ud på?

3. ARTIKEL OM SOL OG HUDKRÆFT

25. OKT. 2013

Solstrid brudt ud: Forskere strides om levealder

To danske forskergrupper strides om leveralder for folk med hudkræft.

Henrik Larsen

Lever såkaldte soldyrkere, som har fået ikkedødelig hudkræft, i gennemsnit seks år længere end resten af den danske befolkning? Det spørgsmål bliver nu genstand for en strid mellem to danske forskergrupper.

Den ene gruppe, under ledelse af professor Børge Nordestgaard fra Herlev Hospital og Københavns Universitet, siger ja. Den anden gruppe, professor Niels Keiding og lektor Theis Lange, begge fra Afdeling for Biostatistik ved Københavns Universitet, siger nej.

Ikke muligt at bevise sammenhæng

Stridens kerne er en videnskabelig artikel, som Børge Nordestgaard sammen med tre danske kolleger har publiceret i det meget anerkendte videnskabstidsskrift International Journal of Epidemiology. En videnskabelig undersøgelse, som Politiken skrev om i sidste uge.

Den videnskabelige undersøgelse bygger på helbredsrelaterede oplysninger om 4,4 millioner danskere mellem 40-100 år i perioden 1980-2006. På basis af undersøgelsen konkluderer Børge Nordestgaard og hans kolleger blandt andet, at mennesker, der har fået ikkedødelig hudkræft – en lidelse, der som regel skyldes, at en person har været udsat for store mængder sollys – i gennemsnit lever seks år længere end resten af befolkningen. Og bliver 86 år, hvor gennemsnitslevetiden for resten af befolkningen er 80 år. I undersøgelsen gør de dog klart opmærksom på, at det ikke er muligt at påvise nogen årsagssammenhæng mellem sollys og levealder.

Personer har i forvejen høj alder

Metoden bag undersøgelsens afsnit om forskellen i levealder på seks år mellem de to grupper holder imidlertid ikke, mener Niels Keiding og Theis Lange, der har bedt International Journal of Epidemiology om at trykke deres argumenter for, at Nordestgaard-gruppen har brugt *en analysemetode, der er absolut forkert*, siger Theis Lange: *Problemet med deres analyse er, at for at komme ind i hudkræftgruppen skal man – naturligvis – have levet længe nok til at udvikle hudkræft – og de fleste, der diagnosticeres med ikkedødelig hudkræft, er over 50 år. For gruppen af personer uden hudkræft forholder det sig lige omvendt, her kan personer af alle aldre indgå. Konsekvensen bliver, at personer i gruppen med hudkræft dør i en højere alder end den anden gruppe, men dette har absolut intet at gøre med deres diagnose. Det er alene fordi gruppen med hudkræft er udvalgt på en måde, så den hovedsageligt indeholder personer, der i forvejen har en høj alder*, siger Theis Lange.

Forskere er klar til at svare på kritikken

Børge Nordestgaard siger, at han og de tre medforfattere ser frem til at modtage Keiding og Langes henvendelse fra International Journal of Epidemiology: *Når det sker, vil vi svare detaljeret på deres kommentarer og kritik – sådan er proceduren. Vi er glade for den store interesse og medieomtale, undersøgelsen har affødt – det viser, at dette emne er vigtigt. Og vi er glade for de kommentarer, vi har modtaget, både de konstruktive og de kritiske, som peger på, hvordan en sådan undersøgelse kan laves endnu bedre*, siger professor Børge Nordestgaard, og tilføjer: *Man må håbe, at andre forskere gennem nye undersøgelser vil vise, om solskin kan have andre positive effekter på livslængde og helbred*. International Journal of Epidemiology oplyser, at man har modtaget og nu vil studere henvendelsen fra Niels Keiding og Theis Lange.

Øvelse 3.3

Antag vi har et land, hvor den demografiske struktur er meget stabil over tid, og fordeler sig således, hvis vi fokuserer på levealder:

Andel af befolkningen	Gennemsnitlig levealder
15%	under 45 år
10%	45-55
20%	55-65
30%	65-75
20%	75-85
5%	85-95

For nyfødte børn gælder altså, at 20% af dem bliver mellem 75 og 85, mens 15% vil dø inden de bliver 45.

a) Hvad skal vi forstå ved gennemsnitlig levealder? Udregn denne!
b) Hvad er den gennemsnitlige levealder for dem, der er blevet 55? Og for dem der er blevet 75?

Øvelse 3.4
Antag du i morgenavisen læser følgende overskrift på en artikel: *Paver og biskopper lever længere end almindelige præster*. Hvad vil din umiddelbare kommentar være?

Sagen fik også en dramatisk afslutning, idet tidsskriftet gav statistikerne fra Biostat ret. Det skete ikke i form af en lille notits, men via hele to artikler, nemlig én artikel, der var indsendt uden viden om den danske kontrovers, og som var en generel analyse af den type fejl, den oprindelige undersøgelse var behæftet med, og hvor redaktøren derfor overtalte forfatterne til at inkludere det danske eksempel som et typisk eksempel på fejlslutninger. Samt en artikel skrevet af redaktøren selv, og som du ser første side af her.

Øvelse 3.5
Læs redaktørens artikel og giv et sammendrag af hans vurdering af den videnskabelige kvalitet af den oprindelige artikel.

Studieretningsprojekt eller anden for projektarbejde
Hele forløbet, som det er beskrevet ovenfor, inklusiv de forskellige øvelser, samt materialet i mappen du kan tilgå via hjemmesiden, hvor du blandt finder de originale tidskriftartikler, kan danne grundlag for firskellige typer af studieretningsprojekter. Man kan koncentrere sig om substansen vedr. soldyrkning og kræft, og hvorledes man med statistiske metoder kan svare på opstillede spørgsmål. Man kan også vælge at lægge vægten på de videnskabsteoretiske sider, der blev afdækket i diskussionen.

International Journal of Epidemiology, 2014, 639–644
doi: 10.1093/ije/dyu108
Editor's choice

Editor's choice

Sun exposure and longevity: a blunder involving immortal time

Jane E Ferrie* and Shah Ebrahim

*Corresponding author. E-mail: jane.ferrie@bristol.ac.uk

Unfortunately we have to start this Editor's Choice with an acknowledgment that we have fallen prey to a common, perennial problem; immortal time bias.

To illustrate the concept we borrow an example from William Farr, as used by James Hanley and Bethany Foster in a full and entertaining exposition of the problem in this issue of the journal.[1] Generals and bishops live longer than corporals and curates—but this is not necessarily because an elevated occupational status makes you live longer—it may simply be because you have to reach a certain age before it is possible to hold such positions. People become generals and bishops in middle age so their deaths arise after this point in time, whereas corporals and curates can die at any age above 20 or so.[2] This difference in time during which an event can occur to one group but not the other produces a bias favouring longer life expectancy—immortal time bias. In the figure on the next page, the problem is evident at a glance (Figure 1).[3]

In the October issue of the *International Journal of Epidemiology* (*IJE*) last year, we published a paper by Peter Brøndum-Jacobsen and colleagues in which they examined the effects of sunlight exposure on mortality among the whole population of Denmark aged above 40 years, using linked data from national registries.[4] They used non-melanoma skin cancer as a proxy for sun exposure, which is a clever idea but it should have been obvious that the findings were 'too good to be true'—an apparent halving of all-cause mortality and reductions in myocardial infarction and hip fracture. The authors concluded: 'Causal conclusions cannot be made from our data. A beneficial effect of sun exposure per se needs to be examined in other studies'.

The Danish media picked up the story and it became front page news—'Sunbathers live longer'.[5] Although the authors never made this claim in their published paper, their interviews with the press did not appear to emphasize their non-causal conclusion. The Danish Cancer Association claims that this paper has undone all their good work in persuading Danes to keep out of the sun to avoid skin cancers.

Commentators on the story identified a likely problem of immortal time bias. People in the 'sun exposure' group had to live long enough to be diagnosed with skin cancer but the comparison group only had to be over 40 years old—the design of the study had built in a potential bias in favour of longevity among those presumed to be more highly exposed to sunlight. Theis Lange and Neils Keiding, in a letter commenting on the paper, pose questions about how such highly improbable findings got through the editorial process at *IJE*.[6]

In response to this criticism, Brøndum-Jacobsen and colleagues argue that their paper used both cohort and case-control analyses, and that the latter should be free from immortal time bias as cases and controls were matched on age.[7] They acknowledge that the case-control analyses—which showed much smaller survival advantage [odds ratio (OR): 0.97, 95% confidence interval (CI) 0.96 to 0.99; vs hazard ratio (HR): 0.52, 95% CI 0.52 to 0.53] —should have been included in their abstract. In addition, they conducted a revised Cox proportional hazards analysis stratified by 10-year, 5-year and 2-year age strata in an attempt to control for immortal time bias, and interpret these findings as similar to those in their original paper. However, they fail to stress that the effect sizes become increasingly attenuated as the age matching becomes more exact, suggesting that the apparent effect of sun exposure may indeed be produced by immortal time bias.

Ironically, in parallel with the review and publication of this paper we had commissioned an 'Education Corner'

4. Dataanalyse baseret på en stikprøve - Walds problem med nedskydning af kampfly (B og A)

Dette afsnit er baseret på: Jordan Ellenberg: How not to be wrong - The hidden maths of everyday life, kapitel 1. Eksemplet kan foldes ud til et større projekt ved at inddrage artikler af og om Walds arbejde, som der linkes til i slutningen af afsnittet

Lærervejledning

Det følgende eksempel stammer fra anden verdenskrig, hvor Amerikanerne bl.a. blev involveret i luftkrig mod tyskerne, og hvor amerikanerne led mange nederlag i form af nedskudte fly.

