3 Usikkerhed og sandsynligheder

De fleste samfundsvidenskabelige problemstillinger involverer usikkerhed. For at kunne analysere disse er det nødvendigt med en dybere forståelse af begrebet usikkerhed samt redskaber til at håndtere usikkerheden. Redskaberne består især af matematiske modeller, som kvantificerer usikkerheden. I dette kapitel diskuterer vi først, hvad man forstår ved usikkerhed. Dernæst ser vi i afsnit 3.2 nærmere på, hvordan udtrækningen af en stikprøve fra en population er forbundet med usikkerhed. Sandsynlighedsmodellen, som vi introducerer i afsnit 3.3, formaliserer denne usikkerhed, så vi kan tilskrive forskellige hændelser sandsynligheder. I afsnit 3.4 skal vi lære at regne med disse sandsynligheder, mens vi i afsnit 3.5 går et skridt videre og kigger på begrebet betinget sandsynlighed, som er sandsynligheden for en hændelse, givet at en anden hændelse allerede er indtruffet. Endelig ser vi i afsnit 3.6 på, hvad vi forstår ved uafhængighed og afhængighed mellem to hændelser, og hvordan afhængighed ofte forveksles med en kausal sammenhæng.

3.1 Usikkerhed

Vi opfatter en situation som usikker, når vi ikke ved, hvad der vil ske. Ofte kender vi godt de mulige begivenheder, vi ved blot ikke hvilken af dem, der vil indtræffe.

Eksempel 3.1: Vejret i morgen

Det er juni måned, og vi ved ikke, hvordan vejret vil blive i morgen. Vi ved dog, at morgendagen vil byde på enten solskin, regn, sne eller blandingsvejr. Andre muligheder er der ikke.

Eksempel 3.2: Obligationskurser

Kursen i morgen kl. 12.00 på en 30-årig byggeobligation er relevant for mange virksomheder i den finansielle sektor, og ikke mindst for de boligejere, som overvejer at omlægge deres realkreditlån. I princippet kan kursen i morgen være en hvilken som helst værdi mellem 0 og uendelig, selvom vi nok på forhånd kan udelukke en del af disse værdier som værende helt usandsynlige, fx kursen 3,21 milliarder.

Eksempel 3.3: Efterspørgsel efter cykler

Cykelproducenterne er interesserede i at vide, hvor mange cykler danskerne vil efterspørge (forbruge) i det kommende år. På forhånd ved de kun med sikkerhed, at efterspørgslen bliver et heltal, som er større end eller lig med nul.

Alle de tre ovenstående eksempler involverer usikkerhed. Usikkerhed kan imidlertid opstå af forskellige årsager. I eksempel 3.3 kan årsagen til usikkerheden være, at folk faktisk ikke ved, hvor mange cykler de vil købe det kommende år, fordi de ikke kender næste års indtægter og udgifter. Årsagen kan imidlertid også være, at selvom den enkelte forbruger ved præcis, hvor mange cykler han eller hun vil efterspørge i det kommende år, så er denne information ukendt for cykelproducenterne. I begge tilfælde er der tale om usikkerhed set fra cykelproducenternes synspunkt.

Uanset årsagen til usikkerheden er vi interesserede i at kunne kvantificere den og lave beregninger på den. Cykelproducenterne fra eksempel 3.3 ønsker måske at beregne sandsynligheden for, at efterspørgslen bliver mindre end 300.000 cykler i det kommende år. Måske vil de også gerne vide, hvordan denne sandsynlighed påvirkes af benzinprisen. Tilsvarende kunne man som boligejer i eksempel 3.2 ønske at kende sandsynligheden for, at kursen er mindre end 100, idet man kun kan optage lån i den pågældende serie, så længe kursen er under denne værdi.

For at håndtere usikkerheden og lave sådanne beregninger kan vi opbygge en matematisk sandsynlighedsmodel. Inden vi kan opstille denne, skal vi kigge lidt nærmere på begreberne population, udvælgelsesmekanisme og stikprøve, som danner rammen for en basal forståelse af begrebet usikkerhed.

3.2 Fra population til stikprøve

I dette afsnit ser vi på sammenhængen mellem en population og en stikprøve. Som nævnt i kapitel 2 er en *population* den statistiske betegnelse for en samling af individer, genstande eller mere abstrakte objekter. Det kunne fx være alle personer i Danmark, hvis vi ønsker at undersøge danskernes indkomst eller alder. Hvis vi i stedet undersøger terningkast, så er det de seks sider på terningen, der udgør den relevante population. En *stikprøve* er en samling af værdier fra populationen. Det kan fx være 1000 indkomster eller en række kast med en terning.

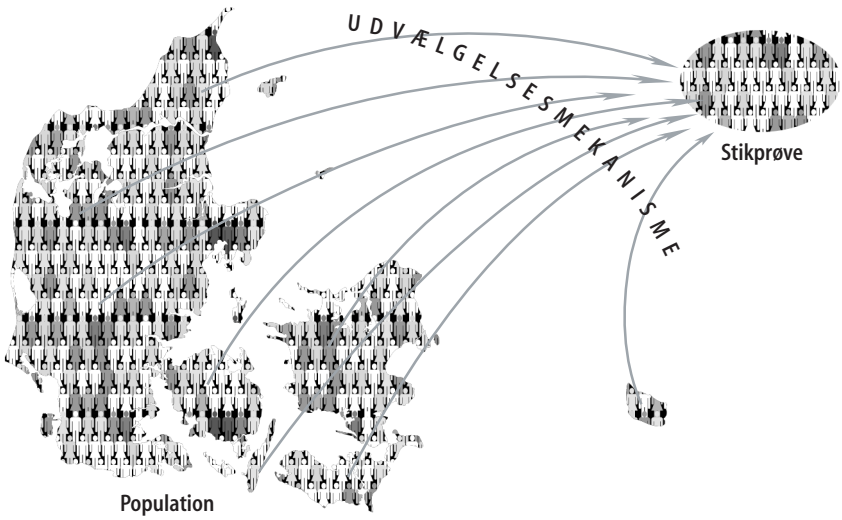
Stikprøven er resultatet af en *udvælgelsesmekanisme,* som skaber stikprøven fra populationen. Hvis fx en terning kastes tre gange, så kunne udvælgelsesmekanismen beskrives som følger: „Læg terningen i raflebægeret, ryst raflebægeret, kast terningen og aflæs resultatet. Gentag dette indtil tre terningkast

foreligger.“ Resultatet af de tre kast er da vores stikprøve. En stikprøve er forbundet med usikkerhed. Usikkerheden opstår, fordi raflebægeret rystes, og man derfor ikke kender udfaldet på forhånd.

Udvælgelsesmekanismen til at opnå en stikprøve på fx 1000 indkomster fra populationen af personer i Danmark er anderledes. Denne udvælgelsesmekanisme kunne være: „Skriv alle personnumre ned på sedler og put dem i en (meget) høj hat. Ryst hatten grundigt og udtræk med lukkede øjne en seddel. Undersøg indkomsten for personen med det udtrukne personnummer og skriv resultatet. Gentag dette 1000 gange.“ På denne måde opnås en stikprøve på 1000 indkomster. Usikkerheden opstår her, fordi den høje hat rystes, og udtrækningen sker med lukkede øjne. I praksis ville den høje hat være erstattet af en computer, som i princippet ville gøre det samme som at ryste hatten og udtrække med lukkede øjne. Udvælgelsesmekanismen beskriver både, hvordan usikkerheden opstår, og hvor mange elementer fra populationen, der skal undersøges.

