



AALBORG UNIVERSITET  
STUDENTERRAPPORT

AALBORG UNIVERSITET

KANDIDATPROJEKT I STATISTIK

---

# Kausal inferens

---

Christoffer GREBJERG  
Søren DAMGAARD NIELSEN

Under vejledning af  
Anne MARIE SVANE



**AALBORG UNIVERSITET**  
STUDENTERRAPPORT

**School of Engineering and Science**

Institut for matematiske fag

Skjernvej 4a

9000 Aalborg

<http://www.math.aau.dk/>

**Titel:**

Kausal inferens

**Tema:**

Statistik

**Projekt:**

Kandidatprojekt

**Projektperiode:**

Forår 2022

**Projektgruppe:**

5.213b

**Gruppemedlemmer:**

Christoffer Grebjerg

Søren Damgaard Nielsen

**Vejleder:**

Anne Marie Svane

**Sidetæl: 85**

**Afsluttet den 3/6-2022**

**Synopsis:**

I dette projekt bliver teori omkring kausal inferens gennemgået. Dette omhandler den gennemsnitlige kausale effekt, kausalitet versus association, randomiserede eksperimenter, standardisering og invers sandsynlighedsvægtning (IP vægtning). Der ses også på begreberne ombyttelighed, positivitet og konsistens af observationsstudier, som benyttes til kausal inferens. Derefter ses der på, hvordan disse modeller grafisk kan afbildes, og hvilke egenskaber, der kommer fra dette. Til sidst behandles et datasæt, hvor den gennemsnitlige kausale effekt af ryggestop på BMI-forskel estimeres.

Dette projekt er udarbejdet som specialeprojekt på 10. semester for kandidatuddannelsen i Matematik på Aalborg Universitet.

Det er udarbejdet på matematikstudiet under Institutet for Matematiske Fag under Det Teknisk-Naturvidenskabelige Fakultet i perioden 1. februar 2022 til 3. juni 2022.

Projektet er forfattet af Christoffer Grebjerg og Søren Damgaard Nielsen.

Projektets overordnede tema er kausal inferens, og herunder behandles den teori, som ligger til grund for anvendelse af data.

Rapporten er opbygget af nummererede kapitler med afsnit og underafsnit. Definitioner, eksempler, propositioner og sætninger er nummereret efter kapitler og fortløbende og henvises til uden parenteser. Ligninger er nummereret på samme måde, og disse henvises til med parenteser. Beviser afsluttes med ■. Nogle resultater bevises ikke. Kilder angives med firkantede parenteser og er nummereret efter rækkefølgen i litteraturlisten.

Det antages, at læseren har viden svarende til en fuldført bacheloruddannelse på matematikstudiet ved Aalborg Universitet.

En tak til vores vejleder Anne Marie Svane.

*Aalborg Universitet 3. juni 2022.*

---

Christoffer Grebjerg

---

Søren Damgaard Nielsen

# Indholdsfortegnelse

---

<b>Forord</b>	<b>iii</b>
<b>Indholdsfortegnelse</b>	<b>iv</b>
<b>Kapitel 1 Indledning</b>	<b>1</b>
1.1 Problemafgrænsning . . . . .	1
1.2 Problemformulering . . . . .	1
<b>Kapitel 2 Teori</b>	<b>2</b>
2.1 En definition af kausal effekt . . . . .	2
2.1.1 Individuelle kausale effekter . . . . .	2
2.1.2 Gennemsnitlige kausale effekter . . . . .	3
2.1.3 Mål af kausal effekt . . . . .	7
2.1.4 Tilfældig variabilitet . . . . .	8
2.1.5 Kausalitet versus association . . . . .	10
2.2 Randomiserede eksperimenter . . . . .	13
2.2.1 Randomisering . . . . .	13
2.2.2 Betinget randomisereing . . . . .	16
2.2.3 Standardisering . . . . .	18
2.2.4 Invers sandsynlighedsvægtning . . . . .	20
2.3 Observationsstudier . . . . .	26
2.3.1 Identificerbarhedsbetingelser . . . . .	26
2.3.2 Ombyttelighed . . . . .	27
2.3.3 Positivitet . . . . .	28
2.3.4 Konsistens: Først, definér det kontrafaktiske udfald . . . . .	31
2.3.5 Konsistens: For det andet, forbind kontrafakta til det observerede data . . . . .	33
2.4 Grafisk afbildning af kausal inferens . . . . .	34
2.4.1 Strukturelle kausale modeller . . . . .	35
2.4.2 Indgreb . . . . .	37
2.4.3 Markov egenskab . . . . .	41
2.4.4 Kausale grafiske modeller . . . . .	45
2.4.5 Beregning af indgrebsfordelinger ved ændring af kovariat . . . . .	46
2.4.6 Potentielle udfald . . . . .	50
<b>Kapitel 3 Databehandling</b>	<b>54</b>
3.1 Estimering ved anvendelse af IP vægtning . . . . .	58
3.2 Standardisering . . . . .	70

3.3 Konklusion af databehandling . . . . .	76
<b>Kapitel 4 Konklusion</b>	<b>78</b>
<b>Litteratur</b>	<b>79</b>

# Indledning 1

---

I dette projekt tages der udgangspunkt i teorien om kausal inferens, hvorefter der ses på anvendelse af teorien på et datasæt, som omhandler Framingham Heart Study. Studiet startede i 1948 og sluttede i 1968, og i starten var der 5209 individer med i studiet. Studiet undersøgte, om individerne havde risikofaktorer, som kunne forårsage hjerte-kar-sygdomme. Et eksempel på risikofaktor er et individs historie om rygning.

## 1.1 Problemafgrænsning

I dette projekt fokuseres der på teorien om kausal inferens, så databehandlingen er til for at afprøve teorien i praksis, hvor der undersøges, om rygestop har nogen effekt på BMI. I teorien ses der kun på to metoder inden for estimation, og hvordan kausale modeller grafisk kan afbildes. Datasættet bliver anvendt til at lave et nyt datasæt ved at udvælge specifikke variabler fra Framingham Heart Study. Dette gøres for at afprøve teorien på et mere simpelt datasæt.

## 1.2 Problemformulering

- Hvordan kan kausalitet estimeres?
- Hvordan kan der tages højde for flere variabler, som muligvis har indflydelse på den interessante variabel?
- Har rygestop en kausal effekt på BMI? Hvis ja, hvor stor er effekten?

## 2.1 En definition af kausal effekt

Dette afsnit er baseret på kapitel 1 i [17], hvis ikke andet er angivet.

### 2.1.1 Individuelle kausale effekter

Betragt Zeus, som er en patient, der venter på en hjertetransplantation. Den 1. januar får han et nyt hjerte, og fem dage senere dør han. Antag, at det kan vides, at hvis Zeus ikke fik en hjertetransplantation den 1. januar, ville han være i live fem dage senere. Med denne information ville de fleste være enige om, at transplantationen var årsagen til Zeus' død. Hjertetransplantationen havde en kausal effekt på Zeus' fem dages overlevelse.

En anden patient, Hera, fik også en hjertetransplantation den 1. januar. Fem dage senere var hun i live. Antag, at det kan vides, at hvis ikke Hera fik et hjerte den 1. januar ville hun stadig have været i live fem dage senere. Dermed havde transplantationen ikke en kausal effekt på Heras fem dages overlevelse.

Dette illustrerer, hvordan menneskers fornuft om kausale effekter fungerer: Når en handling  $A$  finder sted, sammenlignes udfaldet med, hvis handlingen  $A$  er tilbageholdt. Hvis de to udfald er forskellige, har handlingen  $A$  en kausal effekt, forårsagende eller præventiv, på udfaldet. Ellers siges det, at handlingen  $A$  ingen kausal effekt har på udfaldet. Handlingen  $A$  kaldes også et indgreb, en eksponering eller en behandling.

Betragt en binær behandlingsvariabel  $A$  (1: behandlet, 0: ikke behandlet) og en binær udfaldsvariabel  $Y$  (1: død, 0: ikke død). Variabler som  $A$  og  $Y$ , der kan have forskellige værdier for forskellige individer, kaldes *stokastiske variabler*. Lad  $Y^{a=1}$  ( $Y$  under behandling  $a = 1$ ) være udfaldsvariablen, som ville have været observeret under behandling  $a = 1$ , og  $Y^{a=0}$  ( $Y$  under behandling  $a = 0$ ) være udfaldsvariablen, som ville have været observeret under behandling  $a = 0$ .  $Y^{a=1}$  og  $Y^{a=0}$  er også stokastiske variabler. Zeus har  $Y^{a=1} = 1$  og  $Y^{a=0} = 0$ , da han døde efter behandling, men han ville have overlevet uden behandling, mens Hera har  $Y^{a=1} = 0$  og  $Y^{a=0} = 0$ , da hun overlevede efter behandling og også ville have overlevet uden behandling.

**Definition 2.1.1** (Kausal effekt for et individ)

Behandlingen  $A$  har en kausal effekt på et individs udfald  $Y$ , hvis  $Y^{a=1} \neq Y^{a=0}$  for individet.

Dermed har behandlingen en kausal effekt på Zeus' udfald, da  $Y^{a=1} = 1 \neq 0 = Y^{a=0}$ , men ikke på Heras udfald, da  $Y^{a=1} = 0 = Y^{a=0}$ . Variableerne  $Y^{a=1}$  og  $Y^{a=0}$  kaldes *potentielle udfald* eller *kontrafaktiske udfald*.

Et af de kontrafaktiske udfald (det, der svarer til behandlingen, som individet fik) er faktisk faktisk for ethvert individ. For eksempel, da Zeus faktisk var behandlet ( $A = 1$ ), er hans kontrafaktiske udfald under behandling  $Y^{a=1} = 1$  lig med hans observerede udfald ( $Y = 1$ ). Et individ med observeret behandling  $A$  lig med  $a$  har observeret udfald  $Y$  lig med sit kontrafaktiske udfald  $Y^a$ . Denne lighed kan udtrykkes som  $Y = Y^A$ , hvor  $A$  betegner det kontrafaktiske  $Y^a$  evalueret ved værdien  $a$  svarende til individets observerede behandling  $A$ . Ligheden  $Y = Y^A$  kaldes *konsistens*.

Individuelle kausale effekter er defineret som en kontrast af værdierne af kontrafaktiske udfald, men kun et af disse udfald er observeret for hvert individ - det ene svarende til behandlingen, som individet faktisk oplevede. Alle andre kontrafaktiske udfald forbliver uobserverede. På grund af manglende data kan individuelle kausale effekter ikke identificeres, det vil sige, at de ikke kan udtrykkes som en funktion af det observerede data.

I næste underafsnit vises der, hvordan der kan estimeres, om der er en kausal effekt ved anvendelse af det, der kaldes gennemsnitlig kausal effekt.

**2.1.2 Gennemsnitlige kausale effekter**

Der er brug for tre informationer til at definere en individuel kausal effekt: et udfald, handlingerne  $a = 1$  og  $a = 0$  skal sammenlignes, og individets kontrafaktiske udfald  $Y^{a=0}$  og  $Y^{a=1}$  skal sammenlignes. Men da det generelt ikke er muligt at identificere individuelle kausale effekter, introduceres nu den gennemsnitlige kausale effekt i en population af individer. For at definere det er der brug for tre informationer: et udfald, handlingerne  $a = 1$  og  $a = 0$  skal sammenlignes, og en veldefineret population af individer hvis udfald  $Y^{a=0}$  og  $Y^{a=1}$  skal sammenlignes.

	$Y^{a=0}$	$Y^{a=1}$
Rheia	0	1
Kronos	1	0



Demeter	0	0
Hades	0	0
Hestia	0	0
Poseidon	1	0
Hera	0	0
Zeus	0	1
Artemis	1	1
Apollo	1	0
Leto	0	1
Ares	1	1
Athena	1	1
Hephaustus	0	1
Aphrodite	0	1
Cyclope	0	1
Persephone	1	1
Hermes	1	0
Hebe	1	0
Dionysus	1	0

Tabel 2.1

### Eksempel 2.1.2

Betragt Zeus' familie som populationen. Tabel 2.1 viser de kontrafaktiske udfald under både behandling ( $a = 1$ ) og ingen behandling ( $a = 0$ ) for alle 20 individer i populationen. Betragt den sidste søjle: udfaldet  $Y^{a=1}$ , som ville have været observeret for hvert individ, hvis de havde fået behandlingen (en hjertetransplantation). Halvdelen af individerne af populationen (10 ud af 20) ville være døde, hvis de havde fået en hjertetransplantation. Det vil sige, at sandsynligheden kan opskrives som  $P(Y^{a=1} = 1) = 10/20 = 0.5$ . Tilsvarende fra den anden søjle i Tabel 2.1 kan det konkluderes, at halvdelen af individerne i populationen (10 ud af 20) ville være døde, hvis ikke de fik en hjertetransplantation. Det vil sige, at sandsynligheden kan opskrives som  $P(Y^{a=0} = 1) = 10/20 = 0.5$ . Den kontrafaktiske risiko under behandling blev beregnet til at være 0.5 ved at tælle antallet af dødsfald (10) og dividere det med det totale antal af individer (20), hvilket er det samme som at beregne gennemsnittet af de kontrafaktiske udfald over alle individer i populationen. For at se sammenhængen mellem risiko og gennemsnit for et binær udfald kan data i Tabel 2.1 anvendes til at bestemme gennemsnittet af  $Y^{a=1}$ .  $\square$

Nu ses der på definitionen for en gennemsnitlig kausal effekt i en population.

**Definition 2.1.3** (Gennemsnitlig kausal effekt)

En gennemsnitlig kausal effekt af behandling  $A$  på udfald  $Y$  er til stede, hvis

$$P(Y^{a=1} = 1) \neq P(Y^{a=0} = 1)$$

i populationen. For ikke-binære udfald anvendes

$$E[Y^{a=1}] \neq E[Y^{a=0}].$$

Fra eksempel 2.1.2 vides det, at under denne definition har behandling  $A$  ikke en gennemsnitlig kausal effekt på udfald  $Y$  i populationen, da både risikoen for dødsfald under behandling  $P(Y^{a=1} = 1)$  og risikoen for dødsfald under ingen behandling  $P(Y^{a=0} = 1)$  er 0.5. Det har ingen betydning, om alle eller ingen af individerne får en hjertetransplantation: Halvdelen af dem ville dø i begge tilfælde. Hvis den gennemsnitlige kausale effekt i en population er nul, ligesom her, siges det, at *nulhypotesen om ingen gennemsnitlig kausal effekt* er sand.

**Eksempel 2.1.4**

**Flere versioner af behandling.** Definitionen af et kontrafaktisk udfald under behandlingsværdi  $a$  antager også implicit, at der kun er en version af behandlingsværdi  $A = a$ . Det blev for eksempel tidligere sagt, at Zeus ville dø, hvis han fik en hjertetransplantation. Denne udmelding antager implicit, at alle hjertetransplantationer er udført af den samme kirurg med samme procedure og udstyr. Det vil sige, at der kun er en version af behandlingen ”hjertetransplantation”. Hvis der var flere versioner af behandling (eksempelvis kirurger med forskellige evner), er det muligt, at Zeus ville overleve, hvis hans transplantation var udført af Asclepios, og ville dø, hvis hans transplantation var udført af Hygieia. I forhold til flere versioner af behandling er det kontrafaktiske  $Y_i^a$  for et individ  $i$  ikke veldefineret, da et individs udfald afhænger af versionen af behandling  $a$ . Hvis der er flere versioner af behandling, er ”den kausale effekt af hjertetransplantation på Zeus’ udfald” ikke veldefineret. Derfor er det nødvendigt at referere til ”den kausale effekt af hjertetransplantation på Zeus’ udfald hvis Asclepios udfører operationen” eller ”den kausale effekt af hjertetransplantation på Zeus’ udfald hvis Hygieia udfører operationen”. Hvis andre komponenter af behandling (eksempelvis procedure, sted) også er relevant til udfaldet, kan det være nødvendigt at referere til ”den kausale effekt af hjertetransplantation på Zeus’ udfald hvis Asclepios udfører operationen ved anvendelse af hans stang ved templet på Kos”, da den kausale effekt af behandling på Zeus’ udfald kan variere for hver version af behandling. □

Fravær af en gennemsnitlig kausal effekt medfører ikke fravær af individuelle effekter. Tabel 2.1 viser, at behandling har en individuel kausal effekt på 12 individer (herunder Zeus) af populationen da, for hver af disse 12 individer er værdierne af deres kontrafaktiske udfald  $Y^{a=1}$  og

$Y^{a=0}$  forskellige. Ud af de 12 var 6 skadet af behandling, herunder Zeus ( $Y^{a=1} - Y^{a=0} = 1$ ), og 6 blev hjulpet ( $Y^{a=1} - Y^{a=0} = -1$ ). Denne lighed er ikke en fejl: Den gennemsnitlige kausale effekt  $E[Y^{a=1}] - E[Y^{a=0}]$  er altid lig med gennemsnittet  $E[Y^{a=1} - Y^{a=0}]$  af de individuelle kausale effekter  $Y^{a=1} - Y^{a=0}$ . Hvis der ikke er nogen kausal effekt for et individ i populationen, eksempelvis  $Y^{a=1} = Y^{a=0}$  for alle individer, siges det, at den *skarpe kausale nulhypotese* er sand. Den skarpe kausale nulhypotese medfører nulhypotesen om ingen gennemsnitlig effekt.

Det viser sig, at gennemsnitlige kausale effekter nogle gange kan identificeres fra data, selv hvis individuelle kausale effekter ikke kan. Herefter refereres ”gennemsnitlige kausale effekter” som ”kausale effekter” og nulhypotesen om ingen gennemsnitlig effekt som den kausale nulhypotese.

### Eksempel 2.1.5

**Kausale effekter i populationen.** Lad  $E[Y^a]$  være det middel kontrafaktiske udfald hvis alle individer i populationen fik behandlingsværdi  $a$ . For diskrete udfald er middelværdien  $E[Y^a]$  defineret som den vægtede sum  $\sum_y y p_{Y^a}(y)$  over alle mulige værdier  $y$  af den stokastiske variabel  $Y^a$ , hvor  $p_{Y^a}(\cdot)$  er pmf af  $Y^a$ , det vil sige  $p_{Y^a}(y) = P(Y^a = y)$ . For binær udfald er  $E[Y^a] = P(Y^a = 1)$ . For kontinuerte udfald er  $E[Y^a]$  defineret som et integral  $\int y f_{Y^a}(y) dy$  over alle mulige værdier  $y$  af den stokastiske variabel  $Y^a$ , hvor  $f_{Y^a}(\cdot)$  er pdf af  $Y^a$ . En almindelig repræsentation af middelværdien, som anvendes på både diskrete og kontinuerte udfald er  $E[Y^a] = \int y dF_{Y^a}(y)$ , hvor  $F_{Y^a}(\cdot)$  er cdf for den stokastiske variabel  $Y^a$ . Hvis  $E[Y^a] \neq E[Y^{a'}]$  for to værdier  $a$  og  $a'$  siges det, at der er en ikke-nul gennemsnitlig kausal effekt i populationen.

Den gennemsnitlige kausale effekt, som er defineret af en kontrast af middelværdier af kontrafaktiske udfald, er det mest almindeligt anvendte mål for den kausale effekt i en population. En populations kausal effekt kan dog også defineres som en kontrast af funktionaler, herunder medianer, varianser, hazards eller cdf'er af kontrafaktiske udfald. Generelt kan en populations kausale effekt defineres som en kontrast af enhver funktional af de marginale fordelinger af kontrafaktiske udfald under forskellige handlinger eller behandlingsværdier. For eksempel er populations kausal effekten af variansen defineret som  $var(Y^{a=1}) - var(Y^{a=0})$ , hvilket er nul for populationen i Tabel 2.1, da fordelingen af  $Y^{a=1}$  og  $Y^{a=0}$  er identiske (begge har 6 dødsfald ud af 20). Ligheden af disse fordelinger medfører, at for enhver funktional, er populationens kausale effekt på det funktionale nul. Forskellen  $var(Y^{a=1}) - var(Y^{a=0})$  er generelt ikke lig med variansen af individuelle kausale effekter  $var(Y^{a=1} - Y^{a=0})$ . For eksempel, i Tabel 2.1, da  $Y^{a=1} - Y^{a=0}$  ikke er en konstant ( $-1$  for 6 individer,  $1$  for 6 individer og  $0$  for 8 individer), så er  $var(Y^{a=1} - Y^{a=0}) > 0 = var(Y^{a=1}) - var(Y^{a=0})$ . Det er muligt at bestemme  $var(Y^{a=1}) - var(Y^{a=0})$  fra data i et tilfældigt eksperiment, men ikke  $var(Y^{a=1} - Y^{a=0})$ , da det ikke er muligt at observere både  $Y^{a=1}$  og  $Y^{a=0}$  simultant for et individ, og dermed er kova-

riansen af  $Y^{a=1}$  og  $Y^{a=0}$  ikke identificeret. Dette er sandt, ikke kun for variansen, men også for enhver ikke-lineær funktional (eksempelvis median og hazard).  $\square$

Hernæst beskrives forskellige mål for størrelsen af en kausal effekt.

### 2.1.3 Mål af kausal effekt

Det er vist, at behandlingen ”hjertetransplantation”  $A$  ikke har nogen kausal effekt på udfaldet ”Død”  $Y$  i populationen af 20 familiemedlemmer af Zeus. Den kausale nulhypotese holder, da de to kontrafaktiske risikoer  $P(Y^{a=1} = 1)$  og  $P(Y^{a=0} = 1)$  er lig hinanden med værdi 0.5. Der er forskellige måder at repræsentere den kausale nulhypotese. Et eksempel kan være, at risikoen  $P(Y^{a=1} = 1) - P(Y^{a=0} = 1) = 0$ , eller at risikoen  $P(Y^{a=1} = 1)/P(Y^{a=0} = 1) = 1$ .

#### **Definition 2.1.6** (Kausale effektmål)

Repræsentationerne af den kausale nulhypotese ser ud på følgende måder

- $P(Y^{a=1} = 1) - P(Y^{a=0} = 1) = 0$ ,
- $\frac{P(Y^{a=1}=1)}{P(Y^{a=0}=1)} = 1$ ,
- $\frac{P(Y^{a=1}=1)/P(Y^{a=1}=0)}{P(Y^{a=0}=1)/P(Y^{a=0}=0)} = 1$ ,

hvor venstre side af lighederne kaldes henholdsvis den kausale- risiko forskel, risiko ratio og odds ratio.