Det amerikanske militær bad da den til Amerika flygtede matematiker Abraham Wald om hjælp til at forstærke armeringen af deres kampfly. Der var flere problemstillinger Wald skulle tage stilling til: Dels hvor meget armering man kunne give kampflyene uden at hæmme deres manøvredygtighed for meget og bruge for meget brændstof, dels hvilke dele af kampflyet, der var særligt udsat for tysk beskydning og derfor havde behov for særlig beskyttelse: Så hvor på kampflyet skulle man ofre særlig tyk armering, og hvor kunne man nøjes med tyndere armering?

Til støtte for Walds analyser havde det amerikanske militær fremskaffet data om skudhuller, for de amerikanske kampfly, der vendte tilbage til baserne:

Udsnit af flyet	**Antal skudhuller pr kvadratfod**
Motoren	1.11
Skroget	1.73
Brændstoftankene	1.55
Resten af flyet (vinger mm)	1.8

Så her har vi autentiske data fra anden verdenskrig, som kan præsenteres for klassen. Derefter kan klassen deles op i grupper, fx parvis, og diskutere, hvad ville de anbefale vedrørende armeringen af flyene: Hvilken del af flyet ville de ofre mest armering på.

1. del: Data fremlægges og diskuteres

Det er vigtigt, at det fører til en åben diskussion grupperne indbyrdes, så man kan samle op på gruppernes forslag til sidst og høre de forskellige overvejelser, de måtte have gjort og de forskellige anbefalinger, det måtte føre til. Hvis der er konsensus om et af forslagene kan man notere det, men på dette stadie gælder det om at holde alle muligheder åbne og ikke 'røbe' det 'rigtige svar'. Af samme grund er det også vigtigt at finde ud af, om der er elever, der kender problemstillingen på forhånd, for så kan de ikke deltage i diskussionen.

2. del: Datastrukturen analyseres nøjere

I forlængelse af Susanne Ditlevsens video, skal vi nu nøje overveje strukturen af de indsamlede data, og udnytte dette til en mere præcis dataanalyse. Denne del kan evt. føres som en samlet diskussion med klassen, eller grupperne kan igen blive bedt om at overveje fx de følgende spørgsmål:

a) Hvad er **stikprøven** og hvad er **populationen** i de data som Wald fik præsenteret? Hvorfor udgør de ikke data, for alle de kampfly, som Amerikanerne sendte mod tyskerne, og hvad er det særlige kendetegn ved de kampfly, der er med i stikprøven?
b) Hvorfor er stikprøven **ikke** repræsentativ? Hvis den skulle have været repræsentativ, hvilke data skulle man så yderligere have indsamlet?
c) Hvordan kunne en **nulhypotese** om skudhullernes fordeling se ud? Luftkampe mellem amerikanske kampfly og tyske kampfly er ret kaotiske og der er ikke tid til præcisionsindstilling af maskingeværerne! Hvilken fordeling følger skudhullerne ifølge nulhypotesen?
d) Hvorfor afviger skudhullernes observerede fordeling i stikprøven sig fra den forventede fordeling?
e) Hvad fortæller det om skudhullernes fordeling på de fly, der blev skudt ned af tyskerne?
f) Hvad fortæller det om sårbarheden af de forskellige dele på et fly? Hvilken anbefaling ville du nu give det amerikanske militær, vedrørende armeringen af de amerikanske kampfly: Hvor på flyet skal de ofre særligt meget armering?

3. del: Videre arbejde med problemstillingen

I de foregående afsnit har vi undersøgt Walds problemstilling med elementære metoder fra især den deskriptive statistik. Det gør projektet tilgængeligt allerede på C-niveau. Men hvis man arbejder med projektet på A-niveau og gerne vil forankre det i en sandsynlighedsteoretisk ramme, er der flere muligheder for at uddybe projektet.

Dels kan man læse Walds originale rapport: *A method of estimating plane vulnerability based upon damage of survivors.* Du kan hente denne rapport her.
Man kan da arbejde med udvalgte dele af artiklen og prøve at forklare den teori, der ligger bag ved hans beregninger.

Dels kan man læse en moderne statistisk analyse af Walds originale rapport: *Abraham Walds work on aircraft vulnerability.* Du kan hente denne rapport her.
Igen kan man arbejde med udvalgte dele af artiklen og prøve at forklare den teori, der ligger bagved.

5. Er det usundt at ryge? (B og A)

Denne aktivitet, der kan anvendes i et arbejde med hypotesetest med brug af χ^2 – fordelingen, er hentet fra Hvad er Matematik, C bogen kapitel 9, afsnit 6.

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.

Observerede værdier

		Rygevaner		
		ja	nej	I alt
Helbreds-tilstand	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 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 bestemt afgrænset område af England. Der er også mange andre variable, der ikke er taget højde for.

Øvelse 9.26
Nævn tre andre variable, der kunne have indflydelse på undersøgelsens resultat.

Hvis nogle af de variable, der er kommet frem i øvelse 9.26, 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 χ2-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		
		ja	nej	I alt
Helbreds-tilstand	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: 9.27
Gennemfør udregningen af de forventede værdier i den ovenstående tabel i detaljer.

χ^2-teststørrelsen udregnes igen 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. Vi har tidligere omtalt, at dette teoretisk betyder, 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.

Eksperimentel 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 9.28

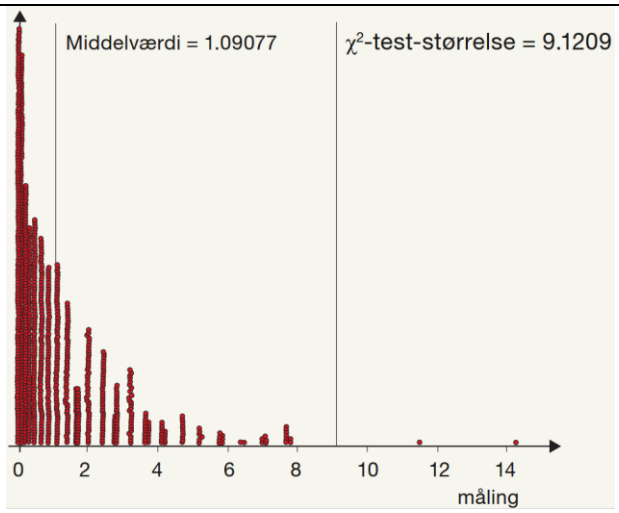
En sådan simulering (omrøring) kan gennemføres i et værktøjsprogram: Den ene variabel holdes fast, mens den anden blandes vilkårligt rundt, og resultatet samles i en ny antalstabel.

a) Gennemfør en sådan omrøring, eller gå ind på hjemmesiden, og benyt den animation, der ligger der.

b) Opstil formlen 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.

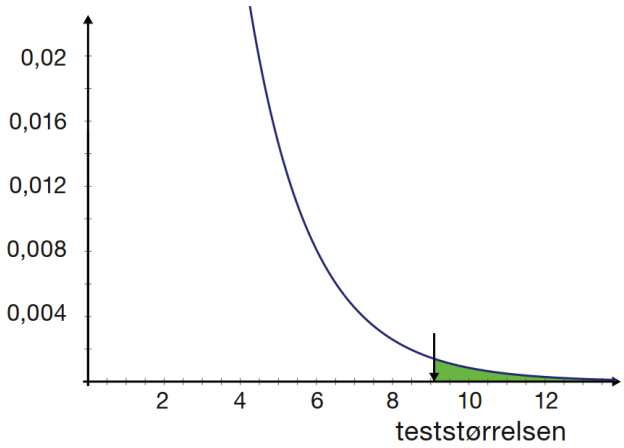
Formelbaseret metode:

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 med 1 frihedsgrad.

Værktøjsprogrammet giver:

χ^2Cdf(9.1209,∞,1) = 0.002527



Øvelse 9.29

a) Benyt dit værktøjsprogram til at finde p-værdien såvel grafisk som ved beregning ud fra den kumulerede χ2-fordeling.

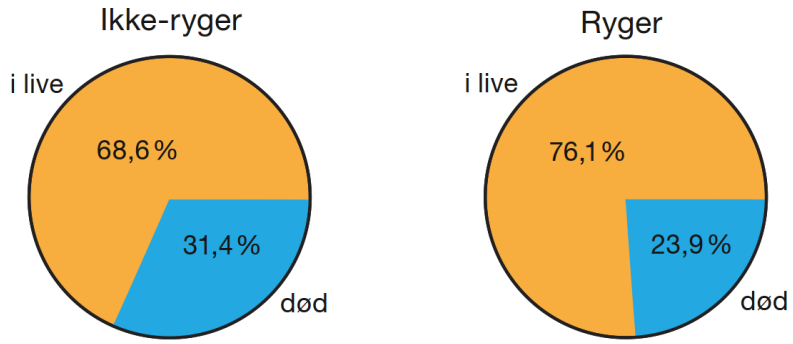
b) Hvor lille skal teststørrelsen være, for at vi ikke længere kan forkaste nulhypotesen?

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.

5.1 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, som vi omtalte i begyndelsen af afsnittet. 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 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 9.30

a) Kopier tabellerne ind i et værktøjsprogram, og udregn række- og søjlesummerne.
b) Udregn overlevelsesprocenterne for rygere og ikke-rygere i de tre aldersgrupper.
c) Illustrer resultatet grafisk.
d) 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 Hvad er Matematik? C, i-bogen kan du dels læse en kommentar til undersøgelsen, der inddrager Simpsons paradoks, dels finde et uddybende materiale om Simpsons paradoks. Der findes også mere materiale om Simpsons paradoks i afsnit 6 om racefordomme 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.