Overordnet kalder man en population, en udvælgelsesmekanisme og en stikprøve for et *eksperiment*. Selvom ordet eksperiment kan lede tankerne hen på et laboratorieforsøg, så dækker begrebet meget bredere i statistikkens verden. Eksemplet med en stikprøve af danskernes indkomster er således et eksperiment i statistisk forstand. Figur 3.1 illustrerer de tre elementer i et eksperiment.

Vi er interesserede i stikprøver af to årsager. For det første er stikprøver små billeder af populationen. De kan derfor bruges til at lære noget om denne. Stikprøven af indkomster er et eksempel på dette. Vi kan her bruge stikprøven



Figur 3.1:
Et eksperiment

til at få et skøn over danskernes (populationens) indkomster. For det andet kan stikprøver være interessante i sig selv. Det er fx tilfældet, hvis man har indgået et væddemål om, hvilken side terningen lander på. Uanset hvilken interesse, man har i stikprøven, opstår der usikkerhed, fordi man ikke kender udfaldet af stikprøven på forhånd.

Alle problemstillinger i denne bog kan opstilles som eksperimentet illustreret i figur 3.1, hvis vi udvider vores fortolkning af, hvad en population er. I det følgende vil vi derfor forklare, hvad vi forstår ved populationer og i den forbindelse skelne mellem to typer: *Virkelige populationer* og *superpopulationer.*

3.2.1 Virkelige populationer

Vi har allerede set på virkelige populationer i kapitel 2. En virkelig population er en population, der eksisterer i virkeligheden. Den har et endeligt antal elementer, og man kan således tælle antallet af elementer og definere en andelsfunktion, som vi gjorde i kapitel 2. Udvælgelsesmekanismen fra en virkelig population har typisk også en fysisk beskrivelse. Når vi udtrækker en person eller kaster en terning, så foretager vi en handling. Vi har altså en vis kontrol over udvælgelsesmekanismen. Som vi skal se i afsnit 3.3, så har udvælgelsesmekanismen direkte konsekvenser for, hvordan sandsynligheder beregnes. I kapitel 9 diskuterer vi forskellige udvælgelsesmekanismer og fejlkilder, som kan opstå i praksis, når man udtager en stikprøve fra en virkelig population.

3.2.2 Superpopulationer

Det viser sig, at det er praktisk også at arbejde med abstrakte, ikke-eksisterende populationer, når man skal beskrive usikkerhed. De følgende eksempler illustrerer dette:

Eksempel 3.4: En superpopulation af obligationskurser

Kursen i morgen klokken 12.00 på en 30-årig byggeobligation er usikker. I morgen klokken 12.00 kender vi kursen, men hvor kommer den fra? Kursen kan i princippet være enhver værdi mellem nul og uendelig, men kun én af disse værdier bliver realiseret. Man kan forestille sig alle de mulige værdier af kursen i morgen klokken 12.00 som en population af kurser. Det er dog en population, man ikke kan finde i virkeligheden. Men ved alligevel at forestille sig denne ikke-eksisterende population af mulige kurser kan man sige, at kursen i morgen klokken 12.00 er en stikprøve fra denne population.

Eksempel 3.5: En superpopulation af regnskabsresultater

En virksomheds resultat i næste regnskabsår er usikkert. Men når man får opgjort regnskabet efter næste år, så står der ét og kun ét tal på bundlinjen. For at kunne formalisere usikkerheden i dag omkring resultatet næste år er det igen nyttigt at tænke på en population bestående af alle mulige resultater

for næste år. Et af disse mulige resultater vil ende med at blive virksomhedens faktiske resultat. I dette eksperiment er stikprøven altså lig med virksomhedens realiserede resultat.

Eksempel 3.6: En superpopulation af vejrlig

I eksempel 3.1 omkring morgendagens vejr kan vi forestille os dette som trukket ud af en population bestående af: Sol, regn, sne og blandingsvejr.

Eksemplerne ovenfor illustrerer, hvordan vi kan overføre tankegangen fra eksperimenter med virkelige populationer til eksperimenter med ikke-eksisterende populationer. For at skelne mellem en virkelig og en ikke-eksisterende population kalder vi sidstnævnte for en *superpopulation*. I modsætning til en virkelig population kan en superpopulation have uendeligt mange elementer.

I et eksperiment med en superpopulation er udvælgelsesmekanismen abstrakt. Vi kan ikke umiddelbart forestille os kursen på en obligation, en virksomheds resultat eller vejret i morgen som noget, der trækkes op af en høj hat, eftersom superpopulationen ikke eksisterer i virkeligheden. Derfor giver det heller ikke mening at tale om andelsfunktioner i forbindelse med superpopulationer.

3.2.3 Udvælgelsesmekanisme og stikprøve

Når man skal analysere egenskaber for en population baseret på en stikprøve, det vi i kapitel 1 benævnte *induktion*, så er viden om udvælgelsesmekanismen essentiel.

Det følgende kan lyde paradoksalt, men er ikke desto mindre korrekt: Jo mere „tilfældigt" udvælgelsesmekanismen udvælger elementer fra populationen, desto nemmere er det at foretage troværdig induktion. Det følgende eksempel illustrerer dette:

Eksempel 3.7: En usædvanlig terning

Forestil dig, at en af dine „venner" har skrevet nogle nye tal på de seks sider af en terning. Du kender endnu ikke disse tal, men skal forsøge at finde ud af hvilke tal, det drejer sig om ved at observere ham slå med terningen (din stikprøve). De seks tal på terningens sider er derfor de ukendte værdier i populationen.

Betragt først den sædvanlige udvælgelsesmekanisme i forbindelse med et terningspil, nemlig at han putter terningen i et raflebæger og ryster det mange gange, inden han vender det, og den øverste side af terningen afsløres. Hvis han fortsætter således tilpas længe, og du dermed får en stor stikprøve, så er

der meget stor chance for, at du vil kunne slutte dig til, hvilke seks tal, der står på terningens sider. Udvælgelsesmekanismen er her karakteriseret ved at være tilfældig, fordi raflebægeret rystes.

Betragt nu følgende udvælgelsesmekanisme: Din ven snyder og lægger hver gang terningen således, at den samme side vender opad. Baseret på en sådan stikprøve må man konkludere, at der står det samme tal på alle seks sider af terningen, eller at der er bly i den ene ende af terningen. Udvælgelsesmekanismen er her fuldstændig deterministisk og indeholder intet tilfældigt.

I ovenstående eksempel er den tilfældige udvælgelsesmekanisme altså bedre end den deterministiske, fordi den tilfældige udvælgelsesmekanisme altid vil afsløre mindst lige så meget om terningens sider som den deterministiske, der kun fortæller os om én af siderne.

Af mangel på gode danske ord kalder man en udvælgelse, der er baseret på tilfældigheder, som er klart beskrevne (som fx det at man ryster raflebægeret mange gange), for *probabilistisk*. Det modsatte af en probabilistisk udvælgelse kaldes en *systematisk* udvælgelse.

Eksemplet ovenfor er en smule karikeret. Man kan således med rette hævde, at en systematisk udvælgelsesmekanisme, der viser os de seks sider på terningen efter tur i dette tilfælde ville være at foretrække. Imidlertid vil en probabilistisk udvælgelse, der viser os de seks sider i tilfældig orden, være lige så god. Og i de situationer, hvor populationen er meget større end den udtrukne stikprøve, vil en probabilistisk stikprøve være klart at foretrække, som vi skal se senere i denne bog. En vigtig ingrediens i en god empirisk undersøgelse er derfor en udvælgelsesmekanisme, som indeholder en tilfældighedsmekanisme i forbindelse med udvælgelsen af elementer fra en population.

I næste afsnit introducerer vi sandsynlighedsmodellen, som man bruger til at formalisere og regne på „tilfældighederne".