Antag nu, at en anden behandling  $A$ , rygning, har en kausal effekt på et andet udfald  $Y$ , lungekræft, i populationen. Den kausale nulhypotese holder ikke, altså  $P(Y^{a=1} = 1) \neq P(Y^{a=0} = 1)$ . I denne situation er den kausale risiko forskel, risiko ratio og odds ratio ikke lig med henholdsvis 0, 1 og 1. I stedet for kvantificerer disse kausale parametre styrken af den samme kausale effekt på forskellige skalaer. Da den kausale- risiko forskel, risiko ratio og odds ratio måler den kausale effekt, kaldes de for *kausale effektmål*, altså mål af den kausale effekt.

Hvert effektmål kan blive anvendt til forskellige formål. For eksempel i en stor population, hvor 3 ud af en million individer udvikler udfaldet, hvis behandlingen gives, og 1 ud af en million individer udvikler udfaldet, hvis de ikke bliver behandlet. Den kausale risiko ratio er 3, og den kausale risiko forskel er 0.000002. Den kausale risiko ratio (multiplikativ skala) er anvendt til at beregne, hvor mange gange behandling, relativ til ingen behandling, øger sygdomsrisikoen. Den kausale risiko forskel (additiv skala) er anvendt til at beregne absolutte antal af tilfælde af sygdommen, der tilskrives behandlingen. Anvendelse af disse mål afhænger af hvad målet er for inferensen.

### Eksempel 2.1.7

**Nødvendigt antal at behandle.** Betragt en population på 100 millioner patienter, hvor 20 millioner dør indenfor fem år, hvis behandlet ( $a = 1$ ), og 30 millioner dør indenfor fem år, hvis ikke behandlet ( $a = 0$ ). Denne information kan summeres på flere ækvivalente måder:

- Den kausale risiko forskel er  $P(Y^{a=1} = 1) - P(Y^{a=0} = 1) = 0.2 - 0.3 = -0.1$
- Hvis de 100 millioner af patienter behandles, så er der 10 millioner færre, der ikke dør, end hvis ikke blev behandlet.
- Det er nødvendigt at behandle 100 millioner patienter, så 10 millioner liv reddes.
- I gennemsnit skal 10 patienter behandles for at redde 1 liv.

Det gennemsnitlige antal af individer, der skal have behandling  $a = 1$  for at reducere antallet af tilfælde  $Y = 1$  med 1, bliver kaldt det nødvendige antal at behandle (NNT, number needed to treat). I eksemplet er NNT lig med 10. For behandlinger, der reducerer det gennemsnitlige antal tilfælde, altså den kausale risiko forskel er negativ, er NNT lig med den reciprokke af den absolutte værdi af den kausale risiko forskel:

$$NNT = \frac{-1}{P(Y^{a=1} = 1) - P(Y^{a=0} = 1)}.$$

For behandlinger, der øger gennemsnittet af tilfælde, kan der symmetrisk defineres et nødvendigt antal for at gøre skade. □

### 2.1.4 Tilfældig variabilitet

Det kan argumenteres, at indtil videre er proceduren til at beregne effektmålene ikke plausibel, da populationen kun består af 20 individer, mens det normalt handler om populationer, der er meget større.

I den lille population er der indsamlet information fra alle individerne. I praksis indsamler forskere kun information fra en stikprøve af den interessante population. Selv hvis de kontrafaktiske udfald af alle individer i studiet var kendte, forhindrer en stikprøve en i at beregne den eksakte proportion af individer i populationen, der havde et udfald under behandlingsværdi  $a$ . Dette vil sige, at sandsynligheden under ingen behandling  $P(Y^{a=0} = 1)$  ikke kan beregnes direkte, men den kan kun estimeres.

Betragt individerne i Tabel 2.1. Tidligere var de hele populationen, men nu antages de at være en tilfældig stikprøve af en meget større, nær uendelig, super-population. Den proportion af individer i stikprøven, som dør, hvis ikke behandlet, betegnes  $\hat{P}(Y^{a=0} = 1) = 10/20 = 0.5$ . Proportionen af stikprøven  $\hat{P}(Y^{a=0} = 1)$  behøver ikke at være lig med proportionen af individer, der ville dø, hvis hele super-populationen ikke blev behandlet,  $P(Y^{a=0} = 1)$ . For eksempel,

hvis  $P(Y^{a=0} = 1) = 0.57$  i populationen, men på grund af tilfældige fejl der skete, da der er variabilitet i stikprøver, er  $\hat{P}(Y^{a=0} = 1) = 0.5$  i stikprøven. Proportionen  $\hat{P}(Y^a = 1)$  fra stikprøven anvendes til at estimere den tilhørende proportion i hele populationen  $P(Y^a = 1)$  under behandlingsværdi  $a$ , hvilket vil sige, at estimatoren skal være konsistent.

**Definition 2.1.8** (Konsistent estimator [5])

Antag, at parametermængden  $\Theta$  er en åben delmængde af  $\mathbb{R}^k$ . En sekvens af estimater  $t_n : \mathbb{R}^n \rightarrow \Theta$  af parameteren  $\theta$  er *konsistent*, hvis

$$\hat{\theta}_n \xrightarrow{P} \theta \quad \text{for } n \rightarrow \infty,$$

hvis  $\theta$  er den sande parameter.

**Sætning 2.1.9** (Store tals lov [14])

Lad  $X_1, X_2, \dots, X_n$  være i.i.d. stokastiske variable med en endelig forventet værdi  $E[X_i] = \mu < \infty$ . Lad også

$$M_n = \frac{X_1 + X_2 + \dots + X_n}{n},$$

så gælder  $M_n \xrightarrow{n.s.} \mu$ .

$\hat{P}(Y^a = 1)$  er en konsistent estimator af  $P(Y^a = 1)$ , da den opfylder Store tals lov 2.1.9, når antallet af individer i stikprøven vokser. Bemærk, at da sandsynligheden  $P(Y^a = 1)$  ikke kan beregnes, men kun estimeres med  $\hat{P}(Y^a = 1)$ , kan der ikke konkluderes med sikkerhed, om der er eller ikke er en kausal effekt.

Indtil videre har der kun været talt om variabilitet i stikprøve, men der kan være en anden kilde af tilfældig variabilitet. Det kunne være at et individs kontrafaktiske udfald ikke er fastsat. Før er der blevet kigget på Zeus, hvor hans kontrafaktiske udfald har været fastsat eller deterministisk, altså  $Y^{a=1} = 1$  og  $Y^{a=0} = 0$ . Med andre ord har Zeus haft 100% chance for at dø hvis behandlet, og 0% chance for at dø hvis ikke behandlet, men der kunne forestilles at være et andet scenarie, hvor Zeus har 90% chance for at dø hvis behandlet, og 10% chance for at dø hvis ikke behandlet. I dette scenarie er de kontrafaktiske udfald stokastiske eller ikke-deterministiske, da Zeus' sandsynlighed for at dø med behandling og uden behandling ikke er enten 0 eller 1. Værdierne i Tabel 2.1 kan være mulige udfald af tilfældige flip af en dødeligheds mønt med disse sandsynligheder. Det forventes, at ikke alle individer har de samme sandsynligheder, da ikke alle har samme modtagelighed for at udvikle udfaldet.

Dette giver, at i kausal inferens kommer de tilfældige fejl fra variabilitet i stikprøve, ikke-

deterministisk kontrafakta eller begge. Indtil videre ses der bort fra disse tilfældige fejl.

### Eksempel 2.1.10

**Ikke-deterministisk kontrafakta.** For ikke-deterministiske kontrafaktiske udfald, er det forventede udfald under behandlingsværdi  $a$ ,  $E[Y^a] = \sum_y y p_{Y^a}(y)$  over alle mulige værdier  $y$  af den tilfældige variabel  $Y^a$ . Sandsynlighedsmasse funktionen er  $p_{Y^a}(\cdot) = E[Q_{Y^a}(\cdot)]$  og  $Q_{Y^a}(y)$  er en tilfældig sandsynlighed for at have udfald  $Y = y$  under behandlingsværdi  $a$ . I eksemplet givet før var  $Q_{Y^{a=1}}(1) = 0.9$  for Zeus. (For kontinuerte udfald bliver den vægtede sum erstattet med et integral.)

Mere generelt giver en ikke-deterministisk definition af kontrafaktiske udfald ikke en bestemt værdi til den tilfældige variabel  $Y^a$  til hvert individ, men snarere en individ-specifik statistisk fordeling  $\Theta_{Y^a}(\cdot)$  af  $Y^a$ . Den ikke-deterministiske definition af kausal effekt er en generalisering af den deterministiske definition, hvor  $\Theta_{Y^a}(\cdot)$  nu er en tilfældig cdf, der kan tage værdier mellem 0 og 1. Det gennemsnitlige kontrafaktiske udfald i populationen  $E[Y^a]$  er lig med  $E[E[Y^a|\Theta_{Y^a}(\cdot)]]$ , derfor er

$$E[Y^a] = E\left[\int y d\Theta_{Y^a}(y)\right] = \int y dE[\Theta_{Y^a}(y)] = \int y dF_{Y^a}(y)$$

hvor  $F_{Y^a}(\cdot) = E[\Theta_{Y^a}(\cdot)]$ .

Hvis det kontrafaktiske udfald er binært og ikke-deterministisk, er den kausale risiko ratio i populationen lig med den vægtede forventede værdi,  $\frac{E[\Theta_{Y^{a=1}}(1)]}{E[\Theta_{Y^{a=0}}(1)]} = E\left[W(\Theta_{Y^{a=1}}(1)/\Theta_{Y^{a=0}}(1))\right]$ , af individets kausale effekt  $\Theta_{Y^{a=1}}(1)/\Theta_{Y^{a=0}}(1)$  på ratio skalaen, med vægt  $W = \frac{\Theta_{Y^{a=0}}(1)}{E[\Theta_{Y^{a=0}}(1)]}$ , givet at  $\Theta_{Y^{a=0}}(1)$  aldrig er lig med 0 for nogle i populationen.  $\square$

I næste underafsnit ses der på, hvornår det kan siges, om der er kausalitet eller association.

## 2.1.5 Kausalitet versus association

Selvfølgelig vil data fra virkelige studier se anderledes ud end i Tabel 2.1. For eksempel ville man ikke kunne forvente, at vide Zeus' udfald hvis behandlet  $Y^{a=1}$  og Zeus' udfald hvis ikke behandlet  $Y^{a=0}$ . I den virkelige verden kan kun et af disse udfald kunne observeres. Det observerede udfald kaldes  $Y$ , så for hvert individ kendes den observerede behandling  $A$  og udfaldet  $Y$  som i følgende tabel.

	A	Y
Rheia	0	0
Kronos	0	1

Demeter	0	0
Hades	0	0
Hestia	1	0
Poseidon	1	0
Hera	1	0
Zeus	1	1
Artemis	0	1
Apollo	0	1
Leto	0	0
Ares	1	1
Athena	1	1
Hephaustus	1	1
Aphrodite	1	1
Cyclope	1	1
Persephone	1	1
Hermes	1	0
Hebe	1	0
Dionysus	1	0

Tabel 2.2

Data i Tabel 2.2 kan anvendes til at beregne proportionen af individer, der udvikler udfaldet  $Y$  blandt de individer i populationen, som har fået behandlingsværdi  $a$ . For eksempel er der 7 individer, der dør ( $Y = 1$ ) blandt de 13 individer, der blev behandlet ( $A = 1$ ). Dette giver en risiko for at individet dør på  $P(Y = 1|A = 1) = 7/13$ . Mere generelt er den betingede sandsynlighed  $P(Y = 1|A = a)$  defineret som den proportion af individer, der udvikler udfaldet  $Y$  blandt de individer i populationen, der har fået behandlingsværdi  $a$ .

*Uafhængighed* er repræsenteret med  $Y \perp\!\!\!\perp A$ , eller ækvivalent  $A \perp\!\!\!\perp Y$ , hvilket læses som at  $Y$  og  $A$  er uafhængige. Nogle ækvivalente definitioner af uafhængighed er

- $P(Y = 1|A = 1) - P(Y = 1|A = 0) = 0$
- $\frac{P(Y=1|A=1)}{P(Y=1|A=0)} = 1$
- $\frac{P(Y=1|A=1)/P(Y=0|A=1)}{P(Y=1|A=0)/P(Y=0|A=0)} = 1$ ,

hvor venstre side af lighederne kaldes henholdsvis associations- risiko forskel, risiko ratio og odds ratio.

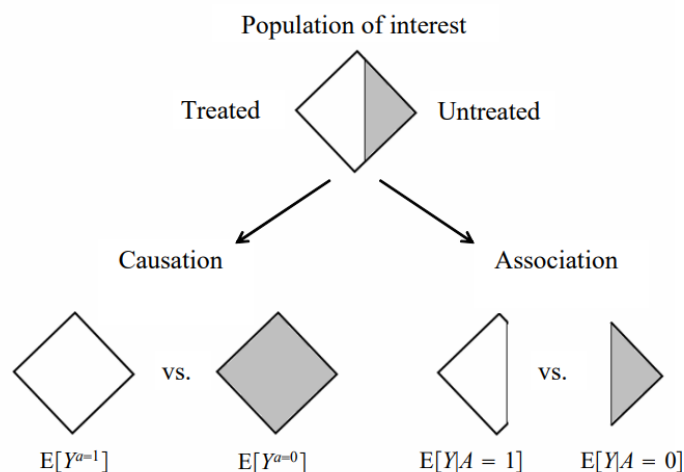


**Definition 2.1.11** (Association)

Behandling  $A$  og udfald  $Y$  kaldes afhængige eller *associeret*, når  $P(Y = 1|A = 1) \neq P(Y = 1|A = 0)$ .

I populationen er behandling og udfald associeret, da  $P(Y = 1|A = 1) = 7/13$  og  $P(Y = 1|A = 0) = 3/7$ . Associations- risiko forskel, risiko ratio og odds ratio kvantificerer styrken af associationen, når den eksisterer, og de måler associationen på forskellige skalaer og kaldes *associationsmål*. Disse mål er også påvirket af tilfældig variabilitet, men det ses der bort fra, og der antages, at populationen i Tabel 2.2 er meget stor.

For binære udfald er risikoen lig med gennemsnittet i populationen, så definitionen af association i populationen kan omskrives til  $E[Y|A = 1] \neq E[Y|A = 0]$ , hvilket også er definitionen for kontinuert  $Y$ . I populationen af 20 individer blev der fundet, at (i) ingen kausal effekt efter at sammenligne risikoen for at dø for alle 20 individer når de blev behandlet, og risikoen for at dø når de ikke blev behandlet, og (ii) en association blev fundet efter at sammenligne risikoen for at dø for de 13 individer, der blev behandlet, med risikoen for at dø for de 7 individer, der ikke blev behandlet. Figur 2.1 viser kausalitet-association forskellen. Populationen repræsenteret med øverste rombe er opdelt i et hvidt område for dem, der blev behandlet og et mindre gråt område for dem, der ikke blev behandlet.



Figur 2.1: Taget fra [17] på side 12.

Definitionen af kausalitet forudsætter en kontrast mellem den hele hvide rombe (alle behandlet) og den hele grå rombe (alle ubehandlet), hvor association forudsætter en kontrast mellem det hvide område og det grå område i den originale rombe. Dette betyder, at inferens med kausalitet arbejder med ”hvad hvis”-spørgsmål i kontrafaktiske verdener, så som ”hvad vil risikoen være, hvis alle blev behandlet?” og ”hvad vil risikoen være, hvis alle ikke blev behandlet?”, hvor inferens med association arbejder med spørgsmål i den virkelige verden, så som ”hvad er risikoen

for dem, der blev behandlet?” og ”hvad er risikoen for dem, der ikke blev behandlet?”

Notationen, der er blevet anvendt tidligere, kan formalisere forskellen mellem kausalitet og association. Risikoen  $P(Y = 1|A = a)$  er en betinget sandsynlighed, der omhandler risikoen for  $Y$  for den delmængde af populationen, der opfylder at have behandlingsværdi  $A = a$ . I kontrast er risikoen  $P(Y^a = 1)$  en ubetinget, også kaldet marginal, sandsynlighed af  $Y^a$  for hele populationen. Derfor er association defineret ved en forskellig risiko i to disjunkte delmængder, der er opdelt baseret på deres behandlingsværdi  $A = 1$  eller  $A = 0$ , hvor kausalitet er defineret ved en forskellig risiko for den samme population under to behandlingsværdier  $A = 1$  og  $A = 0$ .

Disse forskellige definitioner forklarer det velkendte udsagn ”association er ikke kausalitet”. I populationen var der association, da risikoen for at dø for de behandlede ( $7/13$ ) var større end dem, der ikke blev behandlet ( $3/7$ ). Derimod var der ikke nogen kausalitet på grund af, at når alle blev behandlet, var risikoen ( $10/20$ ) den samme, som hvis alle ikke blev behandlet. Denne uoverensstemmelse mellem kausalitet og association ville ikke være overraskende hvis dem, der blev behandlet (fik et ny hjerte) gennemsnitligt er mere syg end dem, der ikke bliver behandlet.

Kausal inferens kræver data som i Tabel 2.1, men alt, der kan forventes i den virkelige verden, er data som i Tabel 2.2. Nu er spørgsmålet under hvilke betingelser kan virkelig data anvendes til kausal inferens. I næste afsnit gives der et svar, hvilket er at lave et randomiseret eksperiment.

## 2.2 Randomiserede eksperimenter

Dette afsnit er baseret på kapitel 2 i [17], hvis ikke andet er angivet.

Hvis et individ kigger op i luften, vil det så gøre, at andre individer også kigger op i luften? Dette spørgsmål har hoveddelene af ethvert kausalt spørgsmål: der vil vides, om en handling (kigge op) har indflydelse på et udfald (andre kigger op) i en specifik population. I dette afsnit ses der på, hvordan et eksperiment kan designes til at besvare disse spørgsmål.

### 2.2.1 Randomisering

I et reelt studie kunne begge Zeus’ potentielle udfald,  $Y^{a=1}$  med behandling, og  $Y^{a=0}$  uden behandling, ikke vides. Det, der vides, er kun det observerede udfald  $Y$  under behandling  $A$ . Tabel 2.3 giver et overblik over, hvordan dette kan se ud, hvor der er manglende data. Data fra denne tabel kan kun anvendes til at beregne associationsmål.

	$A$	$Y$	$Y^0$	$Y^1$
Rheia	0	0	0	?
Kronos	0	1	1	?
Demeter	0	0	0	?
Hades	0	0	0	?
Hestia	1	0	?	0
Poseidon	1	0	?	0
Hera	1	0	?	0
Zeus	1	1	?	1
Artemis	0	1	1	?
Apollo	0	1	1	?
Leto	0	0	0	?
Ares	1	1	?	1
Athena	1	1	?	1
Hephaustus	1	1	?	1
Aphrodite	1	1	?	1
Cyclope	1	1	?	1
Persephone	1	1	?	1
Hermes	1	0	?	0
Hebe	1	0	?	0
Dionysus	1	0	?	0

Tabel 2.3

**Eksempel 2.2.1** (Eksempel på randomiseret eksperiment)

Antag, at populationen i Figur 2.1 var nær uendelig, og at der kastes en mønt for hvert individ i populationen. Individene bliver givet den hvide farve, hvis mønten bliver krone, og grå hvis det er plat. Bemærk, at det ikke er en fair mønt, da sandsynligheden for plat ikke er 0.5. Derefter bedes en forskningsassistent om at give den behandling, der er interessant ( $A = 1$ ) til den hvide gruppe, og placebo ( $A = 0$ ) til den grå gruppe. Fem dage senere ved slutningen af studiet beregnes dødelighedsrisikoen i hver gruppe,  $P(Y = 1|A = 1) = 0.3$  og  $P(Y = 1|A = 0) = 0.6$ . Associations risiko ratioen var  $0.3/0.6 = 0.5$  og associations risiko forskellen var  $0.3 - 0.6 = -0.3$ . I dette tilfælde antages det, at det er et ideelt randomiseret eksperiment, hvor der antages, at der ikke er nogen ukendte variabler, der har indflydelse på eksperimentet. Selvom det ikke er realistisk, er det godt til at introducere nogle vigtige koncepter inden for kausal inferens. □

*Randomiserede eksperimenter* giver data med manglende værdier, ligesom alle andre reelle eksperimenter, men randomisering sikrer, at disse manglende værdier kun fandt sted tilfældigt. Dette giver, at effekt mål kan beregnes, eller mere præcist konsistent estimeres, i randomiserede eksperimenter, selv om der er manglende data.

Hvis nu at der var sket en fejl, og at den gruppe, der fik behandlingen, var den grå gruppe, hvad så? Det vil ikke gøre nogen forskel. Da populationen var randomiseret, ændrer sandsynlighederne sig ikke, da proportionerne af dødsfald blandt dem, der fik behandlingen og ikke fik behandlingen, er de samme. Dette vil sige, at når grupperne er randomiseret, er det irrelevant hvilken gruppe, der fik behandlingen. Formelt kaldes grupperne ombyttelige.

**Definition 2.2.2** (Ombyttelighed)

Der siges at være *ombyttelighed*, når det kontrafaktiske udfald og den aktuelle behandling er uafhængige af hinanden, altså  $Y^a \perp\!\!\!\perp A$  for alle værdier  $a$ .

Bemærk, at der er en forskel mellem  $Y^a \perp\!\!\!\perp A$  og  $Y \perp\!\!\!\perp A$ , da der er forskel på det kontrafaktiske  $Y^a$  og det observerede  $Y$ . Hvis der i et randomiseret eksperiment er ombyttelighed, altså  $Y^a \perp\!\!\!\perp A$ , er der ikke nødvendigvis uafhængighed mellem det observerede udfald og den observerede behandling  $Y \perp\!\!\!\perp A$ , men det kan forekomme.

Randomisering er vigtig, da det forventes at producere ombyttelighed, men det er ikke nødvendigvis sikkert. Denne uafhængighed betyder, at den kontrafaktiske risiko under behandling i den hvide population er lig med den kontrafaktiske risiko under behandling for hele populationen. Bemærk, at den kontrafaktiske risiko for den hvide population ikke er kontrafaktisk, da den hvide population faktisk blev behandlet. Derfor gør det ideelle randomiserede eksperiment det muligt at beregne den kontrafaktiske risiko under behandling for hele populationen  $P(Y^{a=1} = 1)$ , da den er lig med risikoen for de behandlede  $P(Y = 1|A = 1) = 0.3$ . Det samme rationale kan anvendes til at beregne den kontrafaktiske risiko for dem, der ikke bliver behandlet (0.6), den kausale risiko ratio (0.5) og den kausale risiko forskel (-0.3).