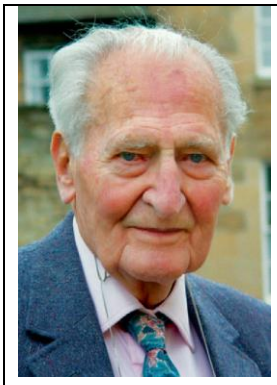
Hvis man er i samarbejde med et andet fag, kan der muligvis ud fra dette fags viden peges på en mekanisme, der kan forklare påvirkningen fra den ene variabel til den anden. Men også dette kan vise sig at være yderst vanskeligt. Havde vi fx ikke haft tabellerne med aldersfordelingen, kunne vi jo ikke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

5.2 The Mortality of Doctors

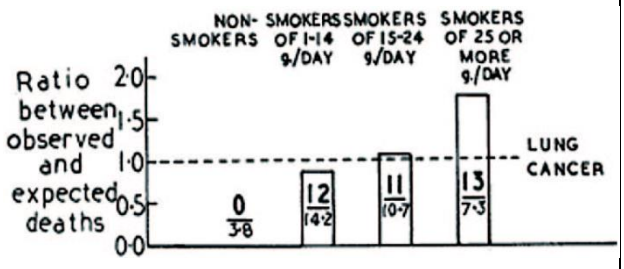
I komplicerede situationer kan det derfor vise sig meget svært at løfte bevisbyrden. Her er sammenhængen mellem rygning og helbred et klassisk eksempel på, hvor svært det kan være.

De første af en lang kæde af indicier på en mulig sammenhæng imellem rygning og helbredblev fundet i midten af 50'erne af et engelsk forskerteam under ledelse af epidemiologi-eksperten Richard Doll. I en banebrydende artikel fra 1954: *"The mortality of doctors in relation to their smoking habits"*, offentliggjort i det anerkendte fagtidsskrift British Medical Journal, påviste de for første gang en ret klar sammenhæng mellem rygning og lungekræft.

Undersøgelsen forløb over to et halvt år og involverede 40.000 læger. Ved starten af undersøgelsen registrerede man deres rygevaner, og ved udløbet af undersøgelsen registrerede man samtlige dødsfald og deres årsag i perioden. Af de 40.000 læger døde 723 i perioden – heraf døde 36 af lungekræft. Alle der døde af lungekræft, var rygere. Ved at sammenholde testpersonernes rygevaner med deres dødelighed for lungekræft så man nu en relativ klar sammenhæng mellem rygevaner og dødelighed.



Richard Doll, engelsk ekspert i epidemiologi, der undersøgte sammenhængen mellem rygning og helbredstilstand. Hans håndtegnede diagram illustrerer forholdet ("ratio") mellem antal observerede og antal forventede døde i forskellige grupper af rygere. Hvis forholdet fx er 1,5 betyder det, at der er 1,5 gange flere døde, end forventet.



Øvelse 9.31

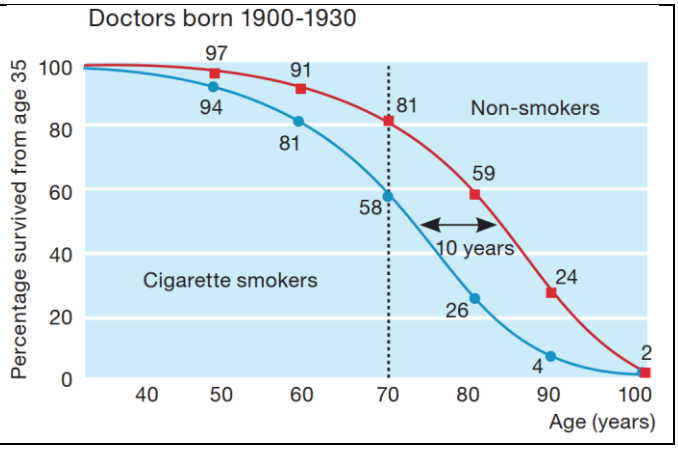
Gennemgå den håndtegnede graf fra 1954. Hvorfor antyder den en sammenhæng mellem rygevaner og lungekræft. Hvorfor er det centralt, at rygerne er yderligere kategoriserede efter deres rygevaner?

Artiklen blev taget som et indicium for en mulig sammenhæng mellem rygning og helbred. Dolls team havde taget mange forholdsregler for at undgå skjulte variable. Fx havde de sikret sig, at alle dødsfald i undersøgelsesgruppen kom med, og at dødsårsagen var så objektiv som mulig, idet den blev uddraget direkte fra dødsattesten. De sikrede sig også mod mulige fejldiagnoser (måske undersøger man ikke dødsårsagen grundigt nok og skriver bare lungekræft på dødsattesten, fordi det er så oplagt en dødsårsag for rygere). Men så skulle andre dødsårsager være underrepræsenterede, hvilket de kunne påvise ikke var tilfældet.

Øvelse 9.32

Ikke alle statistikere lod sig overbevise. *I Hvad er Matematik? C, i-bogen kan du du finde* en artikel af Ronald Fisher, en af det 20. århundredes største statistikere, hvor han argumenterer imod en sammenkædning af rygning og helbred. Fisher var selv storryger.

I en artikel fra 2004, *Mortality in relation to smoking: 50 years' observation on male British doctors*, kunne Doll i stor detalje kortlægge de komplekse sammenhænge mellem rygning og elbred, herunder fx at rygning fra en ung alder alt andet lige forkorter den samlede levetid med 10 år.



Øvelse 9.33

a) Oversæt begreberne, og forklar hvordan de to grafer er tegnet.
b) Forklar, hvad det er, der måles med den lodrette stiplede linje.
c) Forklar, hvad der menes med den vandrette linje hvor der står '10 years'.
d) Hvordan vil du med ord og ud fra graferne beskrive sammenhængen mellem rygning og levetid.

6. Projekt: Racefordomme i USA og Simpsons paradoks (B og A)

(*Dette kapitel er en redigeret udgave af projekt 9.9 i B-bogen: Racefordomme og Simpsons paradoks. Data er hentet fra M. Radelet, "Racial characteristics and imposition of death penalty", American Sociological Review, 46 (1981), pp 918-927*)

I dette projekt vil vi undersøge *racefordomme* i USA: Bliver de sorte diskrimineret i forhold til de hvide? Fx har det været påstået at retssystemet ikke er så farveblindt som det måske burde være. For at undersøge dette har man kigget på dødstallene for 326 retssager, hvor den anklagede risikerede dødsstraf.

Øvelse 6.1. Den anklagedes hudfarve
Den følgende krydstabel viser sammenhængen mellem den anklagedes hudfarve og den dom, der blev fældet i retssagen:

Status\Anklaget	Hvid	Sort
Dømt til døden	19	17
Frifundet	141	149

a) Udbyg tabellen med totalværdier, og oversæt også tabellen til en procenttabel, der viser hvor stor en andel af de anklagede, der dømmes til døden, den såkaldte *dødsrisiko*.
b) Afbild tabellen i passende diagram og skriv en foreløbig konklusion som svar på spørgsmålet: Tyder data fra de amerikanske retssager på at sorte bliver diskrimineret i forhold til hvide?

Øvelse 6.2. Ofrets hudfarve
I denne øvelse inddrager vi endnu en variabel, nemlig ofrets hudfarve. Måske er juryerne påvirket af ofrets hudfarve i deres domfældelse? Det ville jo være en lige så klar racediskrimination. Datamaterialet opdeles derfor yderligere efter ofrets hudfarve. Det giver anledning til de følgende to deltabeller:

Ofret var hvid		
\Anklaget	Hvid	Sort
Dømt til døden	19	11
Frifundet	132	52

Ofret var sort		
Status\Anklaget	Hvid	Sort
Dømt til døden	0	6
Frifundet	9	97

a) Gennemfør nu den samme beregninger som i første øvelse for hver af de to krydstabeller: dvs. undersøg dødsrisikoerne i de forskellige tilfælde
b) Skriv igen en ny foreløbig konklusion på spørgsmålet: Tyder data fra de amerikanske retssager på at sorte bliver diskrimineret i forhold til hvide?

Overvej om der er overensstemmelse mellem de to konklusioner fra første og anden akt? Hvis ikke, hvad kan da være grunden til at du når til to forskellige konklusioner?

6. 1 Simpsons paradoks

I det foregående skulle du gerne have set et eksempel på at en konklusion kan vendes, når man slår to deltabeller sammen til en større tabel, dvs. når man ser bort fra en tredje variabels indflydelse på sagen – den såkaldte skjulte variabel. I denne sammenhæng betyder et paradoks at man er overrasket over noget, at resultatet af undersøgelsen strider mod ens umiddelbare forventning: Slår man to delundersøgelser sammen til en større undersøgelse burde det jo ikke ændre på konklusionen.

I første øvelse tog vi kun hensyn til den anklagedes hudfarve ud fra en forventning om at den var afgørende for dødsrisikoen. Det viste sig da at de hvide anklagede faktisk havde en højere dødsrisiko end de sorte anklagede, om end forskellen ikke var så markant. Men for at være sikker på konklusionen er vi nødt til at forudsætte 'alt andet lige' princippet. Der kunne jo være andre faktorer, der havde indflydelse på dødsrisikoen, skjulte variable som vi ikke har inddraget i undersøgelsen. Men hvis vi har sørget for at sammensætningen for de øvrige variable er den samme i de to typer retssager, dem hvor det er en hvid, der er anklaget og dem, hvor det er en sort, der er anklaget, så burde virkningen af disse skjulte variable være den samme i begge tilfælde, og en eventuel forskel burde derfor kunne tilskrives den anklagedes hudfarve. Konklusionen holder altså kun hvis sammensætningen af de to typer retssager alt andet lige er den samme for alle andre variable, som vi ikke har taget hensyn til (dvs. vi har udført variabelkontrol og holdt alle andre variable på samme niveau i de to typer).

Skulle det mod forventning vise sig at sammensætningen af de to typer retssager faktisk er meget forskellige med hensyn til en tredje skjult variabel, så står vi derimod meget dårligt i vores konklusion, for så kunne forskellen i domfældelserne jo lige så godt skyldes ændringen af den skjulte variabel.