3.3 Sandsynlighedsmodellen

Når vi lægger to tal sammen, er vi ikke nervøse for, om vi får et tal som resultat. Men burde vi egentlig ikke være nervøse? Tal vokser ikke i naturen, de er abstrakte konstruktioner. Børn opdager ikke tal, fordi de kommer til at grave dem op af sandkassen. Der er dog ingen tvivl om nytten af at have tal og af at kunne lægge dem sammen og trække dem fra hinanden. Grunden til, at vi ikke er nervøse for at lave sådanne beregninger med tal, er, at de har et matematisk fundament, som garanterer, at vi kan lægge to tal sammen, uden at vi

behøver at ligge søvnløse over, om resultatet af denne øvelse nu også bliver et tal.

Når vi skal arbejde med usikkerhed, står vi over for en lignende situation. Vi skal have formaliseret usikkerheden således, at vi kan regne på den, som fx sandsynligheden for, at det hverken vil regne eller sne i morgen, eller at kursen på en obligation kommer til at ligge mellem kurs 100 og kurs 105. Et matematisk fundament, som sikrer, at vi kan gøre dette, er *sandsynlighedsmodellen.*

Sandsynlighedsmodellen indeholder den matematiske fortolkning af eksperimentet i figur 3.1, og den giver os mulighed for at regne på den involverede usikkerhed. I det følgende gennemgår vi de tre elementer, som sandsynlighedsmodellen er opbygget af: *Udfaldsrum, hændelsesalgebra og sandsynlighedsmål.* Udfaldsrummet er den direkte matematiske oversættelse af eksperimentet. Hændelsesalgebraen består af de aspekter ved eksperimentet, som vi ønsker at udregne sandsynligheder for. Sandsynlighederne er beskrevet af den tredje del af modellen, sandsynlighedsmålet.

3.3.1 Udfaldsrum

Som beskrevet ovenfor giver et eksperiment anledning til en stikprøve. Et *udfald* er en værdi af en stikprøve. I eksperimentet „at kaste med en terning" er et muligt udfald en „treer". Stikprøven kan altså antage værdien „treer". Til at holde styr på alle udfald definerer man et *udfaldsrum*:

> **Udfaldsrummet, Ω,** for et eksperiment er samlingen af alle mulige udfald af eksperimentet (værdier af stikprøven).

Vi bruger det græske symbol **Ω** (Omega) til at betegne udfaldsrummet. I eksperimentet „at kaste med en terning" er udfaldsrummet: **Ω** = {ener, toer, treer, firer, femmer, sekser}. Tuborgklammerne bruges til at indikere en samling af elementer (udfald), også kaldet en mængde. Resultatet af et eksperiment er ét, og kun ét, af disse elementer, fx en treer.

Eksempel 3.8: Plat og krone

Eksperimentet „at kaste en mønt og observere den øverste side" kan resultere i to mulige udfald: „Plat" og „krone". Udfaldsrummet er derfor **Ω** = {*P, K*}, hvor *P* er „plat", og *K* er „krone". Et mere kompliceret eksperiment kunne være „at kaste en mønt to gange og observere de øverste sider". Et udfald i dette eksperiment består af resultatet for begge kast. Hvis det første kast giver plat og det andet krone, så kan udfaldet skrives (*P, K*). Bemærk, at selvom eksperimentet er sammensat af to kast, så regnes (*P, K*) som ét udfald, ikke som to. Parentesen omkring *P* og *K* markerer, at der er tale om ét sammensat

udfald. Udfaldsrummet for dette eksperiment er: $\Omega = \{(P, P), (P, K), (K, P), (K, K)\}$. Enten viser begge mønter plat, (P, P), eller den første viser plat og den anden krone, (P, K), osv.

Et udfaldsrum består af udfald, der er kombinationer af elementerne i en population. I eksempel 3.8, hvor man kaster en mønt to gange, er populationen $\{P, K\}$, og et udfald er en kombination af disse to populationselementer, fx (P, P) eller (K, P).

Selvom det er fornøjeligt at kaste med mønter og rafle med terninger, er den usikkerhed, vi i dagligdagen er udsat for, af mere kompleks karakter. Den kan imidlertid beskrives med de samme teknikker som mønt- og terningeksperimenterne.

Eksempel 3.9: Salg af paraplyer – del 1

En paraplybutik er interesseret i, hvor mange paraplyer kunderne køber. Den næste kunde, der kommer ind i butikken, kan ende med enten at købe eller ikke købe en paraply. Måske køber han endda to paraplyer, hvis han også vil have en i reserve. Udfaldsrummet for dette eksperiment er derfor: $\Omega = \{0$ paraplyer, 1 paraply, 2 paraplyer$\}$. Med dette valg af udfaldsrum er det på forhånd udelukket, at nogen køber mere end to paraplyer. Hvis det faktisk er et muligt udfald af eksperimentet, skal udfaldsrummet udvides.

Eksempel 3.10: Salg af paraplyer – del 2

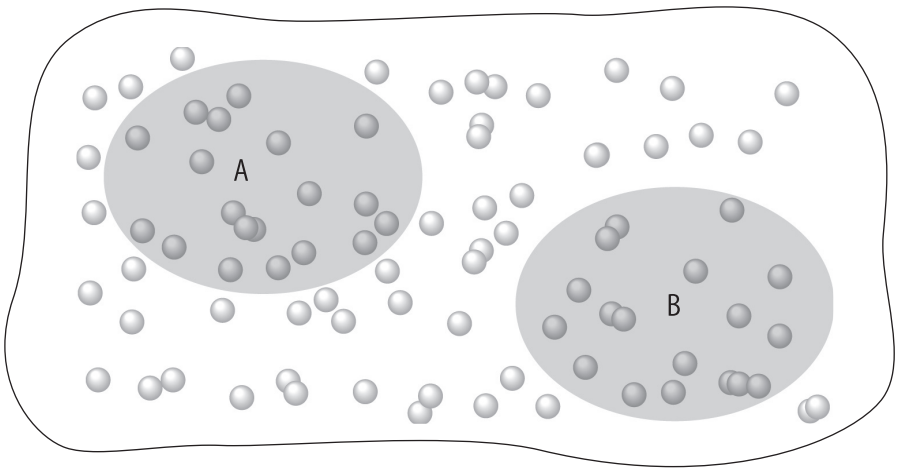
Det kan tænkes, at butikken fra eksempel 3.9 er interesseret i sammenhængen mellem regnvejr og kunders køb af paraplyer. Dette kan også formaliseres som et eksperiment: „at observere om det regner på det tidspunkt, kunden går ind i butikken, samt det antal paraplyer, kunden køber". Udfaldsrummet for dette eksperiment er:

$$\Omega = \{(\text{regn, 0 paraplyer}), (\text{regn, 1 paraply}), (\text{regn, 2 paraplyer}),$$
$$(\text{tørvejr, 0 paraplyer}), (\text{tørvejr, 1 paraply}), (\text{tørvejr, 2 paraplyer})\}$$

Dette er et udfaldsrum, fordi alle mulige udfald af eksperimentet er indeholdt i Ω, og kun ét af dem kan blive resultatet af eksperimentet.

3.3.2 Hændelsesalgebra

Når man analyserer et eksperiment, er man ofte interesseret i en samling af udfald snarere end de enkelte udfald. I eksempel 3.9 kunne man fx være interesseret i, om en kunde køber eller ikke køber paraplyer. Her er det altså udfaldet „0 paraplyer" over for udfaldene „1 paraply" og „2 paraplyer", som er den interessante problemstilling. Man kalder en samling af udfald for en *hændelse:*



Figur 3.2: Udfaldsrum, udfald og hændelser (A og B)

En **hændelse** er en samling af udfald. En hændelse siges at indtræffe, hvis ét af dens udfald indtræffer.