**Sætning 2.2.3**

Hvis der er ombyttelighed i et ideelt randomiseret eksperiment, gælder

$$P(Y^{a=1} = 1) = P(Y = 1|A = 1).$$

**Eksempel 2.2.4**

Er der ombyttelighed i studiet om hjertetransplantation i Tabel 2.3? For at svare på dette, skal der ses, om  $Y^a \perp\!\!\!\perp A$  holder for  $a = 0$  og  $a = 1$ . Der ses på  $a = 0$ . Hvis det kontrafaktiske

data fra Tabel 2.1 er tilgængeligt, kan risikoen for at dø uden behandling beregnes  $P(Y^{a=0}|A = 1) = 7/13$  for de 13 individer, der blev behandlet. Risikoen for at dø uden behandling for de 7 individer, der ikke blev behandlet, er  $P(Y^{a=0}|A = 0) = 3/7$ , da risikoen for at dø uden behandling er større blandt de behandlede individer,  $7/13 > 3/7$ , og det kan konkluderes, at der ikke er ombyttelighed i studiet. Dette kan vises analogt for  $a = 1$ .  $\square$

Det er kun det observerede data fra Tabel 2.3, der er tilgængeligt i den virkelige verden, men data fra denne tabel er ikke tilstrækkeligt til at udregne de kontrafaktiske risikoer. Konklusionen om at der ikke er ombyttelighed i data er ikke tilstrækkeligt til at kunne sige, at der ikke er tale om et randomiseret eksperiment. Der kan være tale om et randomiseret eksperiment, hvor forskere ikke kun anvendte en sandsynlighed (en mønt) til at bestemme behandlingen, men flere sandsynligheder (flere mønter). Dette beskrives i næste underafsnit.

## 2.2.2 Betinget randomisereing

Tabel 2.4 viser data fra det randomiserede hjertetransplantationsstudie. Udover behandling  $A$  og udfaldet  $Y$  er der prognosefaktoren  $L$  (1 hvis individet er i kritisk tilstand, 0 ellers), som blev målt, inden behandling blev tildelt.

	$L$	$A$	$Y$
Rheia	0	0	0
Kronos	0	0	1
Demeter	0	0	0
Hades	0	0	0
Hestia	0	1	0
Poseidon	0	1	0
Hera	0	1	0
Zeus	0	1	1
Artemis	1	0	1
Apollo	1	0	1
Leto	1	0	0
Ares	1	1	1
Athena	1	1	1
Hephaustus	1	1	1
Aphrodite	1	1	1
Cyclope	1	1	1
Persephone	1	1	1
Hermes	1	1	0

Hebe	1	1	0
Dionysus	1	1	0

Tabel 2.4

Der betragtes nu to forskellige studiedesigns, der udelukker hinanden, som kan have givet data fra Tabel 2.4. Design 1 er det, som blev gennemgået i forrige underafsnit, hvor der anvendes en mønt med sandsynlighed 0.65 til at bestemme, om individet bliver behandlet. Design 2 er hvor alle individer bliver klassificeret enten kritisk ( $L = 1$ ) eller ikke-kritisk ( $L = 0$ ) tilstand. Derefter bliver individer udvalgt tilfældigt med sandsynlighed 0.75 til at blive behandlet for individer, der er kritiske, og sandsynlighed 0.5 for ikke-kritiske individer. Dette betyder, at der ikke kun anvendes en mønt til at bestemme behandling, men to mønter med forskellige sandsynligheder. Dette forklarer hvorfor 9 ud af 12 individer i kritisk tilstand blev udvalgt til behandling, og hvorfor 4 ud af 8 individer i ikke-kritisk tilstand blev udvalgt.

Design 2 kaldes for *betingede randomiserede eksperimenter*, da der anvendes flere tilfældige sandsynligheder, der er betinget for værdien af  $L$ . Design 1 kaldes *marginale randomiserede eksperimenter*, hvor der kun anvendes en ubetinget (marginale) tilfældig sandsynlighed, der er fælles for alle individer.

Da det vides, at  $Y^a \perp\!\!\!\perp A$  ikke holder for data, kan det konkluderes, at dette ikke er et marginalt randomiseret eksperiment, men det vides, at det er et randomiseret eksperiment, så det er altså et randomiseret eksperiment betinget på  $L$ . Dette betingede randomiserede eksperiment er en kombination af to marginale randomiserede eksperimenter; en hvor  $L = 1$  og den anden hvor  $L = 0$ . Betragt først mængden af individer, der har  $L = 1$ . I denne delmængde er de behandlede individer og de ikke behandlede individer ombyttelige. Formelt kan dette skrives som

$$P(Y^a = 1|A = 1, L = 1) = P(Y^a = 1|A = 0, L = 1) \quad \text{eller} \quad Y^a \perp\!\!\!\perp A|L = 1 \quad \forall a,$$

hvor  $Y^a \perp\!\!\!\perp A|L = 1$  betyder, at  $Y^a$  og  $A$  er uafhængige givet  $L = 1$ . Analogt gælder også  $Y^a \perp\!\!\!\perp A|L = 0$ , og hvis det gælder for alle værdier af  $L$ , skrives  $Y^a \perp\!\!\!\perp A|L$ . Det vil sige, at selvom betinget randomisering ikke garanterer ubetinget ombyttelighed  $Y^a \perp\!\!\!\perp A$ , så garanterer det *betinget ombyttelighed*  $Y^a \perp\!\!\!\perp A|L$ .

**Definition 2.2.5** (Betinget ombyttelighed)

Der siges at være betinget ombyttelighed, når det kontrafaktiske udfald og den aktuelle behandling er uafhængige af hinanden givet en kovariat  $L$ , altså  $Y^a \perp\!\!\!\perp A|L$  for alle værdier  $a$  og  $l$ .

Nu er spørgsmålet; hvordan udregnes den kausale risiko ratio? Der er to muligheder. Først den gennemsnitlige kausale effekt udregnes for hver delmængde eller strata af populationen. Da association er kausalitet i hver strata, er den strata-specifikke kausale risiko ratio blandt individer i kritisk tilstand lig med den strata-specifikke associations risiko ratio blandt individer i kritisk tilstand, altså

$$\frac{P(Y^{a=1} = 1|L = 1)}{P(Y^{a=0} = 1|L = 1)} = \frac{P(Y = 1|L = 1, A = 1)}{P(Y = 1|L = 1, A = 0)},$$

og analogt for  $L = 0$ . Denne metode kaldes for *stratifikation*. Bemærk, at den kausale risiko ratio for  $L = 1$  kan være forskellig fra  $L = 0$ , og der siges, at effekten af behandlingen er modificeret af  $L$ , eller at der er *effektmodifikation* gennem  $L$ .

Den anden mulighed er, at udregne den gennemsnitlige kausale effekt  $P(Y^{a=1} = 1)/P(Y^{a=0} = 1)$  for hele populationen. Om den første eller anden mulighed benyttes afhænger af praktiske eller teoretiske betragtninger. Et eksempel kunne være, at det ønskes at udregne den gennemsnitlige kausale effekt for hele populationen, da der ikke forventes at have information om  $L$  for fremtidige individer (måske på grund af at  $L$  er dyr at måle). De næste to underafsnit beskriver, hvordan den gennemsnitlige kausale effekt kan udregnes for hele populationen med et betinget randomiseret eksperiment.

### 2.2.3 Standardisering

I dette underafsnit introduceres standardisering, men først ses der på et eksempel.

#### Eksempel 2.2.6

Studiet om hjertetransplantation er et betinget randomiseret eksperiment: forskerne anvendte en randomiseret procedure til at tildele hjerter ( $A = 1$ ) med sandsynlighed 0.5 til de 8 individer i ikke-kritisk tilstand ( $L = 0$ ), og med sandsynlighed 0.75 til de 12 individer i kritisk tilstand ( $L = 1$ ). Betragt de 8 individer i ikke-kritisk tilstand. Bemærk, at de er de gennemsnitlige repræsentanter af 8 milliarder individer for at undgå problemer med estimation. I denne gruppe er risikoen for dødsfald blandt de behandlede  $P(Y = 1|L = 0, A = 1) = \frac{1}{4}$ , og risikoen for dødsfald blandt dem, som ikke er behandlet, er  $P(Y = 1|L = 0, A = 0) = \frac{1}{4}$ . Da behandling var tilfældigt tildelt til individerne i gruppen  $L = 0$ , det vil sige  $Y^a \perp\!\!\!\perp A|L = 0$ , er de observerede risikoer lig med de kontrafaktiske risikoer. I gruppen  $L = 0$  er risikoen blandt de behandlede lig med risikoen, hvis alle havde været behandlet,

$$P(Y = 1|L = 0, A = 1) = P(Y^{a=1} = 1|L = 0),$$

og risikoen blandt de ubehandlede er lig med risikoen, hvis ingen havde været behandlet,

$$P(Y = 1|L = 0, A = 0) = P(Y^{a=0} = 1|L = 0).$$

Det kan konkluderes, at de observerede risikoer er lig med de kontrafaktiske risikoer i gruppen med 12 individer i kritisk tilstand, det vil sige

$$\begin{aligned} P(Y = 1|L = 1, A = 1) &= P(Y^{a=1} = 1|L = 1) = \frac{2}{3} \\ &= P(Y = 1|L = 1, A = 0) = P(Y^{a=0} = 1|L = 1) = \frac{2}{3}. \end{aligned}$$

Antag nu, at målet er at bestemme den kausale risiko ratio  $P(Y^{a=1} = 1)/P(Y^{a=0} = 1)$ . Tælleren af den kausale risiko ratio er risikoen, hvis alle 20 individer i populationen havde været behandlet. Fra ovenstående vides det, at risikoen, hvis alle individer havde været behandlet, er  $\frac{1}{4}$  for de 8 individer med  $L = 0$  og  $\frac{2}{3}$  for de 12 individer med  $L = 1$ . Derfor er risikoen, hvis alle 20 individer i populationen havde været behandlet, et vægtet gennemsnit af  $\frac{1}{4}$  og  $\frac{2}{3}$ , hvor hver gruppe får en vægt proportional til dens størrelse. Da 40% af de 20 individer (8) er i gruppe  $L = 0$ , og 60% af de 20 individer (12) er i gruppe  $L = 1$ , så er det vægtede gennemsnit  $\frac{1}{4} \cdot 0.4 + \frac{2}{3} \cdot 0.6 = 0.5$ . Derfor er risikoen, hvis alle havde været behandlet  $P(Y^{a=1} = 1)$ , lig med 0.5. Tilsvarende kan det udregnes, at risikoen, hvis ingen havde været behandlet  $P(Y^{a=0} = 1)$ , også er lig med 0.5. Den kausale risiko ratio er derfor  $0.5/0.5 = 1$ .

□

Mere formelt er den marginale kontrafaktiske risiko  $P(Y^a = 1)$  et vægtet gennemsnit af de strata-specifikke risikoer  $P(Y^a = 1|L = 0)$  og  $P(Y^a = 1|L = 1)$  med vægte lig med andelen af individer i populationen med  $L = 0$  og  $L = 1$  henholdsvis. Det vil sige, at

$$P(Y^a = 1) = P(Y^a = 1|L = 0)P(L = 0) + P(Y^a = 1|L = 1)P(L = 1).$$

Dette kan også udtrykkes med en mere kompakt notation,

$$P(Y^a = 1) = \sum_l P(Y^a = 1|L = l)P(L = l),$$

hvor  $\sum_l$  er summen over alle værdier  $l$ , som optræder i populationen. På grund af betinget ombyttelighed kan den kontrafaktiske risiko  $P(Y^a = 1|L = l)$  erstattes med den observerede risiko  $P(Y = 1|L = l, A = a)$  i udtrykket før. Det vil sige, at

$$P(Y^a = 1) = \sum_l P(Y = 1|L = l, A = a)P(L = l).$$

Venstre side af ligheden er en uobserveret kontrafaktisk risiko, hvorimod højre side kun inkluderer observerede variabler, hvilket kan bestemmes ved anvendelse af data på  $L$ ,  $A$  og  $Y$ . Når en kontrafaktisk sandsynlighed kan udtrykkes som en funktion af fordelingen af det observerede data, siges det, at den kontrafaktiske sandsynlighed er identificeret eller identificerbar; ellers siges det, at det er uidentificeret eller ikke identificerbar.



**Definition 2.2.7** (Standardisering)

Hvis betinget ombyttelighed  $Y^a \perp\!\!\!\perp A|L$  gælder, så er

$$P(Y^a = 1) = \sum_l P(Y = 1|L = l, A = a)P(L = l),$$

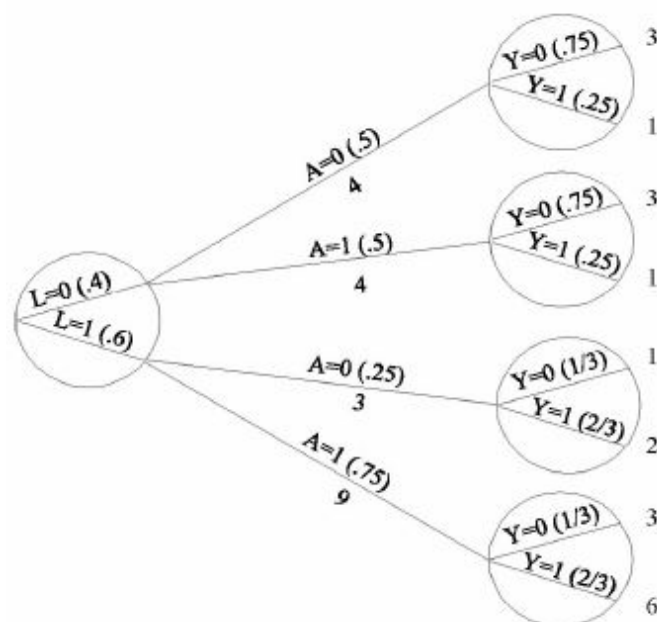
hvilket kaldes *standardisering*. Dette kommer fra loven om total sandsynlighed.

Eksempelvis er tælleren  $\sum_l P(Y = 1|L = l, A = 1)P(L = l)$  af den kausale risiko ratio den standardiserede risiko blandt de behandlede ved anvendelse af populationen som standarden. Med hensyn til betinget ombyttelighed kan denne standardiserede risiko fortolkes som risikoen (den kontrafaktiske risiko), som ville have været observeret, hvis alle individerne i populationen havde været behandlet.

De standardiserede risikoer blandt de behandlede og de ubehandlede er lig med de kontrafaktiske risikoer under behandling og under ingen behandling henholdsvis. Derfor kan den kausale risiko ratio  $\frac{P(Y^{a=1}=1)}{P(Y^{a=0}=1)}$  bestemmes ved standardisering som  $\frac{\sum_l P(Y=1|L=l, A=1)P(L=l)}{\sum_l P(Y=1|L=l, A=0)P(L=l)}$ . I næste underafsnit kigges der på en anden metode.

## 2.2.4 Invers sandsynlighedsvægtning

I forrige afsnit blev den kausale risiko ratio bestemt i et betinget randomiseret eksperiment via standardisering. Betragt nu Tabel 2.4. Dataene i denne tabel kan ses som et træ, hvor alle 20 individer starter til venstre og bevæger sig over tid mod højre, som i Figur 2.2.

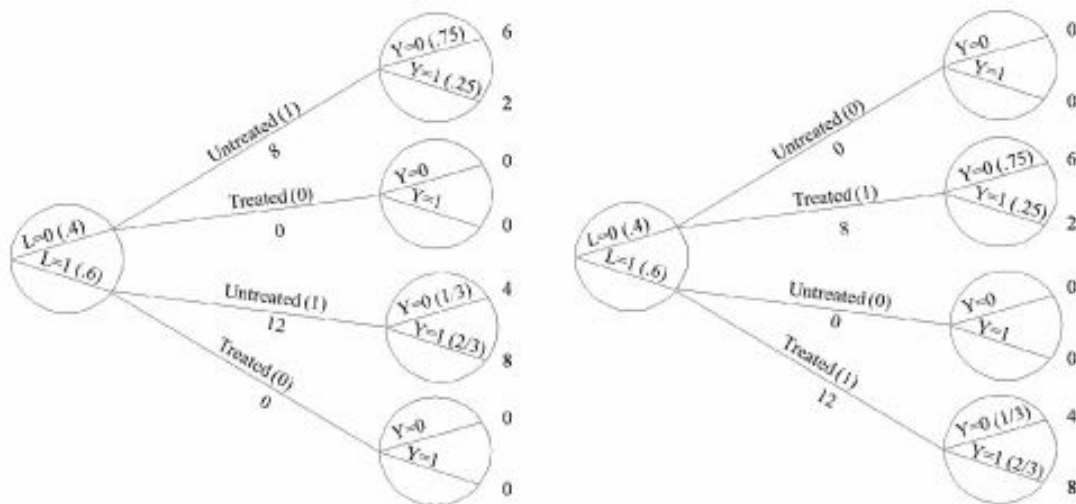


Figur 2.2: Taget fra [17] på side 20.

### Eksempel 2.2.8

Cirklen længst til venstre indeholder dens første forgrening: 8 individer var i ikke-kritisk tilstand ( $L = 0$ ) og 12 i kritisk tilstand ( $L = 1$ ). Tallene i parentes er sandsynlighederne for at være i ikke-kritisk tilstand,  $P(L = 0) = 8/20 = 0.4$ , eller kritisk tilstand,  $P(L = 1) = 12/20 = 0.6$ . Betragt eksempelvis grenen  $L = 0$ . Ud af de 8 individer i denne gren blev 4 ikke behandlet ( $A = 0$ ), og 4 blev behandlet ( $A = 1$ ). Den betingede sandsynlighed for at være ubehandlet er  $P(A = 0|L = 0) = 4/8 = 0.5$ , som vist i parentes. Den betingede sandsynlighed for at være behandlet er også  $P(A = 1|L = 0) = 4/8 = 0.5$ . Den øverste højre cirkel repræsenterer, at ud af de 4 individer i denne gren ( $L = 0, A = 0$ ) overlevede 3 ( $Y = 0$ ), og 1 døde ( $Y = 1$ ). Det vil sige, at  $P(Y = 0|L = 0, A = 0) = 3/4$  og  $P(Y = 1|L = 0, A = 0) = 1/4$ . De andre grene på træet bestemmes analogt. Cirklene indeholder forgreningerne defineret ved ikke-behandlingsvariabler. □

Dette træ anvendes nu i næste eksempel til at bestemme den kausale risiko ratio.



Figur 2.3: Taget fra [17] på side 21.

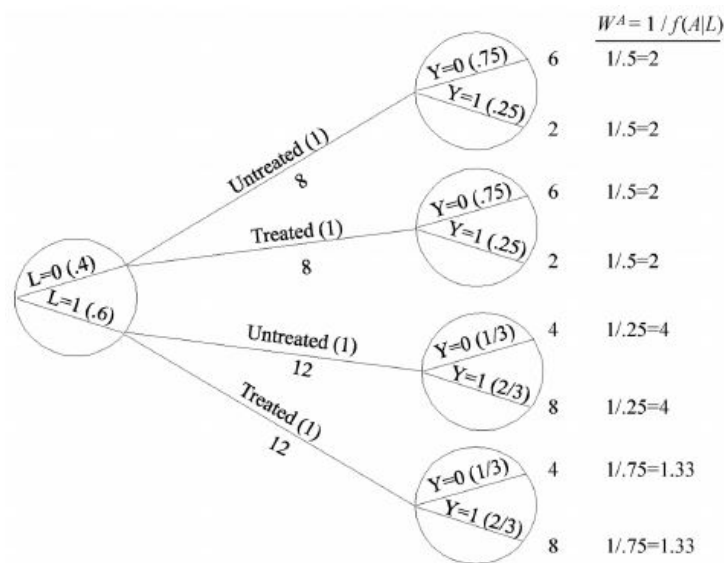
### Eksempel 2.2.9

Nævneren af den kausale risiko ratio,  $P(Y^{a=0} = 1)$ , er den kontrafaktiske risiko for dødsfald, hvis alle i populationen forblev ubehandlet. Denne risiko bestemmes nu. I Figur 2.2 var 4 ud af 8 individer med  $L = 0$  ubehandlede, og 1 af dem døde. Hvor mange dødsfald ville der ske, hvis de 8 individer med  $L = 0$  forblev ubehandlede? Svaret er 2 dødsfald på grund af, at hvis 8 individer i stedet for 4 individer forblev ubehandlede, ville 2 dødsfald i stedet for 1 dødsfald være blevet observeret. Hvis antallet af individer fordobles ville antallet af dødsfald også fordobles. I Figur 2.2 var 3 ud af 12 individer med  $L = 1$  ubehandlede, og 2 af dem døde. Hvor mange dødsfald ville der ske, hvis de 12 individer med  $L = 1$  forblev ubehandlede? Svaret er 8 dødsfald, eller 2 dødsfald gange 4, da 12 er  $3 \cdot 4$ . Det vil sige, at hvis alle  $8 + 12 = 20$  individer i populationen havde været ubehandlede, ville  $2 + 8 = 10$  have været døde. Nævneren af den

kausale risiko ratio,  $P(Y^{a=0} = 1)$ , er  $10/20 = 0.5$ . Det første træ i Figur 2.3 viser populationen, hvis alle forblev ubehandlet. Disse udregninger afhænger af betingelsen, at behandlede individer med  $L = 0$ , hvis de forblev ubehandlet, ville have haft den samme sandsynlighed for dødsfald som dem, der faktisk forblev ubehandlet. Denne betingelse er ombyttelighed givet  $L = 0$ .

Tælleren af den kausale risiko ratio,  $P(Y^{a=1} = 1)$ , er den kontrafaktiske risiko for dødsfald, hvis alle i populationen havde været behandlet. Ud fra samme argumentation som ovenstående bestemmes risikoen til  $10/20 = 0.5$ , under ombyttelighed givet  $L = 1$ . Det andet træ i Figur 2.3 viser populationen, hvis alle havde været behandlet. Ved at kombinere disse resultater med de ovenstående er den kausale risiko ratio  $P(Y^{a=1} = 1)/P(Y^{a=0} = 1)$  er lig med  $0.5/0.5 = 1$ .  $\square$

Nu uddybes metoden. De to træer i Figur 2.3 er en simulation af, hvad der ville ske, hvis alle individer i populationen havde været henholdsvis ubehandlet og behandlet. Disse simulationer er korrekte under betinget ombyttelighed. Begge simulationer kan samles for at lave en hypotetisk population, hvor alle individer fremstår som et behandlet og et ubehandlet individ. Denne hypotetiske population, som er dobbelt så stor som den originale population, er kendt som *pseudo populationen*. Figur 2.4 viser hele pseudo populationen. Under betinget ombyttelighed  $Y^a \perp\!\!\!\perp A|L$  i den originale population er de behandlede og de ubehandlede (ubetinget) ombyttelige i pseudo populationen, da  $L$  er uafhængigt af  $A$ . Det vil sige, at associations risiko ratioen i pseudo populationen er lig med den kausale risiko ratio i både pseudo populationen og den originale population.