I den anden øvelse inddrog vi netop en sådan skjult variabel, ofrets hudfarve, og nu viser der sig pludselig en markant forskel i sammensætningen af de to typer retssager: Der er stort set ingen sorte ofre i de sager, hvor de hvide er på anklagebænken. Hvorfor det er sådan kan også i sig selv vække grund til bekymring: Er det fx sådan at der bare ikke bliver rejst sag i de tilfælde, hvor en sort overfaldes af en hvid? Men det vil vi ikke se nærmere på her. Her holder vi os til data, og når vi inkluderer den skjulte variabel, så tyder data pludselig på at de sortes dødsrisiko er markant større end de hvides.

Hvordan kan det nu være at en sådan skjult variabel kan vende billedet? For at forstå hvordan paradokset kan opstå kan det være en fordel at indføre en simpel model til at forklare hvad der foregår. Vi vil da give to forskellige modeller, en simpel, der viser hvordan man kan konstruere paradokset, og en lidt mere detaljeret, hvor vi både forsøger at forstå oprindelsen til paradokset og få et endeligt svar på spørgsmålet: *Tyder data fra de amerikanske retssager på at sorte bliver diskrimineret i forhold til hvide?*

Fejlslutningen

Vi vender tilbage til tabellerne hvor vi har yderligere opdelt retssagerne efter ofrenes hudfarve

Ofret var hvid			
Status\Anklaget	Hvid	Sort	I alt
Dømt til døden	19	11	30
Frifundet	132	52	184
I alt	151	63	214

Ofret var sort			
Status\Anklaget	Hvid	Sort	I alt
Dømt til døden	0	6	6
Frifundet	9	97	106
I alt	9	103	112

Hvidt offer: I følge den første tabel er dødsrisikoen for hvid givet ved 19/151 = 12.6%, mens den for sort er givet ved 11/63 = 17.5%, dvs. sorts dødsrisiko er størst:

$$p_{\text{sort}\,|\,\text{ofret er hvid}} = \frac{11}{63} > \frac{19}{151} = p_{\text{hvid}\,|\,\text{ofret er hvid}}$$

Sort offer: Ifølge den anden tabel er dødsrisikoen for hvid givet ved 0/9 = 0%, mens den for sort er givet ved 6/103 = 5.8%, dvs. igen er sorts dødsrisiko størst:

$$p_{\text{sort}\,|\,\text{ofret er sort}} = \frac{6}{103} > \frac{0}{9} = p_{\text{hvid}\,|\,\text{ofret er sort}}$$

Men hvorfor kan vi så ikke slutte at det samme så også må gælde for den kombinerede dødsrisiko? Hvis vi fx lagde brøkerne sammen er det jo klart at summen af de to største brøker er større end summen af de to mindste brøker. Men vi lægger ikke brøkerne sammen, når vi kombinerer de to tabeller: Vi lægger tællerne sammen og nævnerne sammen – og det er noget helt andet:

Status\Anklaget	Hvid	Sort	I alt
Dømt til døden	19 = 19 + 0	17 = 11 + 6	36 = 30 + 6
Frifundet	141 = 132 + 9	149 = 52 + 97	290 = 184 + 106
I alt	160 = 151 + 9	166 = 63 + 163	326 = 214 + 112

$$p_{hvid} = \frac{19+0}{151+9} = \frac{19}{160} > \frac{17}{166} = \frac{11+6}{63+163} = p_{sort}$$

Øvelse 6.3. Geometrisk illustration

Dødsrisikoen, dvs. brøken $\frac{\text{antal dømte}}{\text{antal anklagede}}$ kan opfattes som hældningen for den rette linje gennem *Origo* (0,0) og (antal anklagede, antal dømte).

a) Afsæt i et koordinatsystem antal anklagede ud af førsteaksen og antal dømte op af andenaksen. Afsæt heri punkterne hørende til $Hvid_{\text{samlet}}$ og $Sort_{\text{samlet}}$ fra tabellen nedenfor, dvs afsæt punkterne (160,19) og (166,17)

Status\Anklaget	Hvid	Sort
Dømt til døden	19	17
Frifundet	141	149
I alt	160	166

b) Hvad er hældningerne for de linjestykkerne, der forbinder *Origo* (0,0) med de afsatte punkter?

c) Gør det samme med de to deltabeller, hvor den første giver anledning til punkterne $Hvid_{\text{hvidt offer}}$ og $Hvid_{\text{sort offer}}$, mens den anden giver anledning til punkterne $Sort_{\text{hvidt offer}}$ og $Sort_{\text{sort offer}}$.

d) Hvilken figur er knyttet til de fire punkter *Origo*, $Hvid_{\text{hvidt offer}}$, $Hvid_{\text{sort offer}}$ og $Hvid_{\text{samlet}}$? træk figuren op og farv det indre lysegråt (for hvid).

e) Samme spørgsmål til de sorte punkter, hvor figuren farves mørkegrå (for sort)!

f) Hvad fortæller ulighederne $p_{\text{sort}\mid\text{ofret er hvid}} = \frac{11}{63} > \frac{19}{151} = p_{\text{hvid}\mid\text{ofret er hvid}}$

og $p_{\text{sort}\mid\text{ofret er sort}} = \frac{6}{103} > \frac{0}{9} = p_{\text{hvid}\mid\text{ofret er sort}}$

dig om den 'hvide' og den 'sorte' figur?

g) Prøv nu at forklar med dine egen ord, i hvilken forstand figuren fremstiller Simpsons paradoks.

Bemærkning: Hvis du i stedet opretter de 6 punkter $Hvid_{\text{hvidt offer}}$, $Hvid_{\text{sort offer}}$ og $Hvid_{\text{samlet}}$, $Sort_{\text{hvidt offer}}$, $Sort_{\text{sort offer}}$ og $Sort_{\text{samlet}}$ som frie gitterpunkter med heltallige koordinater har du nu en 'maskine' der hurtigt og nemt kan frembringe andre eksempler på antalstabeller, der illustrerer Simpsons paradoks.

Den skjulte variables rolle

Vi vil nu endelig prøve at forstå den skjulte variabels rolle. Det sker igen ved hjælp af en simpel model. Den skjulte variabel handler om ofrenes sammensætning ved de involverede forbrydelser. Det er nemmest at indføre en enkelt variabel *x*, der netop repræsenterer den brøkdel som de hvide ofre udgør. Hvis der slet ingen hvide ofre er har den værdien 0, hvis der er lige mange hvide ofre og sorte ofre har den værdien ½, hvis der kun er hvide ofre har den værdien 1. Variablen *x* antager altså værdier mellem 0 og 1.

Øvelse 6.4

a) Hvilken værdi har variablen *x* i tilfældet med de hvide anklagede? Denne værdi kaldes x_{hvid}.

b) Hvilken værdi har variablen *x* i tilfældet med de sorte anklagede? Denne værdi kaldes x_{sort}.

Dernæst har vi dødsrisikoen, som vi vil betegne med p. Den afhænger af ofrenes sammensætning. Hvis vi går ud fra de to deltabeller:

Ofret var hvid			
Status\Anklaget	Hvid	Sort	I alt
Dømt til døden	19	11	30
Frifundet	132	52	184
I alt	151	63	214

Ofret var sort			
Status\Anklaget	Hvid	Sort	I alt
Dømt til døden	0	6	6
Frifundet	9	97	106
I alt	9	103	112

kan vi omstrukturere disse til en ny krydstabel, der viser forholdene for de hvide anklagede.

Retssager med en hvid på anklagebænken			
	Hvidt offer	Sort offer	I alt
Dømt til døden	19	0	19
Frifundet	132	9	141
I alt	151	9	160

Vi ser da at dødsrisikoen er 0/9, hvis der kun er sorte ofre svarende til $x = 0$ og tilsvarende er den 19/151, hvis der kun er hvide ofre svarende til x = 1. Vi kan altså udfylde den følgende tabel over sammenhængen mellem offerbrøken *x* og dødsrisikoen *p*:

x = offerbrøken	0 = startværdi (ingen hvide)	1 = slutværdi (kun hvide)
p = dødsrisikoen	0/9	19/151

Øvelse 6.5. Sammenligning af dødsrisiko for hvid og sort

a) Tegn den lineære sammenhæng, der hører til denne tabel i et koordinatsystem, hvor du afsætter offerbrøken x (den skjulte variabel!) ud af første aksen og dødsrisikoen p op af andenaksen. Afsæt også den faktiske offerbrøk $x_{hvid} = 151/160$ på x-aksen og find den tilhørende p-værdi grafisk. Kontroller at den stemmer overens med dine tidligere fundne resultater.

b) Gentag derefter den samme øvelse med de retssager, hvor der er en sort på anklagebænken.

c) Du har nu frembragt to lineære sammenhænge i det samme koordinatsystem, hvoraf den ene viser hvordan dødsrisikoen for en hvid anklaget afhænger af offerbrøken, mens den anden viser hvordan dødsrisikoen for en sort anklaget afhænger af offerbrøken.

d) Brug diagrammet til at forklare Simpsons paradoks: Hvordan er det muligt at hvids dødsrisiko kan ligge højere end sorts dødsrisiko på trods af at hvids graf ligger under sorts graf?

e) Hvis du korrigerer for de forskellige offerbrøker kan du nu også besvare det følgende spørgsmål ud fra grafen: Hvor meget højere er dødsrisikoen for en sort anklaget end dødsrisikoen for en hvid anklaget – hvis alt andet er lige?

7. Case om skjulte variable: Optagelsestallene fra Berkeley (B og A).

Hvis man ikke har tid til et projekt som beskrevet i afsnit 6, så kan man vælge at gennemgå denne lille case, der har samme grundlæggende pointe vedr. skjulte variable og Simpsons paradoks.

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.

Øvelse 7.1

Ud af 2691 mandlige ansøgere blev de 1198 optaget. Ud af 1835 kvindelige ansøgere blev 557 optaget.

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

b) Undersøg om der er belæg i tabellen for at rejse anklage om kønsdiskrimination.