Formelt er en hændelse, A, en delmængde af udfaldsrummet, Ω: $A \subset \Omega$.

I figur 3.2 er sammenhængen mellem udfaldsrum, udfald og hændelser illustreret. Udfaldsrummet er „firkanten", som består af de enkelte udfald, boldene. En hændelse er da en mængde af bolde, illustreret ved „cirklerne" i figuren.

Eksempel 3.11: Et terningkast

I eksperimentet „at kaste med en terning" kan vi definere en hændelse, A, som: „terningen viser et ulige antal øjne". Denne hændelse indtræffer, hvis terningen lander på en ener, en treer eller en femmer. Det kan man formelt skrive som: $A = \{\text{ener, treer, femmer}\}$.

I den første kolonne i tabel 3.1 er udfaldsrummet for paraplyeksperimentet i eksempel 3.9 angivet. I den anden kolonne er vist en række hændelser for dette eksperiment. Bemærk, at tuborgklammerne bruges, da der er tale om mængder af udfald. Den tredje hændelse i tabellen er således „højst købe én paraply". Denne hændelse består af udfaldene „0 paraplyer" og „1 paraply". Samlingen af hændelser kaldes en *hændelsesalgebra*.[1]

1. Samlingen af hændelser skal formelt set opfylde nogle særlige betingelser, for at man kan bruge betegnelsen hændelsesalgebra (også kaldet en sigma-algebra). Det er imidlertid for omfangsrigt at komme ind på detaljerne i disse betingelser her.

Mange hændelser kan man give en naturlig fortolkning, som fx „køb"-hændelsen i tabel 3.1. Andre hændelser har knapt så naturlig en fortolkning, som fx hændelsen, der består af udfaldene „0 paraplyer" og „2 paraplyer". Denne hændelse kan fx kaldes „ikke købe én paraply"-hændelsen. Bemærk, at de enkelte udfald også er hændelser i hændelsesalgebraen. Fx består hændelsen „ingen køb" af udfaldet „0 paraplyer".

Tabel 3.1: Udfaldsrum og hændelsesalgebra

Udfaldsrum, Ω	Hændelsesalgebra	
0 paraplyer	{0 paraplyer}	„ingen køb"
1 paraply	{1 paraply, 2 paraplyer}	„køb"
2 paraplyer	{0 paraplyer, 1 paraply}	„højst købe én paraply"
	{2 paraplyer}	„købe to paraplyer"
	{1 paraply}	„købe én paraply"
	{0 paraplyer, 2 paraplyer}	„ikke købe én paraply"
	{0 paraplyer, 1 paraply, 2 paraplyer}	„den sikre hændelse"
	$\varnothing$	„den umulige hændelse"

Blandt hændelserne i tabel 3.1 findes to bemærkelsesværdige mængder. Den første er mængden af alle udfald, som i tilfældet med paraplyer er: {0 paraplyer, 1 paraply, 2 paraplyer}. Ligegyldigt hvilket udfald der indtræffer, så indtræffer denne hændelse altid. Derfor kaldes den også for *den sikre hændelse*. Vi anvender symbolet Ω om den sikre hændelse, fordi den indeholder alle udfald i udfaldsrummet. Det modsatte af den sikre hændelse er *den umulige hændelse*. Den er karakteriseret ved, at ingen udfald tilfredsstiller den. Den betegnes med det specielle symbol $\varnothing$.

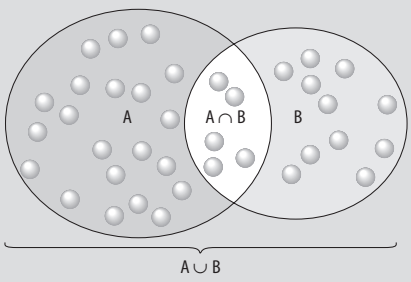
For at kunne beskrive om to hændelser indtræffer samtidig, eller om mindst én af to hændelser indtræffer, definerer man såkaldte *fælleshændelser og foreningshændelser*. De er defineret i nedenstående boks og illustreret i eksempel 3.12.

Fælleshændelser og foreningshændelser:

Lad A og B være to hændelser.

$A \cap B$ kaldes fælleshændelsen og består af de udfald, som indgår i både A og B.

$A \cup B$ kaldes foreningshændelsen og består af de udfald, som indgår i enten A eller B (eller begge).



Et venn-diagram

Eksempel 3.12: Salg af paraplyer – del 3

Betragt hændelserne „køb" og „højst købe én paraply" i paraplyeksperimentet fra tabel 3.1. Fælleshændelsen for disse to hændelser består af udfaldet „1 paraply", da det er det eneste udfald, som findes i begge mængder i tabel 3.1:

$$\text{„køb"} \cap \text{„højst købe én paraply"} = \{1 \text{ paraply}\}$$

Foreningshændelsen for de to hændelser er derimod givet ved:

$$\text{„køb"} \cup \text{„højst købe én paraply"} = \{0 \text{ paraplyer}, 1 \text{ paraply}, 2 \text{ paraplyer}\}$$

og findes ved at tage alle de udfald, der findes i de to hændelser. Bemærk, at udfaldet „1 paraply" findes i begge hændelser. Det indgår dog kun én gang i foreningshændelsen.

Forskellige hændelser kan altså indtræffe samtidig, hvis de har udfald til fælles, hvorimod to hændelser er gensidigt udelukkende: De kan ikke begge indtræffe samtidig, hvis de ingen udfald har til fælles. Figuren i boksen kaldes iøvrigt et Venn diagram.

Til enhver hændelse er der en modsat hændelse. Den modsatte hændelse af hændelsen A er defineret som mængden af alle de udfald, som ikke er i A. Den modsatte hændelse kaldes også *den komplementære hændelse* og betegnes her med toptegnet c (for „complementary"). Den modsatte hændelse af A betegnes derfor A^c. De to hændelser er ikke blot gensidigt udelukkende, man ved også, at enten indtræffer den ene hændelse, eller også indtræffer den anden. Formelt gælder derfor:

$$A \cap A^c = \emptyset \quad og \quad A \cup A^c = \Omega$$

3.3.3 Sandsynlighedsmål

For at kunne måle usikkerhed kan man tilskrive sandsynligheder til de forskellige hændelser. Sandsynligheden angiver chancen for, at hændelsen vil indtræffe. I det følgende ser vi på nogle forskellige fortolkninger af begrebet sandsynlighed samt måder, hvorpå man kan tilskrive sandsynligheder til hændelser.

En måde at fortolke sandsynlighed på er som den andel af gange, en hændelse vil indtræffe, hvis man gentager eksperimentet uendeligt mange gange. Dette er naturligvis en abstraktion, da man jo ikke kan gentage et eksperiment uendeligt mange gange, medmindre man er meget optimistisk med hensyn til sin levealder. Men abstraktionen virker alligevel. Tag eksperimentet „at kaste en mønt". Hvis vi kaster mønten mange gange, vil vi forvente, at den cirka halvdelen af gangene viser plat. Hvorfor? Prøv at vende argumentet om: Hvor-

for ikke? Det er svært at argumentere for, at den ikke skulle vise plat cirka halvdelen af gangene. I så fald må der være et eller andet, som får mønten til at vise fx plat oftere end krone. Mønten kunne være bøjet, man kaster ikke mønten højt nok og med tilstrækkeligt spind, eller indgraveringen af Dronning Margrethe giver mere luftmodstand end indgraveringen på den anden side af mønten. Men hvis vi kan tilbagevise disse forhold som værende irrelevante for eksperimentets udfald, så står vi tilbage uden argumenter for, at vi ikke skulle få plat cirka halvdelen af gangene.