Figur 2.4: Taget fra [17] på side 22.

Denne metode er kendt som *invers sandsynlighed (IP, inverse probability) vægning*. Betragt

de 4 ubehandlede individer med  $L = 0$  i populationen på Figur 2.2. Disse individer anvendes til at lave 8 medlemmer af pseudo populationen på Figur 2.4. Det vil sige, at hver af dem får en vægt på 2, hvilket er lig med  $1/0.5$ . Figur 2.2 viser, at 0.5 er den betingede sandsynlighed af at forblive ubehandlet givet  $L = 0$ . Tilsvarende er de 9 behandlede individer med  $L = 1$  i Figur 2.2 anvendt til at lave 12 medlemmer af pseudo populationen. Det vil sige, at hver af dem får en vægt på  $1.33 = 1/0.75$ . Figur 2.2 viser, at 0.75 er den betingede sandsynlighed for at være behandlet givet  $L = 1$ . Uformelt er pseudo populationen lavet ved at vægte hvert individ i populationen ved den inverse af den betingede sandsynlighed for at få behandlingsniveaet, som de faktisk fik. Disse IP vægte er vist i Figur 2.4.

**Definition 2.2.10** (IP vægtning)

Når der haves behandlingsværdi  $A = a$  og kovariat  $L = l$ , så er et individs IP vægt lig med

$$\frac{1}{P(A = a|L = l)},$$

i det diskrete tilfælde.

Disse vægte kan udtrykkes ved at anvende et enkelt udtryk for alle individer, uanset deres individuelle behandling og kovariatværdier, ved at anvende pdf af  $A$  i stedet for sandsynligheden af  $A$ . Den betingede pdf af  $A$  givet  $L$  evalueret ved værdierne  $a$  og  $l$  er repræsenteret ved  $f_{A|L}(a|l)$ , eller ved  $f(a|l)$ . For diskrete variable  $A$  og  $L$  er  $f(a|l)$  den betingede sandsynlighed  $P(A = a|L = l)$ . I et betinget randomiseret eksperiment er  $f(a|l)$  positiv for alle  $l$ , således at  $P(L = l)$  er ikke-nul.

Da nævneren af vægten for hvert individ er den betingede tæthed evalueret ved individets egne værdier af  $A$  og  $L$ , kan det udtrykkes som den betingede tæthed evalueret ved de stokastiske argumenter  $A$  og  $L$  (i modsætning til de faste argumenter  $a$  og  $l$ ) som  $f(A|L)$ . Denne notation, som blev anvendt i Figur 2.4, er anvendt til at definere IP vægtene  $W^A = 1/f(A|L)$ . Det er nødvendigt at have en samlet notation for vægtene, da  $P(A = A|L = L)$  ikke anses som at være en korrekt notation.

IP vægtning gav det samme resultat som standardisering (kausal risiko ratio lig med 1) i Eksempel 2.2.9. Dette er ikke tilfældigt: standardisering og IP vægtning er matematisk ækvivalente (se bevis for Sætning 2.2.11). Faktisk kan både standardisering og IP vægtning ses som procedurer til at lave et nyt træ, hvor alle individer får behandling  $a$ . Hver metode anvender en forskellig mængde af sandsynligheder for at lave det kontrafaktiske træ: IP vægtning anvender den betingede sandsynlighed af behandling  $A$  givet kovariaten  $L$ , som vist i Figur 2.2, standardisering anvender sandsynligheden af kovariaten  $L$  og den betingede sandsynlighed af udfald  $Y$  givet  $A$

og  $L$ .

Da både standardisering og IP vægtning simulerer det, der ville have været observeret, hvis variabelen (eller variabler i vektoren)  $L$  ikke havde været anvendt til at bestemme sandsynligheden af behandling, siges det ofte, at disse metoder *justeres for  $L$* . Nogle gange siges det, at disse metoder *kontrollerer for  $L$* , men denne ”analytiske kontrol” er forskellig fra den ”fysiske kontrol” i et randomiseret eksperiment. Standardisering og IP vægtning kan generaliseres til betingede randomiserede studier med kontinuerte udfald (se bevis for Sætning 2.2.11).

### Sætning 2.2.11

For både standardisering og IP vægtning gælder

$$E[Y^a] = \sum_l E[Y|A = a, L = l]P(L = l) = E \left[ \frac{I(A = a)}{f(A|L)} Y \right],$$

hvilket giver ækvivalens af standardisering og IP vægtning.

#### Bevis

Antag, at  $A$  er diskret med et endeligt antal værdier, og at  $f(a|l)$  er positiv for alle  $l$ , således at  $P(L = l)$  er forskellig fra nul. Denne positive betingelse holder med garanti i betingede randomiserede eksperimenter. Under positivitet er det standardiserede middelværdi for behandlingsniveau  $a$  defineret som  $\sum_l E[Y|A = a, L = l]P(L = l)$ , og det IP vægtede middelværdi af  $Y$  for behandlingsniveau  $a$  er defineret som  $E \left[ \frac{I(A=a)Y}{f(A|L)} \right]$ . Det vil sige, at middelværdien af  $Y$  er vægtet igen af IP vægten  $W^A = 1/f(A|L)$  for individer med behandlingsværdi  $A = a$ . Indikatorfunktionen  $I(A = a)$  er funktionen, som tager værdi 1 for individer med  $A = a$  og 0 for de andre.

Nu vises ligheden mellem den IP vægtede middelværdi og den standardiserede middelværdi under positivitet. Ud fra definitionen af forventet værdi gælder

$$\begin{aligned} E \left[ \frac{I(A = a)}{f(A|L)} Y \right] &= \sum_{a,l} E \left[ \frac{I(A = a)}{f(A|L)} Y | A = a, L = l \right] P(A = a, L = l) \\ &= \sum_l E \left[ \frac{1}{f(A|L)} Y | A = a, L = l \right] P(A = a | L = l) P(L = l) \\ &= \sum_l \frac{1}{f(a|l)} E[Y | A = a, L = l] f(a|l) P(L = l) \\ &= \sum_l E[Y | A = a, L = l] P(L = l), \end{aligned}$$

hvor  $f(a|l)$  reduceres væk i sidste lighed. I anden lighed er det ikke nødvendigt at summere over de mulige værdier af  $A$ , da der gælder, at for ethvert  $a' \neq a$  er  $I(a' = a)$  lig med nul.

I anden lighed benyttes betinget sandsynlighed på  $P(A = a, L = l)$ . I tredje lighed anvendes definitionen på  $f(a|l)$ . Beviset behandler  $A$  og  $L$  som diskrete, men ikke nødvendigvis binære. For kontinuerte  $L$  erstattes summen over  $L$  med et integral.

Beviset refererer ikke til kontrafakta eller kausalitet. Hvis der derfor antages betinget ombyttelighed, er både den IP vægtede middelværdi og den standardiserede middelværdi lig med den kontrafaktiske middel  $E[Y^a]$ . Der gives to forskellige beviser for det sidstnævnte udsagn. Først bevises lighed mellem  $E[Y^a]$  og den standardiserede middelværdi

$$\begin{aligned} E[Y^a] &= \sum_l E[Y^a|L = l]P(L = l) \\ &= \sum_l E[Y^a|A = a, L = l]P(L = l) \\ &= \sum_l E[Y|A = a, L = l]P(L = l), \end{aligned}$$

hvor andet lighedstegn gælder på grund af betinget ombyttelighed og positivitet, og det tredje lighedstegn gælder på grund af konsistens. Hernæst bevises ligheden mellem  $E[Y^a]$  og den IP vægtede middelværdi som følgende:  $E\left[\frac{I(A=a)}{f(A|L)}Y\right] = E\left[\frac{I(A=a)}{f(A|L)}Y^a\right]$  på grund af konsistens. Da positivitet medfører  $f(a|L)$  aldrig er 0, haves

$$\begin{aligned} E\left[\frac{I(A=a)}{f(A|L)}Y^a\right] &= E\left[E\left[\frac{I(A=a)}{f(a|L)}Y^a\middle|L\right]\right] \\ &= E\left[E\left[\frac{I(A=a)}{f(a|L)}\middle|L\right]E[Y^a|L]\right] \\ &= E[E[Y^a|L]] \\ &= E[Y^a], \end{aligned}$$

hvor andet lighedstegn gælder på grund af betinget ombyttelighed, og tredje lighedstegn gælder på grund af, at  $E\left[\frac{I(A=a)}{f(a|L)}\middle|L\right] = 1$ .

Når behandling er kontinuert, hvilket er et usandsynligt designvalg i betingede randomiserede eksperimenter, er  $E[I(A = a)Y/f(A|L)] \neq \sum_l E[Y|A = a, L = l]P(L = l)$  og er dermed biased for  $E[Y^a]$  selv under ombyttelighed. For at se dette, kan det udregnes, at  $E[I(A = a)/f(a|l)|L = l]$  er lig med 0 frem for 1, hvis  $f(a|l)$  tages for at være (en version af) den betingede tæthed af  $A$  givet  $L = l$  (med hensyn til Lebesguemål). På den anden side, hvis der fortsættes med at tage  $f(a|l)$  til at være  $P(A = a|L = l)$ , så er nævneren  $f(a|L = l)$  lig med nul på en mængde med sandsynlighed 1, så positivitet holder ikke. ■

Hvorfor ikke afslutte her? Der haves et studiedesign (et ideelt randomiseret eksperiment), som tillader at bestemme gennemsnitlige kausale effekter, når der kombineres med den passende

analytiske metode (standardisering eller IP vægtning). Desværre er randomiserede eksperimenter ofte uetiske, upraktiske eller utidige. For eksempel er det tvivlsomt, at en etisk komité ville have godkendt hjertetransplantationsstudiet, som tidligere er beskrevet. Der er mangel på hjerter, og samfundet går ind for at tildele dem til individer, der er mere tilbøjelige til at drage fordel af transplantationen, frem for at tildele dem tilfældigt blandt potentielle modtagere. Der kunne også stilles spørgsmålstejn ved gennemførligheden af studiet, selvom etiske spørgsmål blev ignoreret: dobbeltblind tildeling er umulig, da personer, der er tildelt medicinsk behandling, ikke må sige sig selv fra til at få en transplantation, og der er muligvis ikke kompatible hjerter for dem, der er tildelt transplantation. Selv hvis studiet var gennemførligt, ville det stadig tage flere år at gennemføre det, og der skal træffes beslutninger i mellemtiden. Ofte er det at udføre et observationsstudie den mindst dårlige mulighed.

## 2.3 Observationsstudier

Dette afsnit er baseret på kapitel 3 i [17], hvis ikke andet er angivet.

I stedet for et randomiseret eksperiment kan der laves et observationsstudie. Med udgangspunkt i eksperimentet: ”om andre kigger op, når der er en, der gør det”, før havde det været den, der lavede eksperimentet, der kiggede op først, men nu observeres en population, hvor to individer udvælges, og der ses på, om de kigger op, hvis den anden gør det. Der er nogle betingelser, der skal opfyldes, før disse observationsstudier kan benyttes til at lave kausal inferens, hvilket gennemgås i dette afsnit.

### 2.3.1 Identificerbarhedsbetingelser

Observationsstudier, til forskel fra randomiserede eksperimenter, kan ikke bare benyttes til at beregne den gennemsnitlige kausale effekt. Hovedgrunden til, at observationsassociationer ikke gives en kausal fortolkning er, at den mangler en randomiseret tildeling af behandling. Et eksempel kan være hjertetransplantation, hvor dem, der fik operationen, havde en større risiko for at dø, hvis de ikke fik operationen end dem, der slet ikke fik operationen. Som en konsekvens vil en associations risiko ratio på 1.1 fra et observationsstudie være et kompromis mellem den virkelige gavnlige effekt af hjertetransplantation på dødelighed (hvilket gør, at associations risiko ratioen vil være under 1) og den underliggende større dødelighedsrisiko af dem, der får hjertetransplantation (hvilket gør, at associations risiko rationen er over 1). Så den bedste forklaring på association mellem behandling og udfald i et observationsstudie er ikke nødvendigvis kausalitet.

Randomiserede eksperimenter har en indbygget fordel til kausal inferens, men sommetider er det nødvendigt at anvende et observationsstudie til at besvare kausale spørgsmål. For at gøre dette analyseres data som om, at behandlingen var randomiseret og betinget på en målt kovariat  $L$ , og dette er en approksimation som det bedste. Kausal inferens i observationsstudier kommer fra håbet om, at studiet kan ses som et betinget randomiseret eksperiment.

Uformelt kan et observationsstudie konceptualiseres som et betinget randomiseret eksperiment, hvis følgende betingelser holder:

1. Værdierne af behandling under sammenligning svarer til veldefinerede indgreb, der til gengæld svarer til versioner af behandling i data.
2. Ombyttelighed  $Y^a \perp\!\!\!\perp A|L$  skal holde.
3. Sandsynligheden for at få enhver værdi af behandling betinget på  $L$  er større end nul, altså positiv.

I dette afsnit beskrives disse tre betingelser i sammenhæng med observationsstudier. Betingelse 1 er tidligere refereret til som konsistens, og betingelse 3 er refereret til som positivitet i Eksempel 2.2.11.

Hvis analogien mellem observationsstudier og betingende randomiserede eksperimenter er sand, kan metoder fra tidligere afsnit benyttes (IP vægting og standardisering) til at identificere kausale effekter fra observationsstudiet. Disse 3 betingelser kaldes derfor *identificerbarhedsbetingelser* eller *identificerbarhedsantagelser*.

### 2.3.2 Ombyttelighed

I observationsstudier når behandling ikke er tilfældigt fordelt af forskerne, så er grunden til, at individerne får behandlingen sandsynligvis associeret med en af prædiktorerne. Ligesom i et betinget randomiseret eksperiment vil fordelingen af prædiktorerne variere mellem den behandlede gruppe og den ubehandlede gruppe i et observationsstudie. For eksempel kan data fra Tabel 2.4 være kommet fra et observationsstudie, hvor lægerne giver de sjældne hjerter til dem, der har mest brug for dem  $L = 1$ . Den eneste prædiktor, der ikke er ligeligt fordelt mellem de behandlede og ubehandlede, er  $L$ , så data fra Tabel 2.4 kan beskrives på to måder. Første måde er et observationsstudie, hvor sandsynligheden for at blive behandlet  $A = 1$  er 0.75 blandt dem, der har  $L = 1$ , og 0.5 blandt dem, der har  $L = 0$ . Den anden måde er, at det er et ikke-blindt betinget randomiseret eksperiment, hvor forskerne tilfældigt tildelte behandling  $A = 1$  med sandsynlighed 0.75 blandt dem, der har  $L = 1$ , og 0.5 blandt dem, der har  $L = 0$ . Under begge disse karakteriseringer holder betinget ombyttelighed  $Y^a \perp\!\!\!\perp A|L$ , og standardisering eller IP



vægtning kan benyttes til at identificere kausal effekt.

Det afgørende spørgsmål for et observationsstudie er, om  $L$  er den eneste prædikator, der ikke er ligeligt fordelt blandt de behandlede og ubehandlede. Desværre må dette spørgsmål forblive ubesvaret. Da spørgsmålet ikke kan besvares, må forskerne arbejde under antagelsen, at betinget ombyttelighed  $Y^a \perp\!\!\!\perp A|L$  holder.

Det vigtige ord her er antagelsen. Uanset hvor god antagelsen er, som forskerne er kommet frem til baseret på viden omkring studiet, er der ikke nogen randomisering, og derfor ikke nogen garanti for, at betinget ombyttelighed holder. Et eksempel kan være, at lægerne foretrækker at transplantere hjerte til individer, der ikke ryger. Så hvis to individer med  $L = 1$  har de rigtige HLA gener, som bestemmer, om individet kan modtage et hjerte, men en af dem ryger  $U = 1$ , og den anden ryger ikke  $U = 0$ , så har det individ, der har  $U = 1$ , mindre sandsynlighed for at modtage behandling  $A = 1$ . Når fordelingen af rygning, som er en vigtig prædikator for overlevelse, er forskellig mellem de behandlede (med mindre proportion af  $U = 1$ ) og de ubehandlede (med større proportion af  $U = 1$ ) i strata af  $L = 1$ , så holder  $Y^a \perp\!\!\!\perp A|L$  ikke. Det er vigtigt at vide, at indsamling af data på rygning ikke forhindrer, at der er en mulighed for, at der er en anden ukendt ubalanceret prædikator, som ikke er blevet målt.

Forskere kan benytte deres ekspertviden til at forøge sandsynligheden for, at antagelsen  $Y^a \perp\!\!\!\perp A|L$  holder. De kan måle mange relevante variabler  $L$  og ikke kun den ene variabel i Tabel 2.4 og derefter antage, at  $Y^a \perp\!\!\!\perp A|L$  er approksimativt sandt inden for strata defineret med kombinationen af alle disse variabler  $L$ . Desværre uanset hvor mange variabler der er inkluderet i  $L$ , er der ikke nogen metode til at kunne teste om antagelsen er sand, hvilket gør, at kausal inferens fra et observationsstudie er en risikabel opgave. Gyldigheden af de kausale inferenser kræver, at forskeres ekspertviden er korrekt.

### 2.3.3 Positivitet

Nogle forskere planlægger at udføre et eksperiment for at beregne den gennemsnitlige effekt af hjertetransplantation  $A$  på 5-års dødelighed  $Y$ . Det siger sig selv, at forskere vil tildele nogle individer at modtage behandlingsniveau  $A = 1$  og andre  $A = 0$ . Betragt det alternative: forskere tildeler alle individer enten  $A = 1$  eller  $A = 0$ . Det ville ikke være klogt. Da alle individer modtager den samme behandlingsværdi, ville det være umuligt at bestemme den gennemsnitlige kausale effekt. I stedet skal behandling tildeles, således at nogle individer med næsten sikkerhed vil blive tildelt hver af behandlingsgrupperne. Med andre ord skal det sikres, at der er en sandsynlighed større end nul for at blive tildelt hver af behandlingsniveauerne. Dette er betingelsen for *positivitet*.

Der blev ikke lagt vægt på positivitet, når eksperimenter blev beskrevet, da positivitet tages for givet i disse studier. I marginale randomiserede eksperimenter er sandsynlighederne  $P(A = 1)$  og  $P(A = 0)$  begge positive af design. I betingede randomiserede eksperimenter er de betingede sandsynligheder  $P(A = 1|L = l)$  og  $P(A = 0|L = l)$  også positive af design for alle niveauer af variablen  $L$ , som er egnede til studiet. For eksempel, hvis data i Tabel 2.2 var opstået fra et betinget randomiseret eksperiment, ville de betingede sandsynligheder for tildeling til hjertetransplantation have været  $P(A = 1|L = 1) = 0.75$  for individer i kritisk tilstand, og  $P(A = 1|L = 0) = 0.5$  for de andre individer. Positivitet holder, betinget på  $L$ , da ingen af disse sandsynligheder er 0 (eller 1, hvilket ville medføre, at sandsynligheden for ingen hjertetransplantation  $A = 0$  ville være 0). Det siges, at der er positivitet, hvis  $P(A = a|L = l) > 0$  for alle  $a$  involveret i den kausale kontrast. Faktisk er denne definition af positivitet ikke fuldstændig på grund af, at hvis studiepopulationen var begrænset til gruppen  $L = 1$ , ville der ikke være behov for at kræve positivitet i gruppen  $L = 0$ . Positivitet er kun nødvendig for værdier  $l$ , som er til stede i populationen. Formelt kan positivitet skrives som i følgende definition.

**Definition 2.3.1** (Positivitet)

Positivitet gælder når

$$P(A = a|L = l) > 0 \quad \forall a, l,$$

med  $P(L = l) \neq 0$  for alle værdier  $l$ .

Derudover kræves der kun positivitet for variablerne  $L$ , som er krævet for ombyttelighed. For eksempel, i det betingede randomiserede eksperiment i Tabel 2.2 spørges der ikke, om sandsynligheden for at modtage behandling er større end 0 hos individer med blå øjne, da variablen ”at have blå øjne” ikke er nødvendig for at opnå ombyttelighed mellem den behandlede og den ubehandlede. (Variablen ”at have blå øjne” er ikke en uafhængig prædikator af udfaldet  $Y$  betinget på  $L$  og  $A$  og var ikke anvendt til at tildele behandling). Det vil sige, at den standardiserede risiko og den IP vægtede risiko er lig med den kontrafaktiske risiko efter justering for kun  $L$ ; positivitet gælder ikke for variabler, der ligesom ”at have blå øjne” ikke skal justeres for.

I observationsstudier er hverken positivitet eller ombyttelighed garanteret. For eksempel holder positivitet ikke, hvis læger altid transplanterer et hjerte til individer i kritisk tilstand  $L = 1$ , altså hvis  $P(A = 0|L = 1) = 0$ , som vist i Figur 2.5. En forskel mellem betingelserne ombyttelighed og positivitet er, at positivitet nogle gange kan være empirisk verificeret. For eksempel, hvis Tabel 2.2 svarede til data fra et observationsstudie, konkluderes det, at positivitet holder for

$L$ , da der er individer ved alle behandlingsniveauer ( $A = 0$  og  $A = 1$ ) ved hvert niveau af  $L$  ( $L = 0$  og  $L = 1$ ). Diskussionen om standardisering og IP vægtning tidligere var eksplicit omkring betingelsen for ombyttelighed, men betingelsen for positivitet blev kun antaget implicit (eksplicit i beviset for Sætning 2.2.11). Definitionerne af standardiseret risiko og IP vægtet risiko giver faktisk kun mening, når positivitet holder.

### Proposition 2.3.2

Positivitet er en forudsætning for standardisering og IP vægtning.

*Bevis*

Den standardiserede middelværdi for behandlingsniveau  $a$  er defineret som

$$\sum_l E[Y|A = a, L = l]P(L = l).$$

Dette udtryk kan dog kun bestemmes, hvis  $E[Y|A = a, L = l]$  er veldefineret, hvilket er tilfældet, når den betingede sandsynlighed  $P(A = a|L = l) > 0$  for alle værdier  $l$ , som optræder i populationen. Det vil sige, når positivitet holder. (Bemærk, at udtrykket  $P(A = a|L = l) > 0$  for alle  $l$  med  $P(L = l) \neq 0$  kan ses som at være ækvivalent til  $f(a|L) > 0$  med sandsynlighed 1). Derfor er den standardiserede middelværdi defineret som

$$\sum_l E[Y|A = a, L = l]P(L = l) \quad \text{hvis} \quad P(A = a|L = l) > 0 \quad \text{for alle } l \text{ med } P(L = l) \neq 0,$$

og er udefineret ellers. Den standardiserede middelværdi kan kun bestemmes, hvis der for hver værdi af kovariaten  $L$  i populationen er nogle individer, der modtog behandlingsniveauet  $a$ .