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

Øvelse 7.2

a) Undersøg nu igen sammenhængen mellem køn og optagelsesstatus inden for hvert af de seks hovedområder.

b) Er der stadigvæk belæg for påstanden om kønsdiskrimination i optagelsesproceduren?

Konklusionen fra den oprindelige artikel om Berkeley sagen

Examination of aggregate data on graduate admissions to the University of California, Berkeley, for fall 1973 shows a clear but misleading pattern of bias against female applicants. Examination of the disaggregated data reveals few decision-making units that show statistically significant departures from expected frequencies of female admissions, and about as many units appear to favor women as to favor men. If the data are properly pooled, taking into account the autonomy of departmental decision making, thus correcting for the tendency of women to apply to graduate departments that are more difficult for applicants of either sex to enter, there is a small but statistically significant bias in favor of women. The graduate departments that are easier to enter tend to be those that require more mathematics in the undergraduate preparatory curriculum. The bias in the aggregated data stems not from any pattern of discrimination on the part of admissions committees, which seem quite fair on the whole, but apparently from prior screening at earlier levels of the educational system. Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects.

Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975) Sex bias in graduate admissions: Data from Berkeley. Science, 187, 398–403.

Øvelse 7.3

Hvad siger de til problemstillingen?

8. Testet positiv – men er man syg? (B og A)

Det tredje indledende eksempel fra Susanne Ditlevsens film, hvor substansen er test af om man er syg, kan anvendes i et forløb, der giver en indføring i begreber og metoder, der knytter sig til emnet betingede sandsynligheder. Vælger man at gennemføre et egentligt forløb om betingede sandsynligheder, kan materialet her indgå som en slags paradigmatiske eksempler.

I Susanne Ditlevsens film præsenteres en case, hvor der er udviklet en bestemt test til at afsløre om man har en given sygdom. Den nye metode er åbenbart rigtig god, idet den faktisk fanger langt hovedparten af de der er syge.

Antalstabeller kan hjælpe med til at give overblik over nogle oplysninger og spørgsmål, der i første omgang forekommer lidt indviklede. Så for at få overblik over tallene kan det være en fordel at opstille de oplysninger vi har i en antalstabel:

test \ tilstand	syg	rask	I alt
positivt udslag	a	c	a+c
ikke positivt udslag	b	d	b+d
I alt	a+b	c+d	

Øvelse 8.1

a) Se nu den del af filmen igen og noter dig hvilke oplysninger du får. Hvor i tabellen kan du indsætte disse oplysninger?

b) Læg mærke til, at det er procent-oplysninger. De enkelte procenttal der nævnes må jo være procent af noget. Men af hvad?

Øvelse 8.2

Det spørgsmål vi gerne vil have svar på er: Hvis jeg bliver testet positiv, hvad er så sandsynligheden for at jeg faktisk er syg. Hvilke tal i tabellen skal man kende, for at kunne svare på det?

Se nu igen den del af filmen, hvor Susanne diskuterer med de studerende i auditoriet. Der kommer to svar, som vi åbenbart kan bruge. Noter disse svar. Et af dem taler om et begreb, der hedder *falsk positiv*.

Øvelse 8.3. Begrebet falsk positiv

a) Hvad betyder dette begreb, *falsk positiv*?

b) (*Forudsætter du har arbejdet med hypotesetest*) Når vi arbejder med hypotesetest udregnes en p-værdi på grundlag af en nulhypotese. p-værdien sammenlignes med et på forhånd fastlagt signifikansniveau, og størrelsen af p-værdien afgør om vi forkaster eller accepterer nulhypotesen. Der er således indbygget en subjektivt fastlagt grænse mellem accept og forkastelse i hypotesetest. Dette betyder også, at vi kan begå fejl to typer af fejl: Forkaste noget sandt eller acceptere noget falsk. Læs evt. afsnittet om retssagsmetaforen i *Hvad er matematik*? B og svar på, hvad sammenhængen er mellem *falsk positiv*, *falsk negativ* og de to *fejltyper i hypotestest*.

c) Argumenter for, at hvis vi for et øjeblik glemmer selve tallene, og giver en rent kvalitativ beskrivelse af de enkelte rubrikker i tabellen, så kan det gøres således:

test \ tilstand	syg	rask	I alt
positivt udslag	sand positiv (a)	falsk positiv (c)	a+c
ikke positivt udslag	falsk negativ (b)	sand negativ (d)	b+d
I alt	a+b	c+d	

Giv med dine egne ord en beskrivelse af, hvad der dækker sig bag de andre tre begreber.

Øvelse 8.4
Der kom yderligere et svar fra en af de studerende. Hvor i tabellen er vi henne med dette svar?

Øvelse 8.5
I Susannes opsummering på casen trækkes nogle absolutte tal ind på scenen. Hvor i antalstabellen hører disse til? Overvej hvorfor vi med de nye oplysninger faktisk kan svare på spørgsmålet.

Øvelse 8.6. Case om betydningen af antallet af smittede
Løs nu selv følgende øvelse med brug af antalstabeller som ovenfor.
Antag vi har en test for HIV, der er meget effektiv, idet den fanger alle der er smittede. Testen har en falsk positiv rate på 5%, og en falsk negativ rate på 0%.
Vi har en samlet population på 1000, og ser på to forskellige situationer:
A) 40% er smittede B) 2% er smittede
Du testes positiv. Hvad er sandsynligheden for at du faktisk er smittet i henholdsvis tilfælde A og tilfælde B?

Der findes mange beslægtede problemstillinger, hvor vi så at sige gerne vil regne baglæns:
- fra en viden om hvor mange der testes positivt til en sandsynlighed for, hvor mange der faktisk er syge
- fra en viden om et barns og en potentiel faders blodtyper, til en sandsynlighed for at han faktisk er far til barnet.
- fra en viden om et dna-materiale fundet på et gerningssted og om en mistænkt persons dna-profil, til sandsynligheden for at det fundne materiale faktisk stammer fra den mistænkte.

De sandsynligheder vi her taler om kaldes *betingede sandsynligheder*. Som udgangspunkt har vi stadig et *frekventielt* grundlag for opstilling af sandsynligheder: Sandsynligheden for at en tilfældig nyfødt i Danmark er en dreng er 49,1%, fordi 49,1% af alle nyfødte er drenge, dvs frekvensen af drenge er 49,1%.
Ideerne og metoderne inden for denne del af sandsynlighedsregning behandles i næste afsnit, hvor vi også giver en kort introduktion til *Bayesiansk statistik*. Det er en gren af statistikken, nogle vil sige en skole inden for statistikken, hvor man populært sagt forlader det frekventielle grundlag og i stedet *tager udgangspunkt i opstilling af betingede sandsynligheder*.

9. Betingede sandsynligheder og Bayesiansk statistik (B og A)

Et forløb om betingede sandsynligheder og Bayesiansk statistik kan introduceres via et selvstændigt elevarbejde med materialet i afsnit 8 sammen med eller i stedet for den indledende case om at "scanne for terrorisme".

9.1. Case: Potentielle terrorister og paradokset om de falsk positive

(*Eleverne arbejder selv dette eksempel igennem som en opvarmning til emnet. Eksempel er lånt fra den canadiske forfatter Cory Doctorow, der diskuterer paradokset i sin bog Little Brother*)

Antag man har opdaget en ny og meget sjælden sygdom, som får navnet Super-AIDS. Sygdommen optræder med en hyppighed på ca 1 tilfælde ud af en million. Der udvikles en test der 99% sikker, hvormed menes, at den giver det korrekte resultat 99 ud af 100 gange – sandt, hvis man faktisk er smittet og falsk hvis man ikke er smittet. Eller sagt omvendt: Den giver et falsk resultat i 1% af tilfældene.

Sygdommen anses for ekstremt farlig, så det besluttes at give testen til 1 million indbyggere.

a) Udfyld en tabel som den følgende (med afrundede tal, det er jo hele mennesker):

test \ tilstand	har super-AIDS	har ikke super-AIDS	I alt
positivt udslag	1		
ikke positivt udslag	0		
I alt	1	999.999	1.000.000

Et rimeligt mål for, hvor præcis testen er, kunne være at angive hvor stor en procentandel af de positivt testede, der faktisk er syg.

b) Hvor præcis er denne test?

c) Sammenhold dette tal med, at testen blev præsenteret som 99% sikker. Hvori ligger forklaringen på dette paradox?

Når vi måler på meget små størrelser, skal vores måleudstyr også være meget fintmasket, eller være indrettet på at kunne registrere noget meget småt. Vil man pege på en enkelt pixel på sin skærm, kan en spids blyant godt bruges. Men blyanten er ikke anvendelig, hvis man skulle pege på et enkelt atom.

Lad os trække en parallel fra den sjældne sygdom "super-AIDS" til den aktuelle debat om overvågning af potentielle terrorister. Kan man ved at sammenkøre store datamængder fra mobiltelefoni, banktransaktioner, rejsemønstre, aktiviteter på sociale medier som fx Facebook mv finde potentielle terrorister?

Lad os antage, at efterretningsvæsenet har gode argumenter for at hævde, at hvis bestemte indikatorer er tilstede hos en person, så er der en vis sandsynlighed for, at vedkommende er potentiel terrorist. Og omvendt: Hvis en person faktisk er terrorist, så fanges de i filtret med 99% sikkerhed. Det afgørende i første omgang er ikke, om vi tror på sådanne indikatorer. Vi antager det er korrekt og analyserer metoden ud fra denne antagelse. Da vi ikke kan have sikker viden, men taler om *sandsynligheder*, og da vi taler om *potentielle* terrorister må man acceptere, at der også spottes nogen, der faktisk ikke er terrorister.

Lad os antage, at en organisation som det amerikanske efterretningsvæsen, NSA har skaffet sig adgang til alle bankkonti, til overvågning af alle mobilsamtaler i byen, til at kunne scanne alle facebookprofiler mv. De har lagt filtre ind, der frikender 99,9%, mens 0,1% af befolkningen matcher NSA's definition af "potentielle terrorister".