Nøglen til denne konklusion er symmetri: Alt andet lige, så har alle udfald samme chance for at indtræffe. Derfor tilskriver vi sandsynligheden „1 divideret med antallet af mulige udfald" til hvert udfald. Denne simple nøgle kan løse mange komplekse problemer omkring tilskrivning af sandsynligheder til hændelser. Den er opsummeret i nedenstående boks. Bemærk, at bogstavet P (for „probability") bruges til at angive sandsynlighed.

Tilskrivning af sandsynlighed ved hjælp af symmetri:
Hvis et udfaldsrum består af K udfald, $\Omega = \{\omega_1, \ldots, \omega_k\}$, som alle har lige stor chance for udvælgelse, så er sandsynligheden for enhver hændelse med kun ét udfald lig med $\frac{1}{K}$:

$$P(\omega_1) = \cdots = P(\omega_K) = \frac{1}{K}$$

Sandsynligheden for en hændelse, A, der består af k udfald, er:

$$P(A) = \frac{k}{K}$$

Dette symmetriargument kan man udvide en smule, når eksperimentet involverer en virkelig population. Forestil dig, at vi har en virkelig population bestående af N_{pop} elementer, hvoraf k elementer alle har værdien z af et karakteristikum. Dette kunne fx være populationen af danskere, hvor der er ca. 2,5 millioner elementer, der har værdien, z = *„mand"*, dvs. k = 2,5 millioner og N_{pop} = 5,0 milioner. De k elementer med værdien z udgør andelen k/N_{pop} af populationen. Andelsfunktionen er derfor lig med k/N_{pop} i punktet z: $g(z) = k/N_{pop}$. Hvis alle elementer i populationen har samme chance for udvælgelse, og udfaldet af eksperimentet er givet ved værdien af det udtrukne element, så er sandsynligheden for, at eksperimentet resulterer i udfaldet z, lig med andelen k/N_{pop}. Dvs. der er sandsynlighed 2,5 mill./5,0 mill. = 0,5 for at udvælge en mand fra populationen af danskere. Dette princip er opsummeret i følgende boks.

Tilskrivning af sandsynlighed ved hjælp af andelsfunktionen:
Eksperiment: Udtrækning af et element fra en virkelig population med N_{pop} elementer. Udfaldet af eksperimentet er lig med værdien af det udtrukne element.

Hvis alle elementer i populationen har lige stor chance for udvælgelse, så er sandsynligheden for udfaldet z lig med værdien af andelsfunktionen evalueret i z:

$$P(z) = g(z) = \frac{\text{antal elementer med værdien } z \text{ i populationen}}{N_{pop}}$$

Bemærk forskellen på de to ovenstående bokse. Den første boks er relevant i en situation, hvor alle K udfald af eksperimentet er lige sandsynlige. Dette udnytter man til at tilskrive hvert udfald sandsynligheden $1/K$. Dette er fx tilfældet med et terningkast, hvor hver af de seks sider har sandsynligheden 1/6. I den anden boks behøver de forskellige udfald af eksperimentet ikke at være lige sandsynlige. Her går vi i stedet helt tilbage til populationen og siger, at hvis alle elementer i populationen har lige stor chance for udvælgelse, så kan vi bruge andelen af elementer med værdien z som sandsynligheden for udfaldet z. Hvis vi fx udtrækker én person i en population bestående af 1000 mænd og 500 kvinder, så er sandsynligheden $1000/(1000 + 500) = 2/3$ for, at vi får en mand, hvis alle personer har samme chance for at blive udtrukket.

Chancen for, at det regner i morgen, er 50 %. Her kan vi ikke appellere til symmetri eller gentage eksperimentet „vejret i morgen" uendeligt mange gange, ja ikke engang to gange. Men erfaring kan hjælpe os i dette tilfælde. Der har tidligere været dage med cirka samme temperatur, vindretning, luftfugtighed og skydække, hvor det dagen efter regnede omkring halvdelen af gangene. Vi kan også bruge viden fra forsøg i et laboratorium til at forudsige chancen for skydannelse og dermed regn. Denne viden er ofte samlet i matematiske modeller, som man fx kan bruge til at simulere vejrudviklingen på computere. I kapitlerne om regressionsanalyse senere i bogen gennemgår vi matematiske modeller, som kan være yderst nyttige til forudsigelser. Idéen er, at fortidige observationer bruges i videnskabelige modeller til at forudsige fremtidige situationer og sandsynligheder.

Erfaring kan også bruges til forudsigelser, som ikke er baseret på klare matematiske modeller, som eksempel 3.13 viser.

Eksempel 3.13: Salg af paraplyer – del 4

I paraplybutikken ved man af erfaring, at andelen af kunder, som ikke køber, er 1/4, mens halvdelen af kunderne køber én paraply, og en fjerdel af dem køber to paraplyer. Hvis vi observerer en tilfældig kunde, der kommer ind i butikken, så kan vi bruge den erfarede andel som mål for sandsynligheden for, at kunden

ikke vil købe en paraply. Sandsynligheden for, at kunden ikke køber noget, er således 1/4. Tilsvarende er sandsynligheden for, at kunden køber 1 paraply lig med 1/2, og sandsynligheden for, at han køber 2 paraplyer, er lig med 1/4.

Hvad er chancen for, at hesten Ibrahim vinder på søndag? For nogle vil jockeyens stjernetegn og sangstemme indgå i chancevurderingen. Matematisk er der ikke noget problem i, at man tilskriver sandsynligheder efter forgodtbefindende. Ofte kan man kun støtte sig til sin intuition, uden at man skal være blind for, at intuitionen delvist er et resultat af ens erfaringer. I alle disse tilfælde siger man, at sandsynlighederne er fremkommet subjektivt.

Formelt er *sandsynlighedsmålet* en funktion af en hændelse. Dette gælder uanset, om man tilskriver sandsynlighed vha. symmetriargumenter, ud fra erfaring eller subjektivt.

Tilskrivningen af sandsynligheder til alle hændelserne i hændelsesalgebraen skal opfylde nogle få krav, for at dette matematiske system kan virke. Disse krav er som følger:

Et sandsynlighedsmål, $P(\ )$, skal opfylde**:**

(i) $0 \leq P(A) \leq 1$, hvor A er en hændelse.

(ii) $P(\Omega) = 1$, hvor Ω er den sikre hændelse.

(iii) $P(A \cup B) = P(A) + P(B)$, hvis hændelserne A og B ikke har udfald til fælles, dvs. $A \cap B = \varnothing$.

Disse tre krav medvirker til, at systemet af sandsynligheder er konsistent, dvs. ikke indeholder selvmodsigelser.[2] For det første må udfald ikke have negative sandsynligheder eller sandsynligheder større end 1. For det andet er sandsynligheden for den sikre hændelse lig med 1, fordi denne hændelse altid indtræffer, idet den består af alle udfald i udfaldsrummet, Ω. Endelig er det således, at hvis to hændelser ikke har nogen udfald til fælles, så er sandsynligheden for foreningshændelsen lig med summen af sandsynlighederne for de to hændelser. At A og B ikke har udfald til fælles betyder, at fællesmængden, $A \cap B$, er tom. Som illustreret i det følgende eksempel tilfredsstiller de tilskrevne sandsynligheder i tabel 3.2 de ovenstående tre betingelser.

---

2. Formelt set kræves det også, at $P(A1 \cup A2 \cup A3 \cup \ldots) = P(A1) + P(A2) + P(A3) + \ldots$, hvis ingen af hændelserne, $A1, A2, A3, \ldots$, har nogle udfald til fælles. Dette vil dog typisk være opfyldt, hvis betingelserne (i)-(iii) fra boksen ovenfor er opfyldt.