Den IP vægtede middelværdi  $E\left[\frac{I(A=a)Y}{f(A|L)}\right]$  er ikke længere lig med  $E\left[\frac{I(A=a)Y}{f(a|L)}\right]$ , når positivitet ikke holder. Specifikt er  $E\left[\frac{I(A=a)Y}{f(a|L)}\right]$  udefineret, da  $\frac{0}{0}$  optræder, når den forventede værdi beregnes. På den anden side er den IP vægtede middelværdi  $E\left[\frac{I(A=a)Y}{f(A|L)}\right]$  altid veldefineret, da nævneren  $f(A|L)$  aldrig kan være nul. Dette er dog et biased estimat af den kontrafaktiske middelværdi selv under ombyttelighed. Når positivitet ikke holder, så gælder

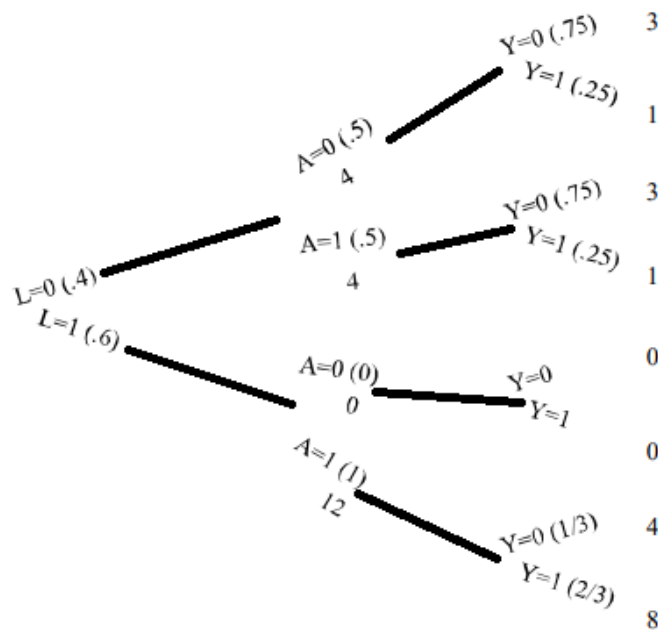
$$\begin{aligned} E\left[\frac{I(A=a)Y}{f(A|L)}\right] &= \sum_{l,a} E\left[\frac{1}{f(A|L)}Y|A = a, L = l, L \in Q(a)\right] P(A = a, L = l, L \in Q(a)) \\ &= \sum_l \frac{1}{f(a|l)} E[Y|A = a, L = l, L \in Q(a)] P(A = a|L = l, L \in Q(a)) \\ &\quad \cdot P(L = l|L \in Q(a))P(L \in Q(a)) \\ &= P(L \in Q(a)) \sum_l E[Y|A = a, L = l, L \in Q(a)]P(L = l|L \in Q(a)), \end{aligned}$$

hvor  $Q(a) = \{l; P(A = a|L = l) > 0\}$  er mængden af værdier  $l$ , hvor  $A = a$  kan observeres med positiv sandsynlighed. Under ombyttelighed gælder der derfor, at

$$E \left[ \frac{I(A = a)Y}{f(A|L)} \right] = E[Y^a | L \in Q(a)]P(L \in Q(a)).$$

Fra definitionen af  $Q(a)$  kan  $Q(0)$  ikke være lig med  $Q(1)$  når  $A$  er binær, og positivitet ikke gælder. I dette tilfælde har kontrasten  $E \left[ \frac{I(A=1)Y}{f(A|L)} \right] - E \left[ \frac{I(A=0)Y}{f(A|L)} \right]$  ingen kausal fortolkning, selv under ombyttelighed, da det er en kontrast mellem to forskellige grupper. Under positivitet gælder, at  $Q(1) = Q(0)$ , og kontrasten er den gennemsnitlige kausale effekt hvis ombyttelighed holder. ■

For intuitivt at forstå hvorfor den standardiserede og IP vægtede risiko ikke er veldefineret, når betingelsen for positivitet fejler, betragt Figur 2.5. Hvis der ingen ubehandlede individer var ( $A = 0$ ) med  $L = 1$ , ville data ikke indeholde information til at simulere, hvad der ville være sket, hvis alle behandlede individer havde været ubehandlede, da der ikke ville være nogen ubehandlede individer med  $L = 1$ , som kunne betragtes som ombyttelige med de behandlede individer med  $L = 1$ .



Figur 2.5: Taget fra [17] på side 31.

### 2.3.4 Konsistens: Først, definér det kontrafaktiske udfald

Konsistens betyder, at  $Y^a = Y$  for alle individer med  $A = a$ . Dette udsagn er tydeligvis sandt, så er der nogle situationer, hvor konsistens ikke holder? Hvis et individ tager aspirin  $A = 1$  og dør ( $Y = 1$ ), er det så ikke tilfældet, at udfaldet  $Y^{a=1}$  under aspirin også er lig med 1? Den tilsyneladende enkelhed af konsistensbetingelsen er vildledende. Derfor beskrives konsistens med

to hovedkomponenter: (1) en præcis definition af de kontrafaktiske udfald  $Y^a$  via en detaljeret specification af  $a$ , og (2) koblingen mellem de kontrafaktiske udfald og de observerede udfald. Dette underafsnit omhandler den første komponent af konsistens.

Betragt igen et randomiseret eksperiment til at bestemme den kausale effekt af hjertetransplantation  $A$  på 5-års dødelighed  $Y$ . Inden der blev indskrevet patienter i studiet, skrev forskere en protokol, hvori de to interventioner - hjertetransplantation  $A = 1$  og medicinsk lægebehandling  $A = 0$  - blev beskrevet i detaljer. For eksempel specificerede forskere, at individer, der blev tildelt hjertetransplantation  $A = 1$ , skulle modtage visse præoperative procedurer, anæstesi, kirurgisk teknik, postoperativ behandling og immunsuppressiv terapi. Hvis protokollen ikke havde specificeret disse detaljer, er det muligt, at hver læge havde udført en anden version af behandlingen "hjertetransplantation" måske ved hjælp af sin foretrukne kirurgiske teknik eller immunsuppressiv terapi.

Der opstår et problem, hvis forskellige versioner af behandlingen har forskellige kausale effekter. For eksempel kan den gennemsnitlige kausale effekt af "hjertetransplantation" i et studie, hvor de fleste læger anvendte en traditionel kirurgisk teknik, afvige fra den i et studie, hvor de fleste læger anvendte en ny kirurgisk teknik. Derfor, når der refereres til "den kausale effekt af hjertetransplantation  $A$  på dødelighed", er det nødvendigt at specificere versionerne  $a$  af behandling  $A$ . Hvis behandlingsværdierne  $a$  ikke er veldefinerede, er de kontrafaktiske udfald  $Y^a$  ikke veldefinerede, hvilket betyder, at  $P(Y^{a=1} = 1) - P(Y^{a=0} = 1)$  ikke er veldefineret. Ideelt set vil protokollerne for randomiserede eksperimenter præcist specificere de behandlingsværdier  $a$ , der er tildelt hvert individ, så deres kontrafaktiske udfald  $Y^a$  er veldefinerede. I observationsstudier er forskere nødt til at specificere værdierne  $a$  så entydigt som muligt. Selvom dette er forholdsvis ligetil for medicinske indgreb, såsom hjertetransplantation, er det meget sværere for behandlinger, der ikke svarer til faktiske indgreb i den virkelige verden.

Det er ikke nødvendigt at have en absolut præcis definition af behandlingen for brugbar kausal inferens. For eksempel, hvad angår den kausale virkning af motion, er forskere enige om, at fordelene ved at løbe med uret rundt om en bane er de samme som ved at løbe mod uret rundt. Derfor, når man beskriver behandlingen "livsvarigt motion", er det ikke nødvendigt at specificere retningen for løbeturen. Dette og andre aspekter af behandlingen anses for at være irrelevante, da en variation af dem ikke ville føre til forskellige kontrafaktiske udfald. Det vil sige, at kun *tilstrækkeligt veldefinerede* indgreb  $a$  er nødvendigt.

### 2.3.5 Konsistens: For det andet, forbind kontrafakta til det observerede data

Betragt en forsker, der er interesseret i følgende indgreb ( $a = 1$ ): ”i en alder af 18 og op til 40 år sættes hver person på en streng obligatorisk diæt, der garanterer, at de aldrig vejer mere end deres vægt i en alder af 18 år”. Hvert individ vejes hver dag efter individet fylder 18 år. Hvis et individs vægt bliver større end baselinevægten efter de første 18 år, bliver individets kalorieindtag begrænset uden at ændre den sædvanlige blanding af kaloriekilder og mikronæringsstoffer indtil det tidspunkt (normalt inden for 1-3 dage), hvor individet falder under baselinevægten. Hvis fejl på et kilogram eller to ignoreres, vil ingen individer nogensinde veje mere end baselinevægten indtil 40 år. Der gives ingen instruktioner eller begrænsninger vedrørende motion på noget tidspunkt eller diæt i perioder med ikke-kaloriebegrænsning. Sammenligningsindgrebet ( $a = 0$ ) er ”grib ikke ind”.

Antag, at der er enighed om, at behandlingsværdierne  $a = 1$  og  $a = 0$  er tilstrækkelig veldefinerede, og derfor at der ikke er meningsfuld uklarhed tilbage i specifikationen af de kontrafaktiske udfald  $Y^{a=1}$  og  $Y^{a=0}$ . Der fokuseres nu på lighedstegnet i konsistensbetingelsen  $Y^a = Y$  for individer med  $A = a$ .

Betragt Ares, som holdt en approksimativt konstant vægt mellem 18 og 40 år på trods af, at han ikke modtog indgrebet, altså  $a = 0$ . Tværtimod bevarede Ares sin baselinevægt. Hans observerede behandlingsværdi var ikke  $A = 1$ , og derfor svarer hans observerede udfald  $Y$  ikke nødvendigvis til det kontrafaktiske udfald  $Y^{a=1}$ , hvilket ville være tilfældet, hvis han modtog indgrebet  $a = 1$ .

For at bevare sammenhængen mellem de kontrafaktiske udfald  $Y^{a=1}$  og de observerede udfald  $Y$  skal vi sikre, at kun individer, der modtager behandlingsversion  $a = 1$ , skal betragtes som behandlede individer ( $A = 1$ ) i analysen, og tilsvarende for de ubehandlede. Implikationen er, at hvis det ønskes at kvantificere den kausale effekt  $P(Y^{a=1} = 1) - P(Y^{a=0} = 1)$  ved anvendelse af observationsdata, er der behov for data, hvor nogle individer modtog behandlingsværdier konsistente med  $a = 1$  og  $a = 0$ , altså er der behov for (ubetinget) positivitet. Det er ikke nyttigt at beskrive et veldefineret indgreb  $a$ , hvis indgrebet ikke kan kobles til de observerede data, det vil sige, hvis ikke det med rimelighed kan antages, at ligheden  $Y^a = Y$  holder for mindst nogle individer.

Begrænsningen til behandlingsværdien  $a$  er dog umulig, hvis data ikke er tilstrækkelig. Dette problem kan opstå i eksempelvis et studie om overvægt, der indsamler data om kropsvægt i alderen 40 år, men ingen data om individets fortid med vægt, motion og diæt.

En vej ud af dette problem er at antage, at effekterne af alle versioner af behandling er identiske - det vil sige, hvis der er *irrelevans af behandlingsvariation*. I nogle tilfælde kan det være en god approksimation. For eksempel, hvis man er interesseret i den kausale effekt af højt blodtryk i forhold til normalt blodtryk ved stroke, så tyder empiriske beviser på, at sænkning af blodtrykket gennem forskellige farmakologiske mekanismer resulterer i lignende udfald. Derfor kan der argumenteres for, at en præcis definition af behandlingen "blodtryk" er unødvendig for at forbinde de potentielle og observerede udfald. I andre tilfælde er gyldigheden af antagelsen mere tvivlsom. For eksempel, hvis der er interesse for den gennemsnitlige kausale effekt af vedligeholdelse af vægt på dødsfald, så tyder empiriske beviser på, at nogle indgreb ville øge risikoen (eksempelvis fortsættelse af rygning), mens andre ville mindske risikoen (eksempelvis moderat motion). I praksis antager mange observationsanalyser implicit irrelevans af behandlingsvariation, når der laves kausale inferenser om behandlinger med flere versioner.

Udefinerede behandlinger som "overvægt" gør fortolkningen af estimater for kausale effekter mere komplekse, men det gør tilstrækkelig veldefinerede behandlinger, der ikke optræder i data også. Opdagelse af et misforhold mellem behandlingsværdierne og data kræver en omhyggelig karakterisering af de behandlingsversioner, der fungerer i populationen. En sådan karakterisering kan være simpel i eksperimenter (det vil sige uanset hvilket indgreb, der anvendes til at tildele behandling) og relativt ligetil i nogle observationsanalyser (eksempelvis dem, der studerer effekterne af medicinske behandlinger), men vanskelig eller umulig i mange observationsanalyser, der studerer effekterne af biologiske og sociale faktorer.

Karakterisationen af behandlingsversionerne i data ville være unødvendig, hvis eksperter enes om, at alle versioner har en lignende kausal effekt. Men da eksperter kan tage fejl, er det bedste, der kan gøres, at gøre disse diskussioner og antagelser så gennemsigtige som muligt, så andre direkte kan udfordre argumenterne.

Der er nu blevet set på observationsstudier, og i næste afsnit ses der på, hvordan kausal inferens grafisk kan afbildes, og hvilke egenskaber der kommer fra det.

## 2.4 Grafisk afbildning af kausal inferens

Dette afsnit er baseret på kapitel 6 i [10], hvis ikke andet er angivet.

I dette afsnit introduceres der, hvordan kausal inferens grafisk kan afbildes. Til dette skal der anvendes nogle matematiske begreber, så som strukturelle kausale modeller (SCM), indgreb

og kontrafakta. Disse begreber defineres matematisk, men først introduceres en definition, som anvendes senere.

**Definition 2.4.1** (Pearls  $d$ -separation)

I en orienteret acyklisk graf (DAG)  $\mathcal{G}$  er en sti mellem knuderne  $i_1$  og  $i_m$  blokeret af en mængde  $S$  (der ikke indeholder hverken  $i_1$  eller  $i_m$ ), når der er en knude  $i_k$ , sådan at en af de to følgende muligheder holder:

- $i_k \in S$  og

$$i_{k-1} \rightarrow i_k \rightarrow i_{k+1}$$

eller  $i_{k-1} \leftarrow i_k \leftarrow i_{k+1}$

eller  $i_{k-1} \leftarrow i_k \rightarrow i_{k+1}$ .

- Hverken  $i_k$  eller nogle af dens efterkommere er i  $S$ , og

$$i_{k-1} \rightarrow i_k \leftarrow i_{k+1}.$$

I en DAG  $\mathcal{G}$  siges det, at to disjunkte delmængder af knuder  $A$  og  $B$  er  $d$ -separeret af en tredje disjunkt delmængde  $S$ , hvis enhver sti mellem knuderne i  $A$  og  $B$  er blokeret af  $S$ . Så skrives der

$$A \perp\!\!\!\perp_{\mathcal{G}} B | S.$$

I definitionen betyder ordet *acyklisk*, at hvis der haves en orienteret graf, findes der ikke en cykel i selve grafen.

Hvis  $i_k \rightarrow i_{k+1}$  for alle  $k$ , så siges der at være en orienteret sti fra  $i_1$  til  $i_m$ , og  $i_1$  kaldes en *forfader* ( $AN_i^{\mathcal{G}}$ ) til  $i_m$ , og  $i_m$  kaldes en *efterkommer* ( $DE_i^{\mathcal{G}}$ ) af  $i_1$ .

I næste underafsnit ses der på en speciel type af kausale modeller, som ofte anvendes.

## 2.4.1 Strukturelle kausale modeller

Strukturelle kausale modeller anvendes til at beregne for eksempel indgrebsfordelinger. I hele dette afsnit antages det, at SCM er givet eller som det mindste dens struktur. Nu gives der en definition på SCM.



**Definition 2.4.2** (Strukturel kausal model)

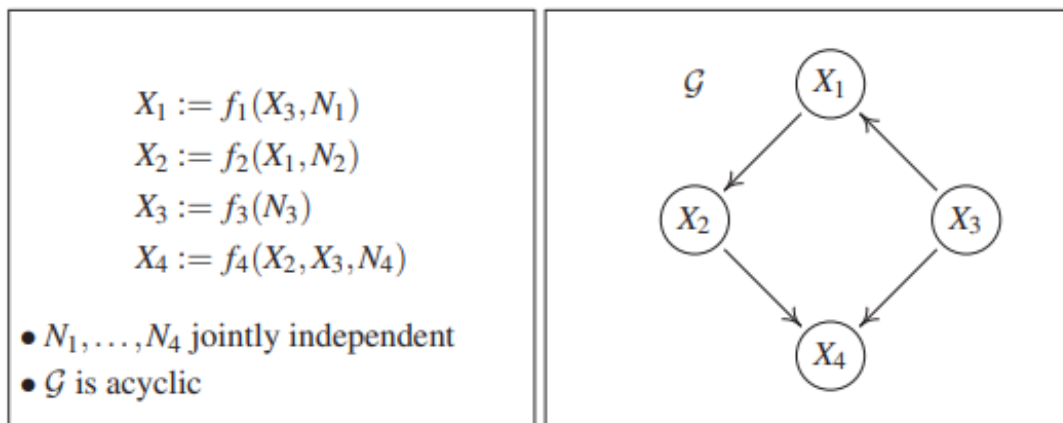
En strukturel kausal model (SCM)  $\mathcal{C} := (\mathbf{S}, P_{\mathbf{N}})$  består af en mængde  $\mathbf{S}$  af  $d$  strukturelle anvisninger

$$X_j := f_j(\mathbf{PA}_j, N_j), \quad j = 1, \dots, d, \quad (2.1)$$

hvor  $\mathbf{PA}_j \subseteq \{X_1, \dots, X_d\} \setminus \{X_j\}$  kaldes *forældre* til  $X_j$ .  $P_{\mathbf{N}} = P_{N_1}, \dots, P_{N_d}$  er en samlet fordeling over støjvariabler, som kræves at være indbyrdes uafhængige, altså er  $P_{\mathbf{N}}$  en produktfordeling.

Grafen  $\mathcal{G}$  af en SCM opnås ved at lave en knude for hver  $X_j$  og lave orienterede kanter fra hver af dens forældre  $\mathbf{PA}_j$  til  $X_j$ , altså fra hver variabel  $X_k$  på højre side af ligning (2.1) til  $X_j$  (se Figur 2.6). Det antages, at denne graf er acyklisk.

En knude  $i$  kaldes en forælder til knude  $j$ , hvis der findes en sti  $i \rightarrow j$ , men ikke en sti  $j \rightarrow i$ . Omvendt hvis der findes en sti  $j \rightarrow i$ , men ikke en sti  $i \rightarrow j$ , kaldes knuden  $i$  for *barnet* af knuden  $j$ . Mængden af knude  $j$ 's børn betegnes  $(\text{CH}_j^{\mathcal{G}})$ .



Figur 2.6: Til venstre er der et eksempel på en SCM, og til højre dens tilhørende graf. Taget fra [10] på side 84.

Nogle gange kaldes elementerne af  $\mathbf{PA}_j$  ikke kun for forældre, men også for *direkte årsager* til  $X_j$ , og  $X_j$  kaldes en *direkte effekt* af hver af dens direkte årsager.

Selvom at terminologien er kausal ("direkte årsag" og "direkte effekt"), så er Definition 2.4.2 udelukkende matematisk. SCM'er er nøglen til at formalisere kausal argumentation og kausal læring. Det vises først, at SCM medfører en observationel fordeling, som er den fordeling, der kan observeres, men til forskel fra de sædvanlige sandsynlighedsmodeller, så medfører den også indgrebsfordelinger, som kommer ind på senere, og kontrafaktum.

**Proposition 2.4.3** (Medførte fordelinger)

En SCM  $\mathcal{C}$  definerer en entydig fordeling over variableerne  $\mathbf{X} = (X_1, \dots, X_d)$ , sådan at  $X_j = f_j(\mathbf{PA}_j, N_j)$ , hvor  $j = 1, \dots, d$ . Den medførte fordeling noteres  $P_{\mathbf{X}}^{\mathcal{C}}$ , og sommetider skrives  $P_{\mathbf{X}}$ .

*Bevis*

Fra Definitionen 2.4.2 er det et krav, at den underliggende graf er acyklisk. Nu kan de strukturelle anvisninger erstattes rekursivt ind i hinanden, og på grund af at grafen er acyklisk, kan enhver knude  $X_j$  skrives som en entydig funktion af alle støjled  $(N_k)_{k \in \mathbf{AN}_j}$ , hvor  $\mathbf{AN}_j$  er alle forfædre af  $X_j$ , altså

$$X_j := g_j((N_k)_{k \in \mathbf{AN}_j}).$$

■

Denne proposition formaliserer proceduren for, hvordan der udtages  $n$  datapunkter fra en samlet fordeling: Først genereres  $n$  i.i.d. støjled  $\mathbf{N}^1, \dots, \mathbf{N}^n \sim P_{\mathbf{N}}$ , og derefter anvendes de strukturelle anvisninger (starter med *kilde knuden* (en knude uden forældre, men med børn), derefter knuder med højst en forældre og så videre) til at generere i.i.d. datapunkter  $\mathbf{X}^1, \dots, \mathbf{X}^n \sim P_{\mathbf{X}}$ . De strukturelle anvisninger (2.1) skal tænkes som en mængde af anvisninger eller funktioner (ikke som matematiske ligninger), der fortæller, hvordan bestemte variabler bestemmer andre.

I næste underafsnit ses der på indgreb, som anvendes ofte i det følgende.

## 2.4.2 Indgreb

Nu kan indgreb modelleres i en SCM. Når der laves indgreb på en variabel  $X_2$ , og sætter det til et binært udfald af en mønt, forventes det, at indgrebet ændrer fordelingen af systemet i forhold til dets tidligere opførsel uden indgreb. Selv hvis variabelen  $X_2$  var kausalt påvirket af andre variabler før, er den nu kun påvirket af mønten: dens kausale forældre er ændret.