Rigtige terrorister, der eksempelvis er villige til at optræde i selvmordsangreb, er sjældne. I en by som New York på 20 millioner indbyggere anslås der at være højst 10 sådanne terrorister. Det betyder mindre for det følgende, om dette tal er fx en faktor 10 større, men der er i vestlige lande trods alt kun set ganske få tilfælde af denne type, selv om det ville være en ret enkel sag at udføre.

d) Anvend de givne oplysninger til at færdigudfylde følgende tabel over "scanningen" af New York for potentielle terrorister:

test \ tilstand	er terrorist	er ikke terrorist	I alt
positivt udslag	10		
ikke positivt udslag	0		
I alt	10		20.000.000

De potentielle terrorister opsamles på en liste, og en whistle-blower lækker listen til pressen. Det viser sig din nabo figurerer op listen.

e) Diskuter i gruppen, hvordan I ville reagere på en sådan oplysning.

f) Hvad er sandsynligheden for at en tilfældig person på listen faktisk er terrorist?

g) Diskuter i gruppen den beskrevne metode til at spotte potentielle terrorister.

En test giver aldrig sikker viden. Derfor opererer man i statistik med følgende fire begreber:
falsk positiv, *falsk negativ, sand positiv, sand negativ*

h) Placer de fire begreber i de fire rubrikker i tabellen og argumenter for hvordan du placerer dem.

9.2 Betingede sandsynligheder

Betingede sandsynligheder er sandsynligheder, der beregnes ud fra en eller anden given betingelse. Hvis man kaster med to terninger, så er sandsynligheden for at få en øje-sum på 10 lig med $\frac{3}{36}=\frac{1}{12}$, fordi der 36 kombinationer af to terninger og præcis kombinationerne (6,4), (5,5) og (4,6) giver summen 10. Hvis man nu ved, at en af terningerne er en 6'er, så er der i alt 11 kombinationer, hvoraf de to, nemlig (6,4) og (4,6) giver summen 10. Dvs med den nye viden er sandsynligheden ændret til $\frac{2}{11}$.

Dette kan vi udtrykke i et formelsprog, der viser sig nyttigt, på følgende måde:

Definitioner og notation vedr sandsynlighedsfelter

1. Den samlede mængde af kombinationer kalder vi *udfaldsrummet* og betegner det U:

U={(1,1), (1,2), ..(1,6), (2,1), (2,2), .., (2,6),…(6,1), (6,2), .. (6,6)}

Generelt betegner u_i et *udfald* og mængden af alle udfald kaldes for *udfaldsrummet*:

$$U=\{u_1,u_2,...,u_n\}$$

2. *Hændelser* er delmængder af udfaldsrummet, og betegnes ofte med store bogstaver:

A = alle kombinationer, der giver øje-summen 10:

A = {(6,4), (5,5), 4,6)}

B = alle kombinationer, hvor en af terningerne viser 6:

B= {(1,6), (2,6), .. (6,6), (6,1), (6,2), ..(6,5)}

3. *Sandsynligheder* angives med en sandsynlighedsfunktion *P*, således:

$$P((5,5))=\frac{1}{36},\ P(\text{A})=\frac{3}{36},\ P(\text{B})=\frac{11}{36},\ P(\text{U})=\frac{36}{36}=1,\ P(\text{Ikke-A})=\frac{33}{36}$$

Generelt gælder, at

$$0\le P(u_i)\le 1 \text{ samt at } \sum_1^n u_i=u_1+u_2+...u_n=1$$

4. Hvis alle udfald har samme sandsynlighed, som det er tilfældet med terningekast med én terning eller med to terninger (sort og rød fx), så siger vi, at vi har et *symmetrisk sandsynlighedsfelt*. I et symmetrisk sandsynlighedsfelt kan vi udregne sandsynligheden af en hændelse, som fx A, ved formlen: $P(A)=\frac{\text{antal udfald i } A}{\text{samlede antal udfald i } U}$

5. Symbolet $A\cap B$ angiver, at to hændelser som A og B *begge indtræffer*. Det læses af og til: "*både A og B*". $A\cap B$ kaldes også fællesmængden af A og B. I vort eksempel er $A\cap B=\{(6,4),(4,6)\}$.

6. Når vi regner *ud fra en given betingelse*, som fx at vi ved at en af terningerne viser 6, dvs at B er indtruffet, så angiver vi det således: $P(A|B)$. Dette betyder: sandsynligheden for *A* når vi ved *B* er indtruffet. Det kaldes også *den betingede sandsynlighed for A givet B*

Hvordan udregnes så $P(A|B)$? Når vi ved, at B er indtruffet, så er de mulige udfald altså ikke hele *U*, men alene *B*. Og når vi ved, at *B* er indtruffet, så er den hændelse, vi spørger om, ikke hele *A*, men $A\cap B$.

Derfor kan vi i symmetriske sandsynlighedsfelter, som eksemplet med kast med to terninger, udregne:

$$P(A\,|\,B)=\frac{\text{antal udfald i } A\cap B}{\text{samlede antal udfald i } B}=\frac{2}{11}$$

Denne formel kan vi omskrive lidt (forkort brøken, dvs divider både tæller og nævner med samme tal):

$$P(A\,|\,B)=\frac{(\text{antal udfald i } A\cap B)\,/\,(\text{antal udfald i } U)}{(\text{samlede antal udfald i } B)\,/\,(\text{antal udfald i } U)}=\frac{P(A\cap B)}{P(B)}$$

Med de givne talværdier ville den sidste udregning give:

$$P(A\,|\,B)=\frac{P(A\cap B)}{P(B)}=\frac{2/36}{11/36}=\frac{2}{11}$$

altså naturligvis det samme som ovenfor.

Men årsagen til, at vi foretager denne omskrivning er, at i den sidste formel har vi sluppet optællingen af antal, som er knyttet til symmetriske sandsynlighedsfeter. Denne formel kan vi derfor anvende som den generelle definition af *den betingede sandsynlighed for A givet B*:

Definition: Betinget sandsynlighed
Den betingede sandsynlighed for A givet B betegnes $P(A\,|\,B)$ og er givet ved:

$$P(A\,|\,B)=\frac{P(A\cap B)}{P(B)}$$

Udtrykket fortolkes som: *Sandsynligheden for A når vi ved B er indtruffet.*

Argumentationen ovenfor fortæller, at dette er en generalisering af tællemetoden fra symmetriske felter.

***Bemærkning om uafhængighed*:**
Betingede sandsynligheder giver anledning til at give en formel definition på et centralt begreb i sandsynlighedsregning, nemlig begrebet *uafhængige hændelser*:
Hændelserne *A* og *B* kaldes uafhængige, hvis der gælder, at $P(A\,|\,B)=P(A)$. Dvs den ekstra oplysning om, at *B* er indtruffet påvirker ikke sandsynligheden for at *A* indtræffer.
Hvis *A* og *B* er uafhængige ser vi af formlen, at der gælder: $P(A\cap B)=P(B)\cdot P(A)$
Nogle lærebøger bruger faktisk den sidste formel som en definition af uafhængighed. Men dermed mister man intuitionen om uafhængighed. Det er imidlertid vigtigt at holde fast i, at det er en formel definition. Begrebet indgår jo også i daglig sproget, og her skal man passe på ikke kun at forlade sig på sin intuition. Begrebet uafhængige hændelser er helt central i behandlingen af de såkaldte binomialmodeller, der er behandlet i B-bogens kapitel 9. Vi vil ikke gå yderligere ind i dette her.

En af de stærke sider ved betingede sandsynligheder er, at man kan "regne baglæns"

Sætning: At regne forlæns og baglæns med betingede sandsynligheder
Hvis A og B er to hændelser i et udfaldsrum U, så gælder:

$$P(A\cap B)=P(A\,|\,B)\cdot P(B)$$

$$P(A\cap B)=P(B\,|\,A)\cdot P(A)$$

$$P(A\,|\,B)=\frac{P(A)}{P(B)}\cdot P(B\,|\,A)$$

Beviset overlades til læseren.

Bemærk, at den tredje formel vender betingelsen om, og dermed giver os mulighed for at "regne baglæns". I de følgende eksempler vil vi både arbejde med formlerne og med tabelopstillingerne, der ofte er et redskab til forholdsvis enkle løsninger.

Øvelse 9.1. Hvilket køn har det andet barn?

(*Vi antager i dette eksempel, at der fødes lige mange piger og drenge i Danmark*)

I et nabohus flytter et par ind, som har to børn. Hvad er sandsynligheden for at begge er drenge? Hvad er sandsynligheden for, at mindst én er en dreng?

Udfaldsrummet her er $U_1 = \{(P,P),(P,D),(D,P),(D,D)\}$, hvor rækkefølgen i talparrene angiver i hvilken rækkefølge børnene blev født. *P* og *D* står for pige og dreng. Svarene er naturligvis:

$$P(\text{to drenge}) = \tfrac{1}{4} = 25\% \qquad P(\text{mindst én dreng}) = \frac{3}{4} = 75\%$$

Du møder parret, der er ude at gå. De har det ene barn med sig. Det er en dreng.

Hvilket køn har det andet barn? Hvad er sandsynligheden for, at det andet barn også er en dreng?

25%? 50%? Et helt andet tal?

Umiddelbart vil mange svare det første – er det andet barn også en dreng er der jo to, og vi har lige udregnet, at sandsynligheden for to drenge er 25%.

Andre hælder måske til 50%: Når vi ikke kender kønnet, så må det være fifty-fifty for dreng-pige.

Men begge svar er forkert. De tager ikke hensyn til at vi har fået en viden i og med vi nu ved, at det ene barn er en dreng.