Eksempel 3.14: Salg af paraplyer – del 5

I eksempel 3.13 var sandsynlighederne for køb af 0, 1 og 2 paraplyer henholdsvis 1/4, 1/2 og 1/4. Disse sandsynligheder kan vi nu bruge til at udregne sandsynlighederne for de resterende hændelser. Fx består hændelsen „køb" af de to udfald: „1 paraply" og „2 paraplyer", som har sandsynlighederne 1/2 og 1/4. Sandsynligheden for hændelsen „køb" kan da udregnes som summen af disse sandsynligheder: 1/2 + 1/4 = 3/4. Tabel 3.2 viser de tilskrevne sandsynligheder for de øvrige hændelser i paraplyeksperimentet.

Tabel 3.2: Udfaldsrum, hændelsesalgebra og sandsynlighedsmål

Udfaldsrum, Ω	Hændelsesalgebra		Sandsynlighedsmål
0 paraplyer	{0 paraplyer}	„ingen køb"	¼
1 paraply	{1 paraply, 2 paraplyer}	„køb"	¾
2 paraplyer	{0 paraplyer, 1 paraply}	„højst købe én paraply"	¾
	{2 paraplyer}	„købe to paraplyer"	¼
	{1 paraply}	„købe én paraply"	½
	{0 paraplyer, 2 paraplyer}	„ikke købe én paraply"	½
	{0 paraplyer, 1 paraply, 2 paraplyer}	„den sikre hændelse"	1
	∅	„den umulige hændelse"	0

3.4 Regneregler for sandsynligheder

En implikation af de tre egenskaber ved sandsynligheder i ovenstående boks er *additionssætningen,* som angiver sammenhængen mellem sandsynlighederne for foreningshændelsen og fælleshændelsen:

Additionssætningen for sandsynligheder:

For hændelser A og B gælder:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Additionssætningen fortæller os, at sandsynligheden for en foreningshændelse er lig med summen af sandsynlighederne for de to involverede hændelser fratrukket sandsynligheden for deres fælleshændelse. Årsagen til, at denne skal fratrækkes, er, at sandsynlighederne for udfald, der indgår i både A og B, er medtaget i både $P(A)$ og $P(B)$. For ikke at tælle sandsynlighederne for disse udfald med to gange trækker vi dem fra via $P(A \cap B)$.

Eksempel 3.15: Salg af paraplyer – del 6

Hændelserne „køb“ og „højst købe én paraply“ i paraplyeksperimentet fra tabel 3.2 har begge sandsynligheden 3/4. For at finde sandsynligheden for deres foreningshændelse, „køb“ ∪ „højst købe én paraply“, skal vi lægge deres individuelle sandsynligheder sammen og trække sandsynligheden for fælleshændelsen fra. Fælleshændelsen består af udfaldet „1 paraply“, som har sandsynligheden 1/2. Sandsynligheden for foreningshændelsen „køb“ ∪ „højst købe én paraply“ er derfor: 3/4 + 3/4 – 1/2 = 1. Dette passer også fint, idet vi fra eksempel 3.12 ved, at:

„køb“ ∪ „højst købe én paraply“ = {0 paraplyer, 1 paraply, 2 paraplyer}

som er hele udfaldsrummet, Ω, og $P(\Omega)$ er jo netop 1.

For en hændelse A og dens modsatte hændelse A^c gælder, at sandsynligheden for, at mindst én af dem indtræffer, er lig med 1. Dette skyldes, at foreningshændelsen mellem de to hændelser er den sikre hændelse. Formelt er:

$$P(A \cup A^c) = P(\Omega) = 1$$

Desuden er A og A^c gensidigt udelukkende, så:

$$P(A \cap A^c) = P(\varnothing) = 0$$

Som en konsekvens af dette får man ved hjælp af additionssætningen for sandsynligheder:

$$P(A) = 1 - P(A^c)$$

Man kan desuden finde sandsynligheden for en hændelse, A, hvis man kender sandsynligheden for fælleshændelsen af A og B (hvor B er en hvilken som helst hændelse) samt sandsynligheden for fælleshændelsen af A og B^c. Da har man nemlig:

$$P(A) = P(A \cap B) + P(A \cap B^c)$$

Eksempel 3.16: Salg af paraplyer – del 7

Lad A være hændelsen, at en kunde køber én paraply, og lad B være hændelsen, at det regner i eksperimentet fra eksempel 3.10. Antag, at sandsynligheden for, at en kunde køber én paraply, er $P(A) = 0{,}15$, og at sandsynligheden for, at det regner, er $P(B) = 0{,}25$. Så kan vi finde sandsynlighederne for de modsatte hændelser, A^c og B^c. Hvis vi også ved, at sandsynligheden for, at en kunde køber én paraply, og det samtidig regner, er $P(A \cap B) = 0{,}1$, da kan vi også finde sandsynlighederne for de øvrige fælleshændelser, der involverer A og B samt A^c og B^c. Dette er gjort i nedenstående tabel:

For eksempel er sandsynligheden for, at kunden køber én paraply og det ikke regner, givet ved:

$$P(A \cap B^c) = P(A) - P(A \cap B) = 0{,}15 - 0{,}1 = 0{,}05$$

Tabel 3.3: Hændelser, modsatte hændelser, fælleshændelser og sandsynligheder

Hændelser	B: (regn)	B^c: (ikke-regn)	
A: (1 paraply)	$P(A \cap B) = 0{,}1$	$P(A \cap B^c) = 0{,}05$	$P(A) = 0{,}15$
A^c: (ikke 1 paraply)	$P(A^c \cap B) = 0{,}15$	$P(A^c \cap B^c) = 0{,}7$	$P(A^c) = 0{,}85$
	$P(B) = 0{,}25$	$P(B^c) = 0{,}75$	$P(\Omega) = 1$

3.5 Betinget sandsynlighed

I mange situationer kan vi anvende information om en hændelse, der er indtruffet, til bedre at kunne forudsige, hvorvidt en anden hændelse indtræffer. Hvis vi fx bruger information om, hvor meget det har regnet i løbet af året, så kan vi bedre forudsige høstudbyttet, end hvis vi ikke bruger denne information. Et andet eksempel er risikoen for at få en arvelig sygdom. Hvis vi kender sygdomshistorien for en given persons forældre, så kan denne bruges til bedre at forudsige personens egen risiko for at få en arvelig sygdom. Teorien om betingede sandsynligheder handler om, hvordan man kan bruge information om én hændelse til at vurdere sandsynligheden for en anden hændelse.

I nogle tilfælde vil information om, hvorvidt én hændelse indtræffer, fjerne usikkerheden omkring, hvorvidt en anden hændelse indtræffer. Dette er tilfældet, når to hændelser, A og B, ikke har nogen fælles udfald (A og B er gensidigt udelukkende). Hvis vi ved, at hændelse B er indtruffet, dvs. ét af udfaldene i B er indtruffet, så er sandsynligheden for, at hændelse A indtræffer, lig med nul.

Vi kalder sandsynligheden for, at hændelsen A indtræffer, givet vi ved, at hændelsen B indtræffer (eller er indtruffet) for *den betingede sandsynlighed for hændelse A givet hændelse B*. Denne sandsynlighed symboliseres med $P(A|B)$, hvor den lodrette streg betyder „givet“ eller „betinget på“.