Formelt konstrueres indgrebsfordelinger fra en SCM  $\mathcal{C}$ . De opnås ved at lave ændringer for  $\mathcal{C}$  og derefter betragtes den nye medførte fordeling. Generelt er der forskel på indgrebsfordelinger og observationsfordelingen.

**Definition 2.4.4** (Indgrebsfordeling)

Betragt en SCM  $\mathcal{C} := (S, P_N)$  og dens medførte fordeling  $P_{\mathbf{X}}^{\mathcal{C}}$ . En (eller flere) af de strukturelle anvisninger erstattes for at opnå en ny SCM  $\tilde{\mathcal{C}}$ . Antag, at anvisningen for  $X_k$  erstattes af

$$X_k := \tilde{f}(\widetilde{\mathbf{PA}}_k, \tilde{N}_k).$$

Så kaldes den medførte fordeling af den nye SCM en *indgrebsfordeling*, og variablerne, hvis strukturelle anvisninger er erstattet, siges at have været *grebet ind på*. Den nye fordeling noteres

$$P_{\mathbf{X}}^{\tilde{\mathcal{C}}} =: P_{\mathbf{X}}^{\mathcal{C}; do(X_k := \tilde{f}(\widetilde{\mathbf{PA}}_k, \tilde{N}_k))}.$$

Mængden af støjvariabler i  $\tilde{\mathcal{C}}$  indeholder nu både nogle ”nye”  $\tilde{N}$ ’er og nogle ”gamle”  $N$ ’er, som alle kræver at være simultant uafhængige.

Når  $\tilde{f}(\widetilde{\mathbf{PA}}_k, \tilde{N}_k)$  er konstant lig med en reel værdi  $a$  skrives  $P_{\mathbf{X}}^{\mathcal{C}; do(X_k := a)}$ , som kaldes et *atomisk* indgreb. Et indgreb med  $\widetilde{\mathbf{PA}}_k = \mathbf{PA}_k$  kaldes *ikke perfekt*. Dette er et specialtilfælde af et stokastisk indgreb [Korb et al., 2004] [7], hvor den marginale fordeling af indgrebsvariablen har positiv varians.

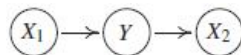
Det kræves, at den nye SCM  $\tilde{\mathcal{C}}$  har en acyklisk graf; mængden af lovlige indgreb afhænger dermed af grafen, der kommer fra  $\mathcal{C}$ .

Det viser sig, at indgrebskonceptet er et godt værktøj til at modellere forskelle i fordelinger og til at forstå kausale forhold. Her er et eksempel.

**Eksempel 2.4.5** (Prædiktorer og indgrebsmål)

Dette eksempel betragter prædiktion. Det viser, at selvom nogle variabler kan være gode prædiktorer for en målvariabel  $Y$ , så kan målvariablen være upåvirket efter et indgreb. Betragt følgende SCM  $\mathcal{C}$ ,

$$\begin{aligned} X_1 &:= N_{X_1} \\ Y &:= X_1 + N_Y \\ X_2 &:= Y + N_{X_2}, \end{aligned}$$



Figur 2.7: *Taget fra [10] på side 90.*

hvor  $N_{X_1}, N_Y \stackrel{iid}{\sim} N(0, 1)$  og  $N_{X_2} \sim N(0, 0.1)$  er simultant uafhængige. Antag, at det ønskes at prædiktere  $Y$  fra  $X_1$  og  $X_2$ . Her er  $X_2$  en bedre prædiktor for  $Y$  end  $X_1$ ; for eksempel, en lineær model uden  $X_2$  fører til en større MSE end en lineær model uden  $X_1$ . Hvis det ønskes at ændre  $Y$ , er indgreb på  $X_2$  ikke brugbare:

$$P_Y^{\mathcal{C}; do(X_2 := \tilde{N})} = P_Y^{\mathcal{C}} \quad \text{for alle variabler } \tilde{N}.$$

Uanset hvordan der laves indgreb på  $X_2$ , så vil fordelingen af  $Y$  forblive upåvirket. Et indgreb på  $X_1$  ændrer derimod fordelingen af  $Y$ :

$$P_Y^{\mathcal{C}; do(X_1 := \tilde{N})} = N(E[N_Y] + E[\tilde{N}], var[N_Y] + var[\tilde{N}]) \neq P_Y^{\mathcal{C}},$$

hvis  $P_{\tilde{N}} \neq P_{N_{X_1}}$ , og hvis  $\tilde{N}$  er normalfordelt og uafhængig af  $N_Y$ . □

Dette eksempel kan også anvendes til at vise, at indgreb ofte er forskellig fra betingelse:

$$p_Y^{\mathcal{C}; do(X_2 := x)}(y) = p_Y^{\mathcal{C}}(y) \neq p_Y^{\mathcal{C}}(y | X_2 = x).$$

#### Definition 2.4.6 (Total kausal effekt)

Givet en SCM  $\mathcal{C}$  er der en *total kausal effekt* fra  $X$  til  $Y$ , hvis og kun hvis

$$X \not\perp\!\!\!\perp Y \quad \text{i } P_{\mathbf{X}}^{\mathcal{C}; do(X := \tilde{N}_X)}$$

for en bestemt tilfældig variabel  $\tilde{N}_X$ .

Der er andre koncepter end den, der er i Definition 2.4.6, som intuitivt beskriver eksistensen af en total kausal effekt. Det viser sig, at de fleste udsagn er ækvivalente.

#### Proposition 2.4.7 (Totale kausale effekter)

Lad en SCM  $\mathcal{C}$  være givet. Så er følgende udsagn ækvivalente:

- (i) Der er en total kausal effekt fra  $X$  til  $Y$ .
- (ii) Der er et  $x'$  og  $x''$  således at  $P_Y^{\mathcal{C}; do(X := x')} \neq P_Y^{\mathcal{C}; do(X := x'')}$ .
- (iii) Der er et  $x'$  således at  $P_Y^{\mathcal{C}; do(X := x')} \neq P_Y^{\mathcal{C}}$ .
- (iv)  $X \not\perp\!\!\!\perp Y$  i  $P_{X,Y}^{\mathcal{C}; do(X := \tilde{N}_X)}$  for ethvert  $\tilde{N}_X$ , hvis fordelingen har fuld support.

I følgende bevis anvendes notation, som introduceres i underafsnit 2.4.5.

#### Bevis

For at forenkle notation skrives  $X_1$  i stedet for  $X$ , og  $X_2$  i stedet for  $Y$ . Ved at anvende følgende trunkeret faktorisering

$$p^{\mathcal{C}; do(X_k := a)}(x_1, \dots, x_d) = \begin{cases} \prod_{j \neq k} p^{\mathcal{C}}(x_j | x_{pa(j)}) & \text{hvis } x_k = a \\ 0 & \text{ellers,} \end{cases}$$

så medfører det, at

$$p_{X_2}^{\mathcal{C};do(X_1:=x_1)}(x_2) = \int \prod_{j \neq 1} p_j(x_j | x_{pa(j)}) dx_3 \cdots dx_d \quad (2.2)$$

$$= \int \prod_{j \neq 1} p_j(x_j | x_{pa(j)}) \frac{\tilde{p}(x_1)}{\tilde{p}(x_1)} dx_3 \cdots dx_d \quad (2.3)$$

$$= p_{X_2|X_1=x_1}^{\mathcal{C};do(X_1:=\tilde{N}_1)}(x_2), \quad (2.4)$$

hvis  $\tilde{N}_1$  sætter positiv masse på  $x_1$ , det vil sige,  $\tilde{p}(x_1) > 0$ . Det kræves yderligere, at følgende to udsagn holder for alle fordelinger  $Q_{X_1, X_2}$  over  $(X_1, X_2)$  med tæthed  $q$ :

$$X_2 \not\perp\!\!\!\perp X_1 \text{ i } Q \iff \exists x'_1, x''_1 \text{ med } q(x'_1), q(x''_1) > 0 \text{ og } Q_{X_2|X_1=x'_1} \neq Q_{X_2|X_1=x''_1}, \quad (2.5)$$

og

$$X_2 \not\perp\!\!\!\perp X_1 \text{ i } Q \iff \exists x'_1 \text{ med } q(x'_1) > 0 \text{ og } Q_{X_2|X_1=x'_1} \neq Q_{X_2}. \quad (2.6)$$

Herefter haves for ethvert  $\tilde{N}_1$  med fuld support, at

$$(i) \xrightarrow{\text{ligning(2.5)}} \exists x'_1, x''_1 \text{ med positiv tæthed under } \tilde{N}_1 \text{ således at } P_{X_2|X_1=x'_1}^{\mathcal{C};do(X_1:=\tilde{N}_1)} \neq P_{X_2|X_1=x''_1}^{\mathcal{C};do(X_1:=\tilde{N}_1)}$$

$$\xrightarrow{\text{ligning(2.2)}} (ii)$$

$$\xrightarrow{\text{ligning(2.2)}} \exists x'_1, x''_1 \text{ med positiv tæthed under } \hat{N}_1 \text{ således at } P_{X_2|X_1=x'_1}^{\mathcal{C};do(X_1:=\hat{N}_1)} \neq P_{X_2|X_1=x''_1}^{\mathcal{C};do(X_1:=\hat{N}_1)}$$

$$\xrightarrow{\text{ligning(2.5)}} (iv)$$

$$\xrightarrow{\text{(trivielt)}} (i).$$

Yderligere haves der  $(ii) \xrightarrow{\text{(trivielt)}} (iii)$ , og at  $P_{X_2}^{\mathcal{C}} = P_{X_2}^{\mathcal{C};do(X_1:=N_1^*)}$ , hvor  $N_1^*$  har fordelingen  $P_{X_1}^{\mathcal{C}}$ .

Med  $\neg(i) \implies \neg(ii)$ , medfører det sidstnævnte, at

$$\neg(i) \implies X_2 \perp\!\!\!\perp X_1 \text{ i } P_{\mathbf{X}}^{\mathcal{C};do(X_1:=N_1^*)}$$

$$\xrightarrow{\text{ligning(2.6)}} P_{X_2|X_1=x'_1}^{\mathcal{C};do(X_1:=N_1^*)} = P_{X_2}^{\mathcal{C};do(X_1:=N_1^*)} \text{ for alle } x' \text{ med } p_1(x') > 0$$

$$\xrightarrow{\text{ligning(2.2)}} P_{X_2}^{\mathcal{C};do(X_1:=x')} = P_{X_2}^{\mathcal{C}} \text{ for alle } x' \text{ med } p_1(x') > 0$$

$$\xrightarrow{\neg(ii)} P_{X_2}^{\mathcal{C};do(X_1:=x')} = P_{X_2}^{\mathcal{C}} \text{ for alle } x'$$

$$\implies \neg(iii).$$

■

Eksistensen af en total kausal effekt er relateret til eksistensen af en orienteret sti i den tilsvarende graf, men dette er dog ikke injektiv. Selvom en orienteret sti er nødvendig for en total kausal effekt, er det ikke tilstrækkeligt.

### Proposition 2.4.8

Antag, at en SCM  $\mathcal{C}$  med tilsvarende graf  $G$  er givet.

- (i) Hvis ikke der er en orienteret sti fra  $X$  til  $Y$ , så er der ingen total kausal effekt.
- (ii) Nogle gange er der en orienteret sti, men ingen total kausal effekt.

*Bevis*

Udsagn (i) følger direkte fra Markov egenskaben for SCM, hvilket behandles i underafsnit 2.4.3. Indgrebet fjerner de indgående kanter i  $X$ , og hvis ikke der er en direkte sti fra  $X$  til  $Y$  i den originale graf, er  $X$  og  $Y$   $d$ -separatoreret. Udsagn (ii) kan bevises med et modeksempel (se Eksempel 6.34 i [10] på side 107). ■

Nu introduceres kontrafakta i SCM termer. Formelt handler det om, at støjfordelingen i en SCM opdateres, og derefter laves et indgreb.

### Definition 2.4.9 (Kontrafakta)

Betragt en SCM  $\mathcal{C} := (\mathbf{S}, P_{\mathbf{N}})$  over knuder  $\mathbf{X}$ . Givet nogle observationer  $\mathbf{x}$ , defineres en kontrafaktisk SCM ved at erstatte fordelingen af støjvariablerne:

$$\mathcal{C}_{\mathbf{X}=\mathbf{x}} := \left( \mathbf{S}, P_{\mathbf{N}}^{\mathcal{C}|\mathbf{X}=\mathbf{x}} \right),$$

hvor  $P_{\mathbf{N}}^{\mathcal{C}|\mathbf{X}=\mathbf{x}} := P_{\mathbf{N}|\mathbf{X}=\mathbf{x}}$ .

Den nye mængde af støjvariabler behøver ikke længere at være indbyrdes uafhængige. Kontrafaktiske udsagn kan nu ses som do-udsagn i den nye kontrafaktiske SCM.

Denne definition kan generaliseres, sådan at der ikke observeres den fulde vektor  $\mathbf{X} = \mathbf{x}$ , men kun nogle af variablerne.

I næste underafsnit ses der på, hvordan Markov egenskaben benyttes i SCM.

## 2.4.3 Markov egenskab

Markov egenskaben er en normalt anvendt antagelse, der giver basen for grafiske modeller. Når en fordeling er Markov med hensyn til en graf, så indkoder denne graf visse uafhængigheder i fordelingen, som kan udnyttes til effektiv beregning eller lagring af data. Markov egenskaben eksisterer for både orienterede og ikke-orienterede grafer, og disse to klasser indkoder forskellige mængder af uafhængigheder. I kausal inferens er der kun interesse for orienterede grafer. Herunder gives en definition på Markov egenskaben.

**Definition 2.4.10** (Markov egenskab)

Givet en DAG  $\mathcal{G}$  og en fælles fordeling  $P_{\mathbf{X}}$ , så siges denne fordeling at overholde

- den globale Markov egenskab med hensyn til DAG  $\mathcal{G}$ , hvis

$$\mathbf{A} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{B} | \mathbf{C} \Rightarrow \mathbf{A} \perp\!\!\!\perp \mathbf{B} | \mathbf{C},$$

for alle disjunkte mængder  $\mathbf{A}, \mathbf{B}, \mathbf{C}$ .

- den lokale Markov egenskab med hensyn til DAG  $\mathcal{G}$ , hvis hver variabel er uafhængig af dens ikke-efterkommere givet dens forældre.
- Markov faktoriseringssegenskaben med hensyn til DAG  $\mathcal{G}$ , hvis

$$p(\mathbf{x}) = p(x_1, \dots, x_d) = \prod_{j=1}^d p_j(x_j | \mathbf{pa}_j^{\mathcal{G}}).$$

For den sidste egenskab skal der antages, at  $P_{\mathbf{X}}$  har tæthed  $p$ , og faktorerne i produktet kaldes *kausale Markov kerner*, som beskriver den betingede fordeling  $P_{X_j | \mathbf{PA}_j^{\mathcal{G}}}$ .

Det viser sig, at så længe den fælles fordeling har en tæthed, er disse tre egenskaber ækvivalente.

**Sætning 2.4.11** (Ækvivalens af Markov egenskaber)

Hvis  $P_{\mathbf{X}}$  har en tæthed  $p$ , så er alle Markov egenskaber i Definition 2.4.10 ækvivalente.

*Bevis*

Dette bevis kan findes som Sætning 3.27 i Lauritzen (1996) [8]. ■

Markov betingelsen relaterer udsagn om graf-separation til betinget uafhængighed. Det er muligt, at forskellige grafer indkoder nøjagtigt den samme mængde af betingede uafhængigheder.

**Definition 2.4.12** (Markov ækvivalens af grafer)

$\mathcal{M}(\mathcal{G})$  betegner mængden af fordelinger, der er Markov med hensyn til  $\mathcal{G}$ :

$$\mathcal{M}(\mathcal{G}) := \{P : P \text{ overholder den globale (eller lokale) Markov egenskab mht. } \mathcal{G}\}.$$

Two DAG'er  $\mathcal{G}_1$  og  $\mathcal{G}_2$  er Markov ækvivalente, hvis  $\mathcal{M}(\mathcal{G}_1) = \mathcal{M}(\mathcal{G}_2)$ . Dette sker, hvis og kun hvis  $\mathcal{G}_1$  og  $\mathcal{G}_2$  har den samme mængde  $d$ -separationer, hvilket betyder, at Markov betingelsen medfører den samme mængde (betingede) uafhængighedsbetingelser.

Bemærk, at mængden af alle DAG'er, der er Markov ækvivalente til en DAG, kaldes *Markov ækvivalensklassen* af  $\mathcal{G}$ . Den kan repræsenteres af en fuld PDAG (partially DAG) som betegnes med  $CPDAG(\mathcal{G}) = (V, \mathcal{E})$ ; den indeholder (den orienterede) kant  $(i, j) \in \mathcal{E}$ , hvis og kun hvis

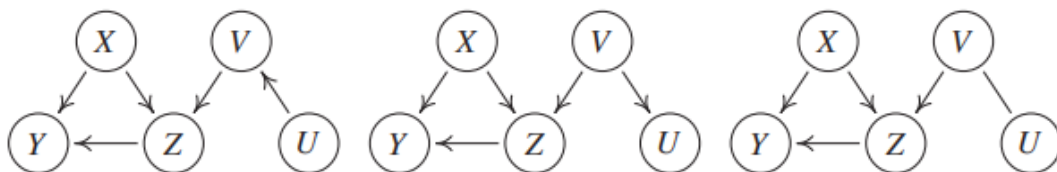
en af medlemmerne af Markov ækvivalensklassen indeholder den. Se Figur 2.8.

Fra denne definition kan det godt se ud som om, at det ikke er trivielt at finde ud af, om to DAG'er er Markov ækvivalente, men Verma and Pearl (1991) [20] giver en præcis karakterisation.

Bemærk, at tre knuder  $A$ ,  $B$  og  $C$  i en DAG danner en  $v$ -struktur, hvis  $A \rightarrow B \leftarrow C$ , og  $A$  og  $C$  er ikke direkte forbundet.

**Lemma 2.4.13** (Grafiske kriterier for Markov ækvivalens)

To DAG'er  $\mathcal{G}_1$  og  $\mathcal{G}_2$  er Markov ækvivalente, hvis og kun hvis de har de samme  $v$ -strukturer og det samme skelet (hvilket betyder, at de er ens, når der ikke er nogen orienterede kanter).



Figur 2.8: Taget fra [10] side 103.

**Eksempel 2.4.14**

Figur 2.8 viser et eksempel på to Markov ækvivalente grafer; grafen til venstre og grafen i midten. Graferne har det samme skelet og kun en  $v$ -struktur  $X \rightarrow Z \leftarrow V$ . I den tilhørende CPDAG til højre i figuren er alle orienterede kanter ikke en del af en  $v$ -struktur. For eksempel er kanten  $Z \rightarrow Y$  nødvendig for at undgå en  $v$ -struktur  $Y \rightarrow Z \leftarrow V$ . Derudover forhindrer kanten  $X \rightarrow Y$ , at der eksisterer en orienteret cykel.  $\square$

Der introduceres nu et grafisk koncept kaldet et Markov blanket fra Pearl (1988) [11], der bliver relevant, når der forsøges at forudsige en målvariabel  $Y$  fra de observerede værdier af alle andre variabler. Der kan spørges, hvad er den mindste mængde af variabler, der gør alle andre variabler irrelevante?

**Definition 2.4.15** (Markov blanket)

Betragt en DAG  $\mathcal{G} = (\mathbf{V}, \mathcal{E})$  og en målknode  $Y$ . *Markov blanket* af  $Y$  er den mindste mængde  $M$ , sådan at

$$Y \perp\!\!\!\perp_{\mathcal{G}} \mathbf{V} \setminus (\{Y\} \cup M) \text{ givet } M.$$



Bemærk, at hvis  $P_X$  er Markov med hensyn til  $\mathcal{G}$ , så er

$$Y \perp\!\!\!\perp \mathbf{V} \setminus (\{Y\} \cup M) \text{ givet } M.$$

Med andre ord, hvis  $M$  er givet, så giver de andre variabler ikke ny information om  $Y$ . I en ideel regression er der kun brug for variablerne i  $M$  for at forudsige  $Y$ . Dette medfører ikke, at de andre variabler er ubrugelige i en endelig prøve. Hvis der er noget underliggende, der ikke vides om fordelingen af  $Y$ , kan det være, at de indsamlede variabler ikke er de helt rigtige til at prædiktere  $Y$ . Dette betyder, at Markov blanket ikke er den bedste mængde variabler at bruge, så der skal bruges andre variabler udenfor den.

For DAG'er vides det, hvordan Markov blanket ser ud. Det indeholder ikke kun forældre, men også børn og forældre til børn [Pearl (1988)] [11].

**Proposition 2.4.16** (Markov blanket)

Betragt en DAG  $\mathcal{G}$  og en målknode  $Y$ . Så inkluderer Markov blanket af  $Y$  dens forældre, dens børn og forældrene til dens børn

$$M = \text{PA}_Y \cup \text{CH}_Y \cup \text{PA}_{\text{CH}_Y}.$$

Indtil nu er der blevet diskuteret omkring Markov egenskaben relateret til fordeling og grafer. Nu kigges der på nogle af dens kausale implikationer. Markov egenskaben kan anvendes til at retfærdiggøre *Reichenbachs common cause principle* (princip 1.1), som siger, at når de stokastiske variabler  $X$  og  $Y$  er afhængige, så må der være en "kausal forklaring" for denne afhængighed:

1.  $X$  forårsager (muligvis indirekte)  $Y$ , eller
2.  $Y$  forårsager (muligvis indirekte)  $X$ , eller
3. der er (muligvis ikke observeret) en fælles årsag  $Z$ , som (muligvis indirekte) forårsager både  $X$  og  $Y$ .

Her er det ikke specificeret, hvad der menes med "forårsage". I den følgende proposition retfærdiggøres Reichenbachs princip med hensyn til et svagt begreb af "forårsage", altså eksistensen af en orienteret kant.

**Proposition 2.4.17** (Reichenbachs common cause principle)

Antag, at ethvert par af variabler  $X$  og  $Y$  kan indlejres i et større system på følgende måde. Der eksisterer en korrekt SCM over samlingen  $\mathbf{X}$  af tilfældige variabler, der indeholder  $X$  og  $Y$  med grafen  $\mathcal{G}$ . Så følger Reichenbach common cause principle fra Markov egenskaben, hvis  $X$  og  $Y$  er (ubetinget) afhængige, så er der

1. enten en orienteret kant fra  $X$  til  $Y$ , eller
2. fra  $Y$  til  $X$ , eller
3. der er en knude  $Z$  med en orienteret kant fra  $Z$  til  $X$  og fra  $Z$  til  $Y$ .