Udregnet ved tællemetoden: Udfaldsrummet er nu $U_2 = \{(P,D),(D,P),(D,D)\}$. Så sandsynligheden for, at den anden også er en dreng er således:

$$P(\text{to drenge} \mid \text{en dreng}) = \tfrac{1}{3} = 33{,}3\%.$$

Da du kommer hjem fortæller du, at du mødte de nye naboer, og at du så, de har en dreng. Ja, og så har en mere, der ligger i barnevogn, lød svaret. Ændrer den nye oplysning på svaret på spørgsmålet: Hvad er sandsynligheden for, at det andet barn også er en dreng?

Mange vil nok svare, at det ikke ændrer noget. Men oplysningen rummer faktisk en ny information. Nu ved vi ikke blot, at den ene af de to børn er en dreng, men også at det er den ældste. Dvs udfaldsrummet er nu $U_3 = \{(D,P),(D,D)\}$. Så sandsynligheden for, at den anden også er en dreng er således:

$$P(\text{to drenge} \mid \text{den ældste er en dreng}) = \tfrac{1}{2} = 50\%$$

Læg mærke til, at sandsynligheden for at hændelsen indtræffer (her: to drenge) stiger med den information vi får.

Øvelse 9.2. Testet positiv - er du syg?
Antag vi har en test for HIV, der er rimelig effektiv, idet den fanger 90% af alle der er smittede. Testen har således en falsk negativ rate på 10%. Men testen fanger ikke kun de syge, den har også en falsk positiv rate på 5%. Vi har en samlet population på 1000, hvoraf i alt 2% er smittede.
Du testes positiv. Hvad er sandsynligheden for at du faktisk er smittet?
Når man skal svare på sådanne spørgsmål, kan det være en fordel at stille oplysningerne op i en tabel som følger:

test \ tilstand	HIV smittet	ikke HIV smittet	I alt
positivt udslag	18	49	67
ikke positivt udslag	2	931	933
I alt	20	980	1000

a) Løs opgaven ved simpel optælling
b) Hvordan vil du definere mængderne A og B, omtalt i definitionen på betingede sandsynligheder? Hvad udgør mængden $A \cap B$?
c) Hvad er $P(A)$, $P(B)$ og $P(A \cap B)$?
d) Udnyt nu formlen $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$ til at løse opgaven.

I et tilfælde hvor vi har givet absolutte tal, som i øvelsen ovenfor, er det klart lettere at optælle. Tabellen er et nyttigt redskab til at skabe overblik, så vi ser, at vi kun behøver at regne i den øverste række.

Hvis vi ikke har absolutte tal, kun %-tal., så kunne vi naturligvis gennemregne med et taleksempel, men kunne vi ikke klare os uden? Det handler næste eksempel om.

Eksempel. Klassikeren fra Harvard Medical School
Mange amerikanske lærebøger om statistik indeholder følgende eksempel på, hvor let det er at slutte forkert i statistik:

> Consider the following problem
>
> *A particular heart disease has a prevalence of 1/1000 people. A test to detect this disease has a false positive rate of 5%. Assume that the test diagnoses correctly every person who has the disease. What is the chance that a randomly selected person found to have a positive result actually has the disease?*
>
> This question was put to 60 students and staff at Harvard Medical School.
>
> Almost half gave the response 95%.
>
> The average answer was 56%.
>
> The correct answer was given by just 11 participants.

1. løsningsmetode

Sygdommen rammer en ud af 1000. Derfor opstiller vi en antalstabel baseret på en population på 1000. Det grønne felt (falsk negativ) rummer 0, da testet fanger alle der faktisk er syg.
Da der er en syg, er der 999 ikke-syge. Antal falsk positive er så 5% af 999. Dette er fundet lig med 50, så i alt testes 51 positivt.
Betragt nu tabellen. En person er testet positiv og er altså blandt de 51. I denne gruppe er der 1 syg.
Så sandsynligheden for at vedkommende har sygdommen er:

$$P(\text{syg}\,|\,\text{testet positiv}) = \frac{1}{51} = 1{,}96\% \approx 2\%$$

test \ tilstand	syg	ikke syg	I alt
positivt udslag	1	50	51
ikke positivt udslag	0	949	949
I alt	1	999	1000

2. Løsningsmetode

Vi formaliserer oplysningerne:

A: Mængden af syge	$P(A) = 0{,}001$	
not A: Mængden af raske	$P(\text{not } A) = 0{,}999$	
B: Mængden der testes positiv.	$P(B\,|\,A) = 1$	$P(B\,|\,\text{not } A) = 0.05$

En person er testet positiv og ønsker at kende sandsynligheden for at vedkommende faktisk er syg.
Dvs. i formelsproget ønsker vi at beregne $P(A\,|\,B)$.

Udnyt formlen: $$P(A\,|\,B) = \frac{P(A)}{P(B)} \cdot P(B\,|\,A) \qquad (*)$$

Vi kan se, at vi her mangler kendskab til $P(B) = P(\text{positiv test})$.

Da alle individer enten er syg eller rask, dvs enten ligger i A eller i not A, så kan vi opdele B i dem der ligger i A, dvs $B \cap A$ og dem der ligger i not A, dvs $B \cap \text{not } A$. Dermed kan vi udregne $P(B)$ således:

$$P(B) = P(B \cap A) + P(B \cap \text{not } A)$$

Udnyt sætningen om at regne baglæns til omskrivningen:

$$\begin{aligned} P(B) &= P(B \cap A) + P(B \cap \text{not } A) \\ &= P(B\,|\,A) \cdot \mathrm{P}(A) + P(B\,|\,\text{not } A) \cdot \mathrm{P}(\text{not } A) \\ &= 1 \cdot 0{,}001 + 0{,}05 \cdot 0{,}999 \end{aligned}$$

Indsæt nu tallene i formlen (*):

$$P(A\,|\,B) = \frac{P(A)}{P(B)} \cdot P(B\,|\,A) = \frac{0{,}001}{1 \cdot 0{,}001 + 0{,}05 \cdot 0{,}999} \cdot 1 = \frac{1}{1 + 49{,}5} \approx 2\%$$

Så sandsynligheden for at vedkommende er ca 2%.

Øvelse 9.3. Elisa testen for screening af blod
I en rapport om udviklingen i bestræbelserne på at bekæmpe AIDS-epidemien kan man læse følgende:

> The ELISA test was introduced in the mid 1980's to screen donated blood for the presence of AIDS antibodies. When antibodies are present, ELISA is positive with a probability of about 0.98; when the blood tested is not contaminated with antibodies, the test gives a positive result with a probability of 0.07. These numbers are conditional probabilities. If one in a thousand of the units of blood screened by ELISA contain antibodies, then (??) of all positive responses will be false positive.
> *(kilde: Lynn Arthur Steen (red.), New approaches to Numeracy)*

Hvad skal der stå på den tomme plads (??, dvs hvor stor en andel falsk positive er det?

Øvelse 9.4. Medicinsk forsøg
Et medicinalfirma tilrettelægger en test af en ny allergimedicin. 1500 testpersoner deltager og opdeles i 3 grupper med 500 i hver. Den ene gruppe får det klassiske produkt, firmaet længe har haft på sin liste. Den anden får et placebo-præparat. Og endelig får den tredje gruppe det nye betydeligt stærkere præparat. De enkelte deltagere ved naturligvis ikke hvilke præparater de får.

	forbedring	**ingen virkning**	**forværrring**	**I alt**
Gruppe 1	159	301	40	500
Gruppe 2	128	342	30	500
Gruppe 3	318	77	105	500
I alt	605	720	175	

Lad os antage, at stikprøven er repræsentativ for den relevante population.

a) Hvilken konklusion – i form af anbefalinger til firmaet - vil du umiddelbart drage om det nye stærkere præparat?
b) En tilfældig valgt person blandt de 1500 får det værre. Hvad er sandsynligheden for at han har fået det nye præparat?
c) En tilfældig valgt person blandt de 1500 får det bedre. Hvad er sandsynligheden for at han har fået det nye præparat?
d) Vil du ændre dine anbefalinger?

Øvelse 9.5. Hvad vej vender betingelsen?
Bayesianske metoder har i stor udstrækning fundet vej til især amerikanske retssale. Det gives vi en kort introduktion til i næste afsnit. I en af de artikler vi henviser til gives følgende eksempel:

> Suppose a crime has been committed and that the criminal has left some physical evidence, such as some of their blood at the scene. Suppose the blood type is such that only 1 in every 1000 people has the matching type. A suspect, let's call him *Fred,* who matches the blood type is put on trial. The prosecutor claims that the probability that an innocent person has the matching blood type is 1 in a 1000 (that's a probability of 0.001). Fred has the matching blood type and therefore the probability that Fred is innocent is just 1 in a 1000.

Analyser anklagerens påstand ved hjælp af betingede sandsynligheder. De to centrale spørgsmål handler om *blodtypen* og om *Fred er uskyldig*, som han påstår. Indfør to hændelser A og B, og opstil et udtryk for, hvad anklageren beregner. Hvad er din konklusion? Du kan evt anvende næste øvelse, som et hint

Øvelse 9.6 $P(A|B)$ **eller** $P(B|A)$**?**
Du får et billede af en kvinde, der er attraktiv og smukt klædt, og bliver spurgt om sandsynligheden for at hun er fotomodel. Hvad er det egentlig for en betinget sandsynlighed vi spørger om?
Du kan evt løse opgaven ved først at definere følgende hændelser:
A: Kvinden er attraktiv og forstår at klæde sig smukt.
B: Kvinden er fotomodel.
Spørger vi om $P(A|B)$ eller $P(B|A)$?