Den betingede sandsynlighed, $P(A|B)$, udregnes på følgende måde. Når hændelsen B indtræffer, så betyder det, at vi kan koncentrere os om de udfald, som indgår i hændelsen B. Af disse udfald er nogle i A, mens andre ikke er i A. Til at udregne den betingede sandsynlighed skal vi derfor bruge sandsynlighe-

den for de udfald, som indgår i både A og B. Denne sandsynlighed er $P(A \cap B)$. Da vi ved, at hændelsen B indtræffer (eller allerede er indtruffet), så skal sandsynligheden $P(A \cap B)$ ses i forhold til sandsynligheden for hændelsen B, dvs. $P(B)$. Dette er intuitionen bag definitionen af den betingede sandsynlighed, som præsenteres i den følgende boks:

Den betingede sandsynlighed for hændelsen A givet hændelsen B er defineret som:

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

forudsat at sandsynligheden for hændelsen B er strengt større end 0, $P(B) > 0$.

Når man ved, at hændelsen B er indtruffet, så kan man tænke på udfaldene i B som et udfaldsrum. Inden for dette udfaldsrum skal man finde sandsynligheden for de udfald, som også tilhører A.

Eksempel 3.17: Salg af paraplyer – del 8

I eksempel 3.16 var sandsynligheden for hændelsen A (at en kunde køber én paraply) givet ved $P(A) = 0{,}15$, mens sandsynligheden for hændelsen B (at det regner), var $P(B) = 0{,}25$. Endelig var sandsynligheden for fælleshændelsen (at en kunde køber én paraply og det regner) givet ved $P(A \cap B) = 0{,}1$. Sandsynligheden for, at en kunde køber én paraply, givet at det regner, er da:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{0{,}1}{0{,}25} = 0{,}4$$

Her ses værdien af informationen om hændelsen B. Hvis vi ikke ved, om det regner, så er vores bedste bud på sandsynligheden for, at en kunde køber én paraply, lig med 0,15. Hvis det derimod regner (hændelsen B er indtruffet), så er sandsynligheden for køb faktisk betydeligt højere, nemlig 0,4.

Man kan udvide begrebet betinget sandsynlighed til en situation, hvor man har viden om flere hændelser. Her definerer man blot den betingede sandsynlighed som sandsynligheden for en hændelse, givet at flere andre hændelser indtræffer (eller er indtruffet). $P(A \mid B, C)$ er således sandsynligheden for hændelsen A, givet at både hændelsen B og hændelsen C indtræffer. Denne sandsynlighed er formelt set givet ved:

$$P(A \mid B, C) = P(A \mid B \cap C))$$

Dvs. som sandsynligheden for hændelsen A givet fælleshændelsen mellem B og C. Hvis vi tænker på denne fælleshændelse som en ny hændelse, D, hvor $D = B \cap C$, så er $P(A \mid B, C) = P(A \mid D)$, hvilket er en sandsynlighed betinget på kun én hændelse ligesom i den oprindelige definition af betinget sandsynlighed. Derfor kan $P(A \mid B, C)$ udregnes som:

$$P(A \mid B, C) = P(A \mid D) = \frac{P(A \cap D)}{P(D)} = \frac{P(A \cap B \cap C)}{P(B \cap C)}$$

3.6 Uafhængighed og spuriøse sammenhænge

En af de største empiriske udfordringer inden for samfundsvidenskab er at undersøge sammenhænge mellem forskellige størrelser og afgøre, om sammenhængene er *kausale*. En kausal sammenhæng mellem to størrelser betyder, at den ene størrelse forårsager den anden størrelse. En forståelse af, om en sammenhæng er kausal eller ej, er essentiel, hvis man fx vil bruge resultaterne af en undersøgelse til at formulere en ny politik på området. Hvis man fx vil undgå, at unge mennesker bliver kriminelle, så er det vigtigt at forstå, hvad der forårsager kriminalitet.

Det er værd at gøre opmærksom på, at kausalitet er et begreb, som ikke er præcist og entydigt defineret. Begrebet har været genstand for diskussion blandt filosoffer i hvert fald de seneste 2500 år. Vi vil derfor afholde os fra at give en præcis definition af kausalitet i denne bog. Løst sagt har vi i tankerne, at en hændelse A forårsager en hændelse B, hvis A indtræffer først og påvirker sandsynligheden for, at B indtræffer. Et væsentligt element i kausalitet er således, at den forårsagende hændelse tidsmæssigt skal forekomme før hændelsen, som forårsages.

Vi vil nu vise, hvordan vi kan bruge redskaberne fra de første afsnit i dette kapitel til at opbygge en ramme, inden for hvilken vi kan undersøge sammenhænge mellem forskellige størrelser under usikkerhed. I dette afsnit definerer vi således, hvad vi statistisk set forstår ved henholdsvis uafhængige og afhængige hændelser. Vi viser desuden, hvorfor afhængighed mellem to hændelser ikke nødvendigvis er det samme som en kausal sammenhæng mellem de to hændelser. Dermed kommer vi tættere på at kunne afgøre, om en given sammenhæng er kausal eller ej.

3.6.1 Uafhængighed

Det centrale statistiske begreb, når man skal se på sammenhænge mellem to eller flere hændelser, er *uafhængighed*. Hvis information om, at en hændelse, B, er indtruffet, ikke ændrer sandsynligheden for, at en anden hændelse, A, indtræffer, så siger vi, at de to hændelser er uafhængige. Formelt svarer dette

til, at: $P(A|B) = P(A)$. Sandsynligheden for hændelse A er således den samme, uanset om vi medtager information om, hvorvidt B er indtruffet, $P(A|B)$, eller vi ikke medtager denne information, $P(A)$.

Det følger da, at hvis to hændelser, A og B, er uafhængige, så er sandsynligheden for fælleshændelsen mellem A og B givet ved:

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

hvor det første lighedstegn følger af definitionen af betinget sandsynlighed fra boksen ovenfor (når der ganges med $P(B)$ på begge sider af lighedstegnet), mens det andet lighedstegn følger af, at de to hændelser er uafhængige og dermed at: $P(A|B) = P(A)$. Vi opsummerer definitionen af uafhængighed i følgende boks:

Uafhængighed: To hændelser, A og B, er uafhængige, hvis og kun hvis:

$$P(A|B) = P(A)$$

hvilket er ensbetydende med:

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

Hvis betingelserne ovenfor ikke er opfyldt, så siges hændelserne at være afhængige. Fordi to hændelser er statistisk afhængige, så betyder det imidlertid ikke, at de to hændelser direkte har noget med hinanden at gøre. I flere lande har det fx vist sig, at der er en statistisk afhængighed mellem antallet af observerede storke og antallet af børnefødsler. Det betyder dog ikke, at det er storken, der kommer med børnene. Hvis to hændelser omvendt er uafhængige, så kan man dog som regel med sindsro konkludere, at de to hændelser ikke har noget med hinanden at gøre.

Når to hændelser ikke har noget med hinanden at gøre, men alligevel statistisk set er afhængige, så skyldes det andre hændelser, som ikke er medtaget i analysen. Fx skyldes sammenhængen mellem storke og børnefødsler den generelle samfundsudvikling med overgangen fra landbrugssamfund til industri- og servicesamfund, som både har ødelagt en del af storkenes oprindelige omgivelser og samtidig har mindsket antallet af børnefødsler. Med andre ord, hvis man tager højde for samfundsudviklingen, så forsvinder sammenhængen mellem antallet af storke og antallet af børnefødsler.

Til at betegne, at en tredje hændelse, C, kan forklare en tilsyneladende sammenhæng mellem to hændelser, A og B, bruger man begrebet *betinget uafhængighed*.

Betinget uafhængighed: Hændelserne A og B er uafhængige betinget på hændelsen C, hvis og kun hvis:

$$P(A \mid B, C) = P(A \mid C)$$

Betinget uafhængighed mellem A og B betyder, at sandsynligheden for hændelsen A ikke påvirkes af, om hændelsen B indtræffer, hvis man allerede ved, om hændelsen C er indtruffet eller ej.