*Bevis*

På grund af Markov egenskaben medfører afhængigheden, at  $\mathcal{G}$  indeholder en ikke-blokeret sti mellem  $X$  og  $Y$ , da den globale Markov egenskab giver, at  $X$  og  $Y$  ikke er disjunkte mængder. Denne sti kan ikke indeholde en  $v$ -struktur, for ellers vil den være blokeret af den tomme mængde. Propositionen følger da for enhver sti mellem  $X$  og  $Y$  uden en  $v$ -struktur må have formen  $X \rightarrow \dots \rightarrow Y$ ,  $X \leftarrow \dots \leftarrow Y$  eller  $X \leftarrow \dots \leftarrow Z \rightarrow \dots \rightarrow Y$ . ■

Pearl (2009) [13] viser i Sætning 1.4.1, at fordelingen  $P_{\mathbf{X}}$ , som kommer fra en SCM, er Markov med hensyn til dens graf.

**Proposition 2.4.18** (SCM medfører Markov egenskab)

Antag, at  $P_{\mathbf{X}}$  kommer fra en SCM med graf  $\mathcal{G}$ . Så er  $P_{\mathbf{X}}$  Markov med hensyn til  $\mathcal{G}$ .

Antagelsen, at fordelingen er Markov med hensyn til den kausale graf, kaldes sommetider den kausale Markov betingelse. Dette har brug for begrebet om en kausal graf, hvilket der ses på i det følgende underafsnit.

## 2.4.4 Kausale grafiske modeller

Det ses senere, at når indgrebsfordelinger skal defineres, kræver det sædvanligt kun kendskab til observationsfordelingen og grafstrukturen. Derfor defineres en kausal grafisk model som et par, der indeholder en graf og en observationsfordeling, sådan at fordelingen er Markov med hensyn til grafen, men der er en diskret teknikalitet. Formelt er det nødvendigt at kende de fulde betingelser, hvis  $p(x_2|x_1 = 3)$  ikke er defineret, for eksempel på grund af, at  $p(x_1 = 3) = 0$ , så kan  $p^{do(X_1:=3)}(x_2)$  muligvis ikke defineres. Dette motiverer til at definere følgende begreb.

**Definition 2.4.19** (Kausal grafisk model)

En kausal grafisk model over stokastiske variable  $\mathbf{X} = (X_1, \dots, X_d)$  indeholder en graf  $\mathcal{G}$  og en mængde af funktioner  $f_j(x_j, x_{\mathbf{PA}_j^{\mathcal{G}}})$ , der integrerer til 1;

$$\int f_j(x_j, x_{\mathbf{PA}_j^{\mathcal{G}}}) dx_j = 1.$$

Disse funktioner inducerer en fordeling  $P_{\mathbf{X}}$  over  $\mathbf{X}$  via

$$p(x_1, \dots, x_d) = \prod_{j=1}^d f_j(x_j, x_{\mathbf{PA}_j^{\mathcal{G}}}),$$

og derfor spiller de en rolle i betingelserne:  $f_j(x_j, x_{\mathbf{PA}_j^{\mathcal{G}}}) = p(x_j | x_{\mathbf{PA}_j^{\mathcal{G}}})$ .

Hvis en fordeling  $P_{\mathbf{X}}$  over  $\mathbf{X}$  er Markov med hensyn til en graf  $\mathcal{G}$  og tillader en strengt positiv kontinuert tæthed  $p$ , så definerer parret  $(P_{\mathbf{X}}, \mathcal{G})$  en kausal grafisk model med  $f_j(x_j, x_{\mathbf{PA}_j^{\mathcal{G}}}) = p(x_j | x_{\mathbf{PA}_j^{\mathcal{G}}})$ .

**2.4.5 Beregning af indgrebsfordelinger ved ændring af kovariat**

I dette underafsnit gøres der brug af et noget trivielt, men meget kraftfuldt invariansudsagn. Givet en SCM  $\mathcal{C}$ , og  $pa(j) := \mathbf{PA}_j^{\mathcal{C}}$ , haves

$$p^{\tilde{\mathcal{C}}}(x_j | x_{pa(j)}) = p^{\mathcal{C}}(x_j | x_{pa(j)}), \quad (2.7)$$

for alle SCM  $\tilde{\mathcal{C}}$ , som er konstrueret fra  $\mathcal{C}$  ved indgreb på et  $X_k$ , men ikke på  $X_j$ . Ligning (2.7) viser, at kausale forhold er autonome under indgreb; denne egenskab kaldes nogle gange "autonomi". Hvis der indgribes på en variabel, forbliver de andre mekanismer invariante.

Der udledes en formel fra ligning (2.7), som er kendt under tre forskellige navne: trunkeket faktoriseringsformel [Pearl, 1993] [12], G-computation formel [Robins, 1986] [16], og manipulations-sætning [Spirtes et al., 2000] [3]. Vigtigheden kommer fra, at det tillader at bestemme udsagn om indgrebsfordelinger, selvom der aldrig er blevet set på data.

Betragt en SCM  $\mathcal{C}$  med

$$X_j := f_j(X_{pa(j)}, N_j), \quad j = 1, \dots, d,$$

og tæthed  $p^{\mathcal{C}}$ . På grund af Markov egenskaben haves

$$p^{\mathcal{C}}(x_1, \dots, x_d) = \prod_{j=1}^d p^{\mathcal{C}}(x_j | x_{pa(j)}).$$

Bemærk, at de betingede  $p^{\mathcal{C}}(x_j|x_{pa(j)})$  kan defineres selv for værdier  $x_{pa(j)}$  således at  $p^{\mathcal{C}}(x_{pa(j)}) = 0$ . Betragt nu SCM  $\tilde{\mathcal{C}}$ , som udvikles fra  $\mathcal{C}$  efter  $do(X_k := \tilde{N}_k)$ , hvor  $\tilde{N}_k$  giver tætheden  $\tilde{p}$ . Igen følger det fra Markovantagelsen i indgrebsfordelingen, at

$$p^{\mathcal{C};do(X_k:=\tilde{N}_k)}(x_1, \dots, x_d) = \prod_{j \neq k} p^{\mathcal{C};do(X_k:=\tilde{N}_k)}(x_j|x_{pa(j)}) \cdot p^{\mathcal{C};do(X_k:=\tilde{N}_k)}(x_k) \quad (2.8)$$

$$= \prod_{j \neq k} p^{\mathcal{C}}(x_j|x_{pa(j)}) \tilde{p}(x_k). \quad (2.9)$$

I sidste lighed anvendes invariansen fra ligning (2.7). Ligning (2.8) tillader at bestemme et indgribende udsagn (venstre side) fra observationsstørrelser (højre side). Som et specialtilfælde haves

$$p^{\mathcal{C};do(X_k:=a)}(x_1, \dots, x_d) = \begin{cases} \prod_{j \neq k} p^{\mathcal{C}}(x_j|x_{pa(j)}) & \text{hvis } x_k = a, \\ 0 & \text{ellers.} \end{cases} \quad (2.10)$$

Ofte er betingelser og indgreb med  $do()$  forskellige operationer. Det vises nu, at disse operationer bliver identiske for variabler, som ingen forældre har. Antag uden tab af generalitet, at  $X_1$  er sådan en kilde knude. Så haves

$$p^{\mathcal{C}}(x_2, \dots, x_d|x_1 = a) = \frac{p(x_1 = a) \prod_{j=2}^d p^{\mathcal{C}}(x_j|x_{pa(j)})}{p(x_1 = a)} \quad (2.11)$$

$$= p^{\mathcal{C};do(X_1:=a)}(x_2, \dots, x_d). \quad (2.12)$$

Ligning (2.8) og (2.10) er bredt anvendelige, men nogle gange lidt besværligt at anvende.

**Definition 2.4.20** (Gyldig justeringsmængde)

Betragt en SCM  $\mathcal{C}$  over knuder  $V$  og lad  $Y \notin \mathbf{PA}_X$  (ellers haves  $p^{\mathcal{C};do(X:=x)}(y) = p^{\mathcal{C}}(y)$ ). En mængde  $Z \subseteq V \setminus \{X, Y\}$  kaldes en *gyldig justeringsmængde* for det ordnede par  $(X, Y)$ , hvis

$$p^{\mathcal{C};do(X:=x)}(y) = \sum_z p^{\mathcal{C}}(y|x, z) p^{\mathcal{C}}(z).$$

Her er summen (hvilket også kunne være et integral) over alle værdier  $z$ , som  $Z$  kan antage.

Nogle gange kan simple betingelser lede til falske konklusioner. Dette leder frem til følgende definition.

**Definition 2.4.21** (Confounding)

Betragt en SCM  $\mathcal{C}$  over knuder  $V$  med en orienteret sti fra  $X$  til  $Y$ , hvor  $X, Y \in V$ . Den kausale effekt fra  $X$  til  $Y$  kaldes *confounded*, hvis

$$p^{\mathcal{C};do(X:=x)}(y) \neq p^{\mathcal{C}}(y|x).$$

Det kunne nogle gange tænkes at gøre justeringsmængden så stor som muligt for at reducere indflydelsen af potentielle confounders. Det viser sig, at dette ikke altid er en god idé (Berkson's paradox [Berkson, 1946] [1]). Ikke alle mængder er gyldige justeringsmængder, og nogle gange er det bedre ikke at inkludere en kovariat i justeringsmængden. Det undersøges nu hvilke mængder, som kan anvendes for justering. For en mængde  $\mathbf{Z}$  haves

$$\begin{aligned} p^{\mathcal{C};do(X:=x)}(y) &= \sum_{\mathbf{Z}} p^{\mathcal{C};do(X:=x)}(y, \mathbf{z}) \\ &= \sum_{\mathbf{Z}} p^{\mathcal{C};do(X:=x)}(y|x, \mathbf{z}) p^{\mathcal{C};do(X:=x)}(\mathbf{z}). \end{aligned}$$

Hvis der haves

$$p^{\mathcal{C};do(X:=x)}(y|x, \mathbf{z}) = p^{\mathcal{C}}(y|x, \mathbf{z}) \quad \text{og} \quad p^{\mathcal{C};do(X:=x)}(\mathbf{z}) = p^{\mathcal{C}}(\mathbf{z}), \quad (2.13)$$

følger det som før, at  $\mathbf{Z}$  er en gyldig justeringsmængde. Egenskaben fra ligning (2.13) betyder, at betingelserne forbliver de samme selv efter indgreb på  $X$ ; det siges, at de er invariante. Det næste spørgsmål er, hvilke betingelser forbliver invariante under indgribningen  $do(X := x)$ . Til dette ses der på følgende eksempel.

#### Eksempel 2.4.22 (Karakterisation af invariante betingelser)

Betragt en SCM  $\mathcal{C}$  med

$$X_j := f_j(\mathbf{P}\mathbf{A}_j, N_j)$$

og et indgreb  $do(X_k := x_k)$ . Analogt til hvad der gøres i Pearl [2009, Chapter 3.3.2] [13] for eksempel, kan der nu konstrueres en ny SCM  $\mathcal{C}^*$ , som er lig med  $\mathcal{C}$ , men med endnu en variabel  $I$ , som indikerer, om indgrebet fandt sted eller ej. Mere præcist er  $I$  en forælder til  $X_k$  og har ingen andre naboer. De tilhørende strukturelle anvisninger er

$$\begin{aligned} I &:= N_I \\ X_j &:= f_j(\mathbf{P}\mathbf{A}_j, N_j) \quad \text{for } j \neq k \\ X_k &:= \begin{cases} f_k(\mathbf{P}\mathbf{A}_k, N_k) & \text{hvis } I = 0 \\ x_k & \text{ellers,} \end{cases} \end{aligned}$$

hvor  $N_I$  eksempelvis har en Bernoulli fordeling med  $P(I = 0) = P(I = 1) = 0.5$ . Det vil sige, at  $I = 0$  svarer til observationsmiljøet, og  $I = 1$  svarer til indgrebsmiljøet. Ved at anvende ligning (2.11) fås

$$\begin{aligned} p^{\mathcal{C}^*}(x_1, \dots, x_d | I = 0) &= p^{\mathcal{C}^*;do(I:=0)}(x_1, \dots, x_d) \\ &= p^{\mathcal{C}}(x_1, \dots, x_d), \end{aligned}$$

og tilsvarende

$$p^{C^*}(x_1, \dots, x_d | I = 1) = p^{C; do(X_k := x_k)}(x_1, \dots, x_d).$$

Ved at anvende Markov egenskaben for  $C^*$  følger det for variable  $A$  og en mængde af variable  $B$ , at

$$\begin{aligned} A \perp\!\!\!\perp_{\mathcal{G}^*} I | B &\implies p^{C^*}(a | \mathbf{b}, I = 0) = p^{C^*}(a | \mathbf{b}, I = 1) \\ &\implies p^C(a | \mathbf{b}) = p^{C; do(X_k := x_k)}(a | \mathbf{b}). \end{aligned}$$

Højre side siger, at fordelingen  $P_{A|B}$  af den betingede  $A$  givet  $B$  forbliver invariant under et indgreb på  $X_k$ .

□

Ligning (2.13) er opfyldt for mængder  $Z$ , hvor der haves

$$Y \perp\!\!\!\perp_{\mathcal{G}^*} I | X, Z \quad \text{og} \quad Z \perp\!\!\!\perp_{\mathcal{G}^*} I.$$

Notationen  $\mathcal{G}^*$  betyder, at  $d$ -separationsudsagnet skal holde i  $\mathcal{G}^*$ . Dette leder frem til følgende proposition.

**Proposition 2.4.23** (Gyldige justeringsmængder)

Betragt en SCM over variable  $X$  med  $X, Y \in X$  og  $Y \notin PA_X$ . Så er følgende tre udsagn sande.

(i) **”Forælder justering”**:

$$Z := PA_X$$

er en gyldig justeringsmængde for  $(X, Y)$ .

(ii) **”Bagdørskriterie”**: Enhver  $Z \subseteq X \setminus \{X, Y\}$  med

(a)  $Z$  indeholder ingen efterkommere af  $X$ , og

(b)  $Z$  blokerer alle stier fra  $X$  til  $Y$ , der kommer ind i  $X$  gennem bagdøren ( $X \leftarrow \dots$ , se Figur 2.9)

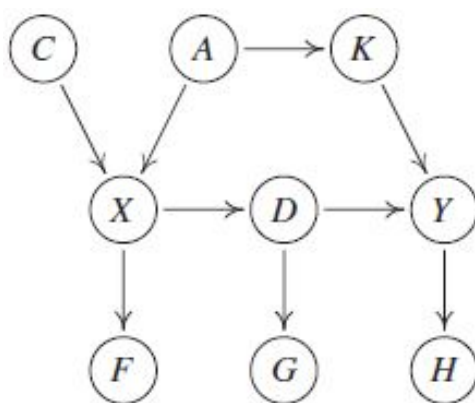
er en gyldig justeringsmængde for  $(X, Y)$ .

(iii) **”Hen imod nødvendighed”**: Enhver  $Z \subseteq X \setminus \{X, Y\}$  med

(a)  $Z$  indeholder ingen efterkommere af nogen knude på en orienteret sti fra  $X$  til  $Y$  (bortset fra efterkommere af  $X$ , som ikke er på en orienteret sti fra  $X$  til  $Y$ ), og

(b)  $Z$  blokerer alle ikke-orienterede stier fra  $X$  til  $Y$

er en gyldig justeringsmængde for  $(X, Y)$ .



Figur 2.9: Taget fra [10] på side 115.

Betragt Figur 2.9. Kun stien  $X \leftarrow A \rightarrow K \rightarrow Y$  er en ”bagdørsvej” fra  $X$  til  $Y$ . Mængden  $Z = \{K\}$  opfylder bagdørskriteriet (se Proposition 2.4.23 (ii)); men  $Z = \{F, C, K\}$  er også en gyldig justeringsmængde for  $(X, Y)$ ; se Proposition 2.4.23 (iii).

Tredje udsagn uddybes. Betragt eksempelvis en gyldig justeringsmængde  $Z$ , som er opnået via bagdørskriteriet. En knude  $Z_0$  kan tilføjes til  $Z$ , som opfylder  $Z_0 \perp\!\!\!\perp Y|X, Z$ , og så gælder

$$\begin{aligned} \sum_{Z, z_0} p(y|x, z, z_0)p(z, z_0) &= \sum_Z p(y|x, z) \sum_{z_0} p(z, z_0) \\ &= \sum_Z p(y|x, z)p(z). \end{aligned}$$

Det viser sig, at (iii) fra Proposition 2.4.23 karakteriserer alle gyldige justeringsmængder [Shpitser et al., 2010] [19].

## 2.4.6 Potentielle udfald

I dette underafsnit introduceres en alternativ tilgang til kausal inferens, som ikke er baseret på SCM’er. Dette kaldes potentielle udfald, hvilket svarer til kontrafakta, som blev anvendt i de tre første afsnit i dette kapitel.

### Definitioner og eksempler

Betragt et eksempel om en øjenlæge. I stedet for tilfældige variable har vi nu en gruppe af  $n$  patienter (eller enheder)  $u = 1, \dots, n$ , hvor hver patient enten får behandling eller ej. Der tildeles to *potentielle udfald* til hver patient  $u$ :  $B_u(t = 1)$  betegner, om patienten ville blive blind ( $B = 1$ ) eller blive helbredt ( $B = 0$ ), hvis behandling modtages ( $T = 1$ ). Analogt betegner  $B_u(t = 0)$ , hvad der ville ske uden behandling ( $T = 0$ ). Begge af disse potentielle udfald antages at være deterministiske. For hver patient hjælper behandlingen, eller også gør den ikke: der

er ingen tilfældighed involveret. Hvis  $B_u(t = 1) = 0$  og  $B_u(t = 0) = 1$  siges det, at behandlingen har en positiv effekt for enhed  $u$ .

I praksis er det ikke muligt at tjekke disse betingelser. ”The fundamental problem of causal inference” [Holland, 1986] [6] siger, at for hver enhed  $u$  kan enten  $B_u(t = 1)$  eller  $B_u(t = 0)$  observeres, men aldrig begge på samme tid. Dette er på grund af, at efter en person er behandlet, er det ikke muligt at gå tilbage i tiden og fortryde det. Dette gælder også den anden vej. Hvis det besluttet ikke at give en behandling, kan behandlingen stadig gives på et senere tidspunkt, men dette kan ikke længere fortolkes som et udfald af variabelen  $B_u(t = 1)$ . Patienten kan for eksempel muligvis være helbredt i mellemtiden af sig selv. Derfor kan kun en af de potentielle udfald observeres, og de uobserverede bliver kontrafakta.

Figur 2.10 viser et (hypotetisk) datasæt for forrige eksempel om øjenlægen. Ofte antages *the stable unit treatment value assumption* (SUTVA) [Rubin, 2005] [18] implicit, hvilket siger, at enhederne ikke forstyrrer hinanden (det potentielle udfald af en enhed afhænger ikke af hvilken behandling enhver anden enhed modtog, eksempelvis) [Cox, 1958] [2]; yderligere kræves det, at de potentielle udfald ikke afhænger af hvordan eller hvorfor behandlingen er modtaget. I næste underundersnit ses det, at SUTVA er opfyldt, når data er genereret fra en SCM.

Unit $u$	Treatment $T$	Pot. Outcome $B_u(t = 0)$	Pot. Outcome $B_u(t = 1)$	Unit-Level Causal Effect $B_u(t = 1) - B_u(t = 0)$
1	1	1	0	-1
2	0	1	0	-1
3	1	1	0	-1
⋮				
43	1	1	0	-1
44	0	0	1	1
45	0	1	0	-1
⋮				
119	1	1	0	-1
120	1	0	1	1
121	0	1	0	-1
⋮				
200	0	1	0	-1

Figur 2.10: Taget fra [10] på side 124.

Tabellen i Figur 2.10 viser det nævnte eksempel med øjenlægen ved anvendelse af potentielle udfald. For hver patient (eller enhed), observeres kun en af de to potentielle udfald. Den observerede information har en grå baggrund. Behandlingen  $T$  er nyttig for næsten alle patienter. Kun i 2 ud af 200 tilfælde skader behandlingen patienten og gør individet blind,  $B = 1$ . Selvom at



tildele behandlingen ( $T = 1$ ) er en god idé i de fleste tilfælde, så var det den forkerte beslutning for patient  $u = 120$ .

De potentielle udfald viser effekten af en behandling på et individ. Den *enheds kausale effekt* defineres som  $B_u(t = 1) - B_u(t = 0)$  og en gennemsnitlig kausal effekt

$$CE = \frac{1}{n} \sum_{u=1}^n (B_u(t = 1) - B_u(t = 0)). \quad (2.14)$$

”The fundamental problem of causal inference” forhindrer at udregne ligning (2.14) direkte. Antag, at i et fuldstændigt randomiseret eksperiment modtog enheder  $u \in U_0 \subset \{1, \dots, n\}$  behandling  $T = 0$  og enheder  $u \in U_1 = U_0^C$  behandling  $T = 1$ . Neyman [1923] [9] viser, at

$$\widehat{CE} := \frac{1}{\#U_0} \sum_{u \in U_0} B_u(t = 1) - \frac{1}{\#U_1} \sum_{u \in U_1} B_u(t = 0) \quad (2.15)$$

er en unbiased estimator for ligning (2.14). Tilfældigheden i  $\widehat{CE}$  kommer fra de tilfældige tildelinger, som afgør, hvilken af de to potentielle udfald for enheden, der observeres; udfaldene selv anses som skjulte, ikke tilfældige. Bemærk, at ligning (2.15) kun indeholder observerede størrelser og kan derfor udregnes efter studiet er gennemført.