Den generelle udgave af Bayes formel

I analysen af "klassikeren fra Harvard" anvendte vi følgende formel:

$$P(B) = P(B \cap A) + P(B \cap \text{not } A)$$

Formlen bygger på den enkle iagttagelse, at enten indtræffer hændelsen A eller også gør den ikke, dvs. de to situationer $B \cap A$ og $B \cap \text{not } A$ udtømmer alle muligheder. Men dette kan vi generalisere til situationer, hvor alle muligheder kan opdeles i adskilte mængder $A_1, A_2, \ldots, A_n$. Det kan fx være situationen, hvor vi opdeler befolkningen i indkomstgrupper, svarende til $A_1, A_2, \ldots, A_n$, og hvor hændelsen *B* kunne være holdningen til om vi i Danmark skal gå over til Euroen. Da er:

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + \ldots + P(B \cap A_n) \qquad (*)$$

Hvis vi nu ønsker at beregne $P(A_k|B)$, så kan vi tage udgangspunkt i definitionen:

$$P(A_k|B) = \frac{P(A_k \cap B)}{P(B)}$$

Heri indføres nu først (*):

$$P(A_k|B) = \frac{P(A_k \cap B)}{P(B \cap A_1) + P(B \cap A_2) + \ldots + P(B \cap A_n)}$$

Ofte har vi situationer, hvor vi kender sandsynligheder for de omvendte betingelser, fx $P(B|A_k)$, og ved at anvende de tidligere formler, kan vi nu omskrive til følgende

$$P(A_k|B) = \frac{P(B|A_k) \cdot P(A_k)}{P(B|A_1) \cdot P(A_1) + P(B|A_2) \cdot P(A_2) + \ldots + P(B|A_n) \cdot P(A_n)}$$

Sætning: Bayes formel
Hvis hændelserne A og B er to hændelser i et udfaldsrum U og hændelsen A kan opdeles i n adskilte hændelser / delmængder, $A_1, A_2, \ldots, A_n$, så gælder:

$$P(A_k|B) = \frac{P(B|A_k) \cdot P(A_k)}{P(B|A_1) \cdot P(A_1) + P(B|A_2) \cdot P(A_2) + \ldots + P(B|A_n) \cdot P(A_n)}$$

Øvelse 9.7. Bliver der regn på bryllupsdagen
En amerikansk kvinde Marie skal giftes ved en spektakulær udendørs ceremoni i et ørkenområde uden for Las Vegas. De senere år har det kun regnet 5 dage om året. Uheldigvis har en af TV stationernes meteorologer forudsagt, at det bliver regn, netop på bryllupsdagen. Når det faktisk regner, har meteorologen forudsagt dette i 90 % af tilfældene. Men også i 10% af de dage, hvor det ikke regner, har han forudsagt regn. Hvad er sandsynligheden for at det vil regne på bryllupsdagen?

9.3 Bayesiansk statistik

Sandsynlighedsregningen er som udgangspunkt bygget op på et *frekventielt* grundlag. Hermed menes, at sandsynligheden for at en tilfældig nyfødt i Danmark er en dreng er 49,1%, fordi 49,1% af alle nyfødte er drenge, dvs frekvensen af drenge er 49,1%. Frekventielle sandsynligheder bygger på tanken om et eksperiment, der kan gentages et antal gange.

Men man behøver ikke lede efter eksotiske eksempler for at se, at det ofte ikke giver mening at tale om frevenser som grundlag. Mord og andre emner for alvorlige retssager er én type af eksempler. Meteorologers vejrprognoser kan også kun vanskeligt modelleres ud fra en ren frekventiel model. Når eksperimentalfysikere leder efter nye fænomener, eller når der skal laves risikovurderinger i en industriproduktion, giver det ikke rigtig mening af tale om, at bestemme sandsynligheder ud fra en frekventiel tilgang.

I stedet træder den Bayesianske matematik og statistik ind på banen med en ny tilgang til sandsynlighedsbegrebet. Sandsynlighedsregning og statistik handler jo altid om fænomener, hvor vi ikke har sikker viden. Den grundlæggende ide er nu den ret enkle, at man udtrykker sin usikkerhed i form af sandsynligheder. Står man over for et fænomen, en bestemt hændelse, og ønsker at vurdere sandsynligheden for at denne indtræffer, starter man med *subjektivt* at tilegne hændelsen en sandsynlighed, en talværdi mellem 0 og 1. Det er naturligvis her, at modstanderne af denne skole inden for statistikken sætter angrebet ind – det subjektive skurrer i ørene. Blev man stående der, ville det også være meningsløst. Men man begynder jo netop at indsamle oplysninger, indsamle data om fænomenet. Og for hver ny relevant oplysning opdaterer man sandsynligheden. Og dette kan kun ske på bais af Bayes formel. Efterhånden som der indsamles data, så viser det sig, at selv om man havde vidt forskellige subjektive udgangspunkter, så sker der en tilnærmelse mellem de efterfølgende vurderinger af sandsynlighederne.

Eksempel: Bayesiansk matematik i retssalen

Bayesianske metoder er ofte blevet inddraget i især amerikanske retssager. Vi vil her se på det tilfælde, hvor DNA-materiale fra den formodede gerningsmand er det eneste konkrete bevismateriale.
Vi definerer hændelserne
S: Den anklagede person er skyldig
D: Den anklagedes DNA-profil matcher det DNA, som blev fundet på gerningsstedet.

Hvis man nu antager at man har fundet en matchende DNA-profil på gerningsstedet, så kan man regne på sandsynligheden for, at den anklagede er skyldig:

$$P(S \mid D) = \frac{P(S)}{P(D)} \cdot P(D \mid S)$$

På højre side er leddet $P(D \mid S) = 1$, da man netop har fundet en matchende DNA-profil. $P(S)$ er sandsynligheden for at den anklagede er skyldig, alt taget i betragtning. Her kan man anvende fakta som, hvor mange der bor i lokalområdet, sandsynligheden for at vidneudsagnene er rigtigt osv., men det er klart, at her bliver bevisførelsen og udregningerne mere uklare.

Leddet , dvs. sandsynligheden for, at der findes et match mellem den anklagedes DNA-profil og DNA-materialet på gerningsstedet beregnes ved formlen:

$$P(D) = P(D \mid S) \cdot \mathrm{P}(S) + P(D \mid \text{not } S) \cdot \mathrm{P}(\text{not } S)$$

Det sidste led på højresiden beskriver det tilfælde, hvor den anklagede er så uheldig at være uskyldig, men at hans DNA-materiale alligevel findes på gerningsstedet.

Vi går nu over til et konkret eksempel: *Regina versus Adams*. Du kan finde sagen beskrevet *her* (http://en.wikipedia.org/wiki/R_v_Adams)
I retssagen mod Adams i 1996, hvor han var anklaget for voldtægt, var det eneste bevis mod ham, at hans DNA fandtes på gerningsstedet. Offeret kunne ikke genkende ham, og Adams havde også et alibi, som dog afhang af hans kæreste. Anklagernes argument var, at chancen for at en tilfældig mands DNA-profil matchede DNA-materialet fra gerningsstedet var:

$$P(D) = \frac{1}{200.000.000}.$$

Dette virker ved første indtryk som et overvældende godt bevismateriale. Adams' forsvarere valgte på den anden side at inddrage Bayes formel i et håb om at vise, at det kan være problematisk at dømme en person udelukkende ud fra DNA-materiale.

Adams' forsvarere så på det bevismateriale i retssagen, som ikke var knyttet til DNA-undersøgelser og påviste, at sandsynligheden for at Adams var skyldig ud fra dette materiale blot var:

$$P(S) = \frac{1}{3.600.000}$$

Dette tal fandt de frem til ved at opstille nogle spørgsmål/hændelser, hvorpå jurymedlemmerne kunne hæfte deres egen sandsynlighedsvurdering. Forsvarerens bud på disse sandsynligheder var:

- 75 % sandsynlighed for at gerningsmanden var fra lokalområdet og i 18-60.
- 90 % sandsynlighed for at den anklagede ikke ville blive genkendt af offeret, hvis den anklagede var uskyldig.
- 25 % sandsynlighed for at den anklagedes alibi holdt vand, hvis han var skyldig. 50 % i tilfældet, hvis den anklagede var skyldig.

Disse oplysninger kan vha. Bayes formel bruges til at give en vurdering af sandsynligheden for, hvorvidt Adams var skyldig:

$$P(S \mid D) = \frac{P(S)}{P(D)} \cdot P(D \mid S) = \frac{P(D \mid S) \cdot P(S)}{P(D \mid S) \cdot \mathrm{P}(S) + P(D \mid \text{not } S) \cdot \mathrm{P}(\text{not } S)} = \frac{1 \cdot \frac{1}{3.600.000}}{1 \cdot \frac{1}{3.600.000} + \frac{1}{200.000.000} \cdot (1 - \frac{1}{3.600.000})} = 0{,}9823.$$

Altså er sandsynligheden for at Adams var gerningsmanden reduceret til at være omkring 54/55. Stadig rimelig sandsynligt, men alligevel en væsentlig forbedring, idet der her dog må siges at findes en realistisk sandsynlighed for, at Adams er uskyldig (ca. 1,8 %).

Jurymedlemmerne kunne selvfølgelig komme med deres egen vurdering over sandsynlighederne i de tre punkter her ovenfor. På denne måde bliver Bayes formel et redskab hos juryen til at regne på sandsynligheden for, hvorvidt den anklagede var skyldig.

Der viste sig at være en del bekymring og forvirring over, hvorvidt denne anvendelse af Bayes formel lod sig gøre på en rimelig måde. Det endte dog med, at juryen fik et spørgeskema, hvor de skulle angive forskellige vurderinger i %. I spørgeskemaet var der så inkluderet en udregningsformel.
Sagen endte med, at Adams blev dømt skyldig.

Du kan **her** finde en række materialer om Bayesiansk statistik, om anvendelser af Bayesianske metoder i autentiske retssager, samt oplæg til studieretningsprojekter inden for disse emner
Det kan være en omstændelig proces at foretage de mange beregninger i Bayesiansk statistik, men med moderne maskiner er dette blevet mere overkommeligt, og har åbnet for flere anvendelser.