3.6.2 Direkte, indirekte og spuriøse sammenhænge

Der findes utallige empiriske undersøgelser, der konkluderer, at der statistisk set er en sammenhæng mellem to størrelser. Vi skal senere i denne bog se, hvordan man kan lave sådanne undersøgelser. Når man hører, at der er en sammenhæng mellem to størrelser, så bør man imidlertid altid overveje, om undersøgelsen har korrigeret for andre størrelser, som kan tænkes at udgøre den bagvedliggende årsag til den fundne sammenhæng.

Der er grundlæggende tre situationer, hvor hændelserne A og B statistisk set udviser afhængighed. De tre situationer er:

1. A og B *afhænger direkte* af hinanden: $A \rightarrow B$ eller $B \rightarrow A$.
2. A og B *afhænger indirekte* af hinanden gennem en hændelse C: $A \rightarrow C \rightarrow B$ eller $B \rightarrow C \rightarrow A$.
3. A og B *afhænger ikke* af hinanden, men begge afhænger af en tredje hændelse C: $A \leftarrow C \rightarrow B$.

I alle tre situationer er $P(A \mid B) \neq P(A)$. Når man siger, at A og B afhænger af hinanden, så er det typisk underforstået, at det er situation 1, man mener, gør sig gældende. Det er også typisk underforstået, at en af hændelserne forårsager den anden hændelse. Men baseret på statistisk afhængighed mellem A og B kan det lige så godt være situation 3, som er gældende. I så fald er der reelt ingen afhængighed mellem A og B i den forstand, at en af dem forårsager den anden. Man kalder en sådan tilsyneladende sammenhæng for en *spuriøs sammenhæng*.

Man kan afsløre, om man befinder sig i situation 1 eller i en af de to andre situationer ved at undersøge, om der findes en hændelse, C, hvorom det gælder, at A og B er uafhængige betinget på C: $P(A \mid B, C) = P(A \mid C)$. I så fald er det en af de to sidste situationer, der gør sig gældende. For at kunne afgøre, om dette er tilfældet, skal man imidlertid have et godt bud på, hvad C kan være.

I praksis fremgår det typisk af en undersøgelses formål, hvad hændelserne A og B repræsenterer. Hændelsen C er derimod mindre oplagt. Det er her, man skal bruge sit kendskab til en undersøgelses problemstilling samt sin sunde fornuft til at vurdere, hvad hændelsen C kan være. Det følgende eksempel illustrerer dette.

Eksempel 3.18: Kvalitet af skoler

Forestil dig et eksperiment, hvor man udvælger en skole tilfældigt og observerer, om den er privatejet (B) eller ej (B^c), og om karaktergennemsnittet på skolen er højt (A) eller ej (A^c). Det viser sig, at $P(A|B) = 0{,}7$ mens $P(A|B^c) = 0{,}4$. Dermed kan man konkludere, at hændelserne A og B ikke er uafhængige statistisk set. Man kan dernæst spørge sig selv, hvad årsagen til denne sammenhæng er. Skyldes den forskellig kvalitet af undervisningen på de to typer af skoler?

Forskning tyder på, at en elevs præstation i skolen afhænger af, hvor meget fx forældre har haft tid til at stimulere eleven sprogligt de første seks leveår. Lad hændelsen C betegne, om skolens elever har modtaget tilstrækkelig sproglig stimulering de første seks leveår. Sandsynligheden for, at karaktergennemsnittet i en skole bliver højt, givet at eleverne har modtaget sproglig stimulering, antages at være: $P(A|C) = 0{,}8$.

Man kan dernæst finde sandsynligheden for, at karaktergennemsnittet er højt betinget både på typen af skole og mængden af stimulering. Hvis man her finder, at $P(A|B, C) = 0{,}8$, så er $P(A|B, C) = P(A|C)$, og A og B er da uafhængige betinget på C. Det betyder, at der ikke er en direkte sammenhæng mellem A og B. I stedet kan sammenhængen mellem typen af skole og karaktergennemsnittet være spuriøs. Det kan fx skyldes, at velstimulerede børn i højere grad sendes på privatskole, hvilket igen kan hænge sammen med, at velstimulerede børn kommer fra familier med relativt høje indkomster.

Som eksemplet viser, så kan man nogle gange vise, at en tilsyneladende afhængighed mellem to hændelser, A og B, ikke holder, hvis man tager højde for en tredje hændelse, C. Ofte kan det dog være svært at vide, hvad denne tredje hændelse skal være, og man kan da forsøge sig med forskellige bud på C. Og selvom man ikke finder en tredje hændelse, der kan forklare sammenhængen mellem A og B, så er det jo ikke det samme som, at en sådan hændelse ikke findes.

Hvis man finder, at C kan forklare sammenhængen mellem A og B, så kan man derfor heller ikke umiddelbart konkludere, at der er en direkte afhængighed mellem C og A og mellem C og B. Det kan være, at en helt fjerde hændelse, D, kan forklare begge sammenhænge. Som nævnt i eksemplet kan det

være familiens socioøkonomiske status, der bestemmer den sproglige stimulering, valget af skole samt karaktergennemsnittet. Dette illustrerer også, hvor vanskeligt det kan være at afdække direkte afhængighed og i sidste ende kausale sammenhænge inden for samfundsvidenskaberne.

3.7 Opgaver

1) Repetitionsspørgsmål:
   a) Hvad består et eksperiment af?
   b) Giv eksempler på virkelige populationer og superpopulationer.
   c) Forklar kort de tre elementer i sandsynlighedsmodellen: Udfaldsrum, hændelser og sandsynligheder.
   d) Gør rede for sammenhængen mellem udfaldsrum og stikprøve.
   e) Forklar kort om de forskellige måder at tilskrive sandsynligheder på.
   f) Hvilke tre egenskaber skal sandsynligheder opfylde?
   g) Gør rede for additionssætningen for sandsynligheder.
   h) Forklar hvad man forstår ved betinget sandsynlighed.
   i) Hvad betyder det, at to hændelser er henholdsvis uafhængige og betinget uafhængige?

2) En håndboldanalytiker vurderer, at hjemmeholdet til en forestående kamp har 40% chance for at vinde og 70% chance for ikke at tabe.
   a) Hvad er populationen, stikprøven og udfaldsrummet i dette eksperiment?
   b) Forklar ud fra egenskaberne ved sandsynligheder, hvad sandsynligheden er for et uafgjort resultat.
   c) Hvad er sandsynligheden for, at et af holdene vinder?

3) I en aftenskoleklasse for orgelspil er tre kursister under 40 år, fire er mellem 40 og 65 år og to er over 65 år. Der skal trækkes lod om en fribillet til en Ole Erling-koncert. Alle navnene på de ni personer lægges i en spand, som rystes, og ét navn udvælges tilfældigt.
   a) Hvad er populationen og den tilhørende andelsfunktion?
   b) Hvad er sandsynligheden for hændelsen, at en kursist mellem 40 og 65 år udvælges?
   c) Hvad er sandsynligheden for hændelsen, at en kursist over 40 år udvælges?

4) Foulum Fodbold Uniteds (FFU) første og eneste hold deltager i Viborgmesterskaberne. Der spilles tre kampe i indledende runde. Sandsynligheden for at Foulum taber en kamp er 0,9, og resultaterne af de tre kampe er uafhængige.
   a) Hvad er sandsynligheden for at FFU ikke taber deres anden kamp?
   b) Hvad er sandsynligheden for, at FFU taber alle tre kampe?
   c) Hvad er sandsynligheden for, at FFU taber to kampe?
   d) Hvad er sandsynligheden for, at FFU ikke har tabt anden kamp givet at man kun ved, at FFU har tabt to kampe i indledende runde? Sammenlign svaret med dit svar til spørgsmål a).