### Relation mellem potentielle udfald og SCM'er

I SCM'er kan potentielle udfald repræsenteres ved hjælp af kontrafakta. I eksemplet med øjenlægen opfylder SCM  $\mathcal{C}$ ,  $T \rightarrow B$ , at  $T = N_T$  og  $B = T \cdot N_B + (1 - T) \cdot (1 - N_B)$ . Derfor kan hver patient repræsenteres ved specifikke værdier for  $N_B$  og  $N_T$ . I Figur 2.10 er eksempelvis patient 43 karakteriseret af  $N_T = 1, N_B = 0$ , mens patient 44 opfylder  $N_T = 0, N_B = 1$ . De to led  $t = 0$  og  $t = 1$  svarer til indgreb på  $T$ . I alt haves

$$B_u(t = \tilde{t}) = B \text{ i SCM } \mathcal{C} | \mathbf{N} = \mathbf{n}_u; do(T := \tilde{t}), \quad (2.16)$$

hvor  $\mathbf{n}_u$  karakteriserer enhed  $u$  [Pearl, 2009, Equation (3.51)] [13]. Venstre side af ligning (2.16) er det potentielle udfald, og højre side er det kontrafaktiske SCM. Da alle støjled er deterministiske i den kontrafaktiske SCM, er den medførte fordeling af  $B$  også degenereret, og  $B$  er deterministisk (som krævet). I eksemplet vist i Figur 2.10 er der blevet taget 200 i.i.d. enheder ved anvendelse af Bernoullifordelinger  $N_T \sim Ber(0.6)$  og  $N_B \sim Ber(0.01)$ . I dette tilfælde er SUTVA opfyldt. I.i.d. antagelsen medfører, at enhederne ikke forstyrrer hinanden og modulariteten (indgreb på  $T$  ændrer kun den strukturelle anvisning for  $T$ ) giver det, at måden behandlingen modtages ikke har nogen indflydelse på resultatet.

Nu diskuteres et resultat, der viser, i hvilken forstand begge repræsentationer i ligning (2.16) er ækvivalente. Til dette følges repræsentationen i Pearl [2009, 7.3.1] [13] og Halpern [2000] [4]. Argumentationen er baseret på følgende trin:

1. Definér egenskaberne (aksiomer): (C0)-(C5) og (MP) [Halpern, 2000, Section 3] [4].
2. Disse aksiomer er opfyldt i begge repræsentationer ("soundness")
3. Det kan vises, at disse egenskaber er fuldstændige for kontrafaktiske SCM'er. Ethvert kontrafaktisk udsagn følger fra et af disse aksiomer.
4. Det kan konkluderes, at enhver sætning, som holder for kontrafaktiske SCM'er, holder for potentielle udfald og omvendt (for at det gælder omvendt, kræves det, at den potentielle udfaldsramme ikke antager mere end aksiomerne, som er nævnt). Det følger også fra trin 3, at ethvert datasæt (ligesom i Figur 2.10), som opfylder de tre aksiomer, kunne være modelleret med en kontrafaktisk SCM. Hvis ingen SCM kan generere dette datasæt, betyder det, at kontrafakta fra SCM'er ville opfylde en anden egenskab, som ikke er medført fra de tre aksiomer, nemlig den egenskab, at dette datasæt ikke kan genereres.

De to verdener adskiller sig i deres sprog. Selv hvis hver sætning holder i begge rammer, kan nogle sætninger være nemmere at bevise i en verden end i den anden. Enhver antagelse i en sætning medfører restriktioner på den underliggende datagenererende proces; afhængig af applikationen kan en formulering muligvis forenkle vurderingen af disse restriktioner. Ses der på, hvor den gennemsnitlige kausale effekt er nul, men de individuelle kausale effekter er ikke-nul, så ser det ud til at være nemmere for potentielle udfald. På den anden side kan den grafiske repræsentation af SCM'er muligvis være nyttig til at udnytte antagelser på de kausale forhold mellem tilfældige variabler.

Richardson and Robins [2013] [15] foreslår at anvende *single world intervention graphs*. Disse grafer tillader at sætte variabler til bestemte værdier, og derfor konstruere grafiske sammenhænge til kontrafaktiske variabler. Disse modificerede grafer giver mulighed for at aflæse betingede uafhængighedsudsagn, der involverer både faktuelle og kontrafaktiske variabler. Derfor kan disse grafer ses som et nyttigt værktøj til at oversætte grafiske antagelser til kontrafaktiske udsagn, som ofte anvendes af potentielle udfaldsanalytikere.

# Databehandling 3

---

Dette kapitel er baseret på kapitel 12 og 13 i [17], hvor koden tilhørende disse kapitler er tilpasset dette projekts datasæt.

I dette kapitel tages der udgangspunkt i datasættet ”Framingham Heart Study”, som startede i 1948, hvor hvert individ blev fulgt i 24 år, såfremt individerne ikke blev censureret eller døde. Hvert sjette år blev individerne undersøgt, altså tre perioder i alt. I vores datasæt er baseline for målingerne fra periode 1. For at gøre det nemmere at afprøve teorien på datasættet, ser vi kun på de individer, som var med i både periode 1 og 3, så slutmålingerne på individerne er fra periode 3.

## Indblik i data

```
1 frmgham2 <- read_csv("csv/frmgham2.csv")
2
3 frmgham2 <- frmgham2[, -c(3, 11, 13, 22, 23)]
4 frmgham2 <- na.omit(frmgham2)
5
6 frmg_per1 <- subset(frmgham2, PERIOD==1)
7 frmg_per3 <- subset(frmgham2, PERIOD==3)
```

Vi indlæser datasættet og fjerner de variabler, som har mange manglende værdier og ikke er relevante, så vi har et mere rent datasæt at arbejde med. Derefter deles datasættet op i to. Det ene er ”frmg\_per1”, som indeholder de målinger, der blev taget i periode 1, og ”frmg\_per3”, som indeholder de målinger, der blev taget i periode 3.

De variabler, der tages med i analysen, er:

- ”RANDID”, som er identifikationen på individerne i datasættet.
- ”SEX”, som er kønnet på individerne. 1 for mænd og 2 for kvinder.
- ”AGE”, som er alderen for individerne målt i år i periode 1.
- ”educ”, som er uddannelsesniveaueet i periode 1. 1 for 0 – 11 års uddannelse, 2 for high school diploma eller GED, 3 for lidt college eller vocational school og 4 for college/bachelorgrad eller højere.
- ”BMI”, som er individernes BMI i periode 1.

- "CURSMOKE", som er om individerne ryger eller ej i periode 1. 0 for ikke-rygere og 1 for rygere.
- "CIGPDAY", som er antallet af cigaretter per dag i periode 1.
- "DEATH", som er om individet døde efter periode 3.
- "bmidef", som er forskellen på BMI i periode 3 og periode 1.
- "qsmk", som er om individet stoppede med at ryge. 0 for hvis der ikke var nogen ændring, og 1 hvis de stoppede med at ryge.

Den følgende kode sikrer os, at vi kan samle variableerne i ét datasæt til sidst.

```

1 dat1.1 <- data.frame(frmg_per3$RANDID, frmg_per3$SEX, frmg_per3$AGE,
2                     frmg_per3$educ, frmg_per3$BMI, frmg_per3$CURSMOKE,
3                     frmg_per3$CIGPDAY, frmg_per3$DEATH)
4
5 dat2 <- subset(frmg_per1, frmg_per1$RANDID%in%dat1.1$frmg_per3.RANDID)
6 dat1.1 <- subset(dat1.1, dat1.1$frmg_per3.RANDID%in%dat2$RANDID)
7
8 dat1.1$bmidef <- dat1.1$frmg_per3.BMI-dat2$BMI

```

I linje 5-6 i den overstående kode sikrer vi os for kun at inkludere individer, der er med i begge perioder, sådan at forskellen på BMI kan beregnes. Herefter beregnes variabelen "qsmk", hvor hvis "CURSMOKE" er lig med 1 i periode 1 og 0 i periode 3, så fås værdien 1, når forskellen bestemmes, hvilket betyder, at de er stoppet med at ryge, altså "qsmk" lig med 1.

```

1 dat1.1$qsmk <- 0
2
3 for (i in 1:length(dat1.1$frmg_per3.RANDID)) {
4   if (dat2$CURSMOKE[i]-dat1.1$frmg_per3.CURSMOKE[i]==1)
5     {dat1.1$qsmk[i] <- 1}
6   else {dat1.1$qsmk[i] <- 0}
7 }

```

Nu samles variableerne i ét datasæt, hvor de er taget fra de passende perioder, som i beskrivelsen af variableerne.

```

1 RANDID <- dat1.1$frmg_per3.RANDID
2 SEX <- dat1.1$frmg_per3.SEX
3 AGE <- dat2$AGE
4 educ <- dat2$educ
5 BMI <- dat2$BMI
6 CURSMOKE <- dat2$CURSMOKE
7 CIGPDAY <- dat2$CIGPDAY

```

```

8 DEATH <- dat1.1$frmg_per3.DEATH
9 qsmk <- dat1.1$qsmk
10 bmidef <- dat1.1$bmidef
11
12 dat1 <- data.frame(RANDID, SEX, AGE, educ, BMI, CURSMOKE, CIGPDAY,
13                    DEATH, qsmk, bmidef)

```

Det, der gerne vil udregnes nu, er associationsforskellen  $E[Y|A = 1] - E[Y|A = 0]$  på, om man er stoppet med at ryge  $A = 1$ , eller om der ikke er nogen ændring  $A = 0$  og indflydelsen af dette på BMI  $Y$ . Det første, der gøres, er at tilpasse en lineær model for at kunne estimere  $E[Y|A = 1]$  og  $E[Y|A = 0]$ .

```

1 modell <- lm(bmidef ~ qsmk, data = dat1)
2 summary(modell)
3 #Result
4 Call:
5 lm(formula = bmidef ~ qsmk, data = dat1)
6
7 Residuals:
8      Min       1Q   Median       3Q      Max
9 -17.1186  -1.2686   0.0814   1.3264  12.0114
10
11 Coefficients:
12             Estimate Std. Error t value Pr(>|t|)
13 (Intercept)  0.06862    0.04571   1.501   0.133
14 qsmk         0.61499    0.11980   5.133 3.02e-07 ***
15 ---
16 Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
17
18 Residual standard error: 2.362 on 3123 degrees of freedom
19 Multiple R-squared:  0.008367, Adjusted R-squared:  0.00805
20 F-statistic: 26.35 on 1 and 3123 DF, p-value: 3.02e-07

```

Det følgende er estimatet af henholdsvis  $E[Y|A = 1]$  og  $E[Y|A = 0]$ .

```

1 predict(modell, data.frame(qsmk=1))
2 #Result
3 0.6836044

```

```

1 predict(modell, data.frame(qsmk=0))
2 #Result
3 0.06861798

```

Dette betyder, at associationsforskellen er  $0.68 - 0.07 = 0.61$ . Det kan ses ud fra summary af ”modell”, at associationsforskellen er signifikant, så der er lille sandsynlighed for, at den er lig med 0, så det er næsten sikkert, at man tager på i vægt, når man stopper med at ryge. Det ser dog ikke ud til, at man tager særlig meget på i vægt, når man stopper med at ryge, og der kan være confounding variabler, der også har indflydelse på dette. Denne associationsforskel giver os et groft estimat på kausalitetsforskellen  $E[Y^{a=1}] - E[Y^{a=0}]$ , men dette estimat kan underestimere den sande værdi.

I de følgende tabeller ses der på, hvordan rygestop er fordelt mellem køn.

```

1 table(dat1$qsmk, dat1$SEX)
2 prop.table(table(dat1$qsmk, dat1$SEX), 2)
3 #Result
4      1      2
5 0 1048 1622
6 1  275  180
7
8      1      2
9 0 0.79213908 0.90011099
10 1 0.20786092 0.09988901

```

Det kan ses ud fra vores data, at hvis man er en mand, har man sandsynlighed 0.2 for at stoppe med at ryge, mens at sandsynligheden for kvinder er 0.1. Dette betyder ikke nødvendigvis, at der er flere kvinder, der ryger. Det kan lige så godt være, at der var færre kvinder, der røg til at starte med. Det, der kan siges ud fra disse tabeller, er, at ud af de mænd og kvinder der røg, så stoppede 20% af mændene og 10% af kvinderne.

Nu ses der på fordelingen af rygestop mellem uddannelsesniveauer.

```

1 table(dat1$qsmk, dat1$educ)
2 prop.table(table(dat1$qsmk, dat1$educ), 2)
3 #Result
4      1      2      3      4
5 0 1077  823  459  311
6 1  179  128   76   72
7
8      1      2      3      4
9 0 0.8574841 0.8654048 0.8579439 0.8120104
10 1 0.1425159 0.1345952 0.1420561 0.1879896

```

Det eneste vi kan konkludere ud fra disse tabeller er, at der ikke er stor forskel på rygestop mellem uddannelsesniveauer. Den eneste, der skiller sig lidt ud, er niveau 4, men det er ikke meget.

I næste del estimeres IP vægtene, hvor der tages højde for de confounding variabler. Det antages, at de målte variabler er nok til at justere for confounding.

### 3.1 Estimering ved anvendelse af IP vægtning

I hele dette afsnit estimeres parametrene i modellerne ved at anvende IP vægtning og udvidelser på dette. På grund af den måde datasættet er blevet behandlet, er der nogle antagelser, der gøres. Den vigtigste antagelse er, at der kun ses på individer, der ikke er censureret, altså  $C = 0$ . Så når der skrives  $E[Y|A = a]$ , er det indforstået, at det er  $E[Y|A = a, C = 0]$ . Dette gælder hver gang, der betinges på noget.

#### Estimering af parametre med IP vægtning

Det første, der gøres, er at tilpasse en generaliseret lineær model (glm) med den binomiale familie, så det er logistisk regression, da variabelen "qsmk" er binær. Responsvariablen i denne model er "qsmk", og de andre variabler i modellen er "SEX", "AGE", "AGE<sup>2</sup>", "educ", "CIGPDAY", "CIGPDAY<sup>2</sup>", "BMI" og "BMI<sup>2</sup>". De variabler, der kvadreres, er de kontinuerte variabler. Dette anvendes til at udregne IP vægtene på alle individer i datasættet, som senere anvendes til at estimere parameteren for rygestop.

```
1 fit <- glm(qsmk~ SEX + AGE+ I(AGE^2)+ as.factor(educ)+ CIGPDAY+
2           I(CIGPDAY^2)+ BMI + I(BMI^2) ,
3           family = binomial() ,
4           data = dat1)
5 summary(fit)
6 #Result
7 Call:
8 glm(formula = qsmk ~ SEX + AGE + I(AGE^2) + as.factor(educ) +
9     CIGPDAY + I(CIGPDAY^2) + BMI + I(BMI^2), family = binomial(),
10    data = dat1)
11
12 Deviance Residuals:
13     Min       1Q   Median       3Q      Max
14 -1.2704  -0.5928  -0.3595  -0.2854   2.8186
15
16
```

```

17 Coefficients:
18             Estimate Std. Error z value Pr(>|z|)
19 (Intercept)   -3.309e+00  2.446e+00  -1.353 0.176083
20 SEX           -4.179e-01  1.183e-01  -3.533 0.000411 ***
21 AGE            2.041e-02  8.140e-02   0.251 0.802012
22 I (AGE^2)      9.704e-05  8.136e-04   0.119 0.905060
23 as.factor(educ)2 -6.739e-02  1.363e-01  -0.494 0.621036
24 as.factor(educ)3  1.487e-01  1.609e-01   0.924 0.355391
25 as.factor(educ)4  3.516e-01  1.672e-01   2.103 0.035442 *
26 CIGPDAY        1.704e-01  1.277e-02  13.345 < 2e-16 ***
27 I (CIGPDAY^2)  -3.272e-03  3.493e-04  -9.367 < 2e-16 ***
28 BMI           -2.762e-02  1.011e-01  -0.273 0.784736
29 I (BMI^2)       7.629e-04  1.819e-03   0.419 0.674924
30 ---
31 Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
32
33 (Dispersion parameter for binomial family taken to be 1)
34
35 Null deviance: 2593.8 on 3124 degrees of freedom
36 Residual deviance: 2244.3 on 3114 degrees of freedom
37 AIC: 2266.3
38
39 Number of Fisher Scoring iterations: 5

```

For at udregne estimatet af IP vægtene

$$\hat{W} = \frac{1}{\hat{P}(A = 1|L)}$$

for dem, der stopper med at ryge, og

$$\hat{W} = \frac{1}{1 - \hat{P}(A = 1|L)}$$

for dem, der ikke stopper, benyttes en "ifelse" funktion til først at udregne sandsynlighederne for individerne. Derefter tages den reciprokke af sandsynlighederne og så fås de estimerede IP vægte.

```

1 p.qsmk.obs <- ifelse(dat1$qsmk==0, 1- predict(fit, type = "response"),
2                   predict(fit, type = "response"))
3
4 dat1$w <- 1/p.qsmk.obs
5 summary(dat1$w)
6

```



```

7 #Result
8   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
9   1.001  1.054  1.095  1.943  1.424  53.100

```

Middelværdien af disse IP vægte estimeres til at være 1.943, og den mindste IP vægt er 1.001, mens den største er 53.1. Det kan muligvis siges, at 53.1 er en stor værdi for en IP vægt, men den er ikke urealistisk. Det betyder bare, at der var et individ, der blev forudsagt at have sandsynlighed 0.019 for at stoppe med at ryge, men fortsatte alligevel. Eller omvendt en der blev forudsagt med stor sandsynlighed for at stoppe med at ryge, som ikke stoppede. Dette sker også i virkeligheden, så disse IP vægte giver mening at benytte til parameter estimation.

For at estimere parameteren tilpasses en lineær model på formen  $E[Y|L] = \beta_0 + \beta_1 A$ . Her er  $A$  "qsmk" og  $Y$  er "bmidef", hvor der tages højde for individernes IP vægte. For at finde et 95% konfidensinterval omkring vores parameter estimat,  $\hat{\beta}_1$ , anvendes funktionen "geeglm", som tilpasser en GEE model. Dette gøres, da GEE modeller anvender en robust varians estimator, når korrelationsstrukturen er uafhængig.

```

1 library(geepack)

1 fitgee.w <- geeglm(bmidef~ qsmk, data = dat1, weights = w,
2                   id =RANDID, corstr = "independence" )
3 summary(fitgee.w)
4 #Result
5 Call:
6 geeglm(formula = bmidef ~ qsmk, data = dat1, weights = w,
7        id = RANDID, corstr = "independence")
8
9 Coefficients:
10             Estimate Std.err  Wald Pr(>|W|)
11 (Intercept)   0.0426  0.0453   0.88    0.35
12 qsmk          0.6598  0.1537  18.43  1.8e-05 ***
13 ---
14 Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
15
16 Correlation structure = independence
17 Estimated Scale Parameters:
18
19             Estimate Std.err
20 (Intercept)    5.49    0.311
21 Number of clusters: 3125 Maximum cluster size: 1

```

```

1 beta <- coef(fitgee.w)
2 SE <- coef(summary(fitgee.w))[, 2]
3 lcl <- beta - qnorm(0.975) * SE
4 ucl <- beta + qnorm(0.975) * SE
5 cbind(beta, lcl, ucl)
6 #Result
7           beta      lcl    ucl
8 (Intercept) 0.0426 -0.0462 0.131
9 qsmk        0.6598  0.3585 0.961

```

Konfidensintervallet for  $\beta_1$  er (0.3585, 0.961), hvilket er et konservativt estimat, som dækker den sande værdi for super-populationen mere end 95% af gangene. Vores estimat er  $\hat{\beta}_1 = 0.6598$ , hvilket betyder, at hvis man stopper med at ryge, stiger ens BMI med 0.6598. For et individ med højde 181cm, hvilket svarer til en gennemsnitlig mand i Danmark, vil det sige, at de har taget 2.16kg på i vægt. Konfidensintervallet svarer til (1.17, 3.15)kg. Dette betyder, at nulhypotesen om at rygestop ikke har nogen effekt på vægt, kan forkastes.

### Estimering af stabiliserede IP vægte

Nu estimeres parameteren  $\beta_1$  ved anvendelse af stabiliserede IP vægte. De er forskellige fra IP vægte  $W^A = \frac{1}{f(A|L)}$  ved, at der ikke står 1 i tælleren, men der står  $f(A)$ ,  $SW^A = \frac{f(A)}{f(A|L)}$ . Disse vægte estimeres næsten på samme måde som tidligere, med at der tilpasses en glm til start for at finde nævneren.

```

1 denom.fit <- glm(qsmk~ as.factor(SEX) + AGE+ I(AGE^2)+
2               as.factor(educ)+ CIGPDAY+
3               I(CIGPDAY^2)+ BMI + I(BMI^2),
4               family = binomial(),
5               data = dat1)
6 summary(denom.fit)
7 #Result
8 Call:
9 glm(formula = qsmk ~ as.factor(SEX) + AGE + I(AGE^2)
10     + as.factor(educ) + CIGPDAY + I(CIGPDAY^2) + BMI + I(BMI^2),
11     family = binomial(), data = dat1)
12
13 Deviance Residuals:
14    Min       1Q   Median       3Q      Max
15 -1.270  -0.593  -0.359  -0.285   2.819
16

```

```

17 Coefficients:
18             Estimate Std. Error z value Pr(>|z|)
19 (Intercept)   -3.726602   2.426814   -1.54  0.12464
20 as.factor(SEX)2 -0.417914   0.118286   -3.53  0.00041 ***
21 AGE           0.020411   0.081402    0.25  0.80201
22 I(AGE^2)      0.000097   0.000814    0.12  0.90506
23 as.factor(educ)2 -0.067391   0.136314   -0.49  0.62104
24 as.factor(educ)3  0.148660   0.160855    0.92  0.35539
25 as.factor(educ)4  0.351600   0.167169    2.10  0.03544 *
26 CIGPDAY       0.170372   0.012766   13.35 < 2e-16 ***
27 I(CIGPDAY^2)  -0.003272   0.000349   -9.37 < 2e-16 ***
28 BMI           -0.027618   0.101109   -0.27  0.78474
29 I(BMI^2)      0.000763   0.001819    0.42  0.67492
30 ---
31 Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
32
33 (Dispersion parameter for binomial family taken to be 1)
34
35 Null deviance: 2593.8 on 3124 degrees of freedom
36 Residual deviance: 2244.3 on 3114 degrees of freedom
37 AIC: 2266
38
39 Number of Fisher Scoring iterations: 5

```

Herunder kan det ses, hvordan nævneren "pd.qsmk" estimeres. Herefter tilpasses en glm for at estimere tælleren i  $SW^A$ .

```

1 pd.qsmk <- predict(denom.fit, type = "response")
2 numer.fit <- glm(qsmk ~ 1, family = binomial(), data = dat1)
3 summary(numer.fit)
4 #Result
5 Call:
6 glm(formula = qsmk ~ 1, family = binomial(), data = dat1)
7
8 Deviance Residuals:
9   Min       1Q   Median       3Q      Max
10 -0.561  -0.561  -0.561  -0.561   1.963
11
12 Coefficients:
13             Estimate Std. Error z value Pr(>|z|)
14 (Intercept)  -1.7695     0.0507  -34.9  <2e-16 ***

```

```

15 ---
16 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
17
18 (Dispersion parameter for binomial family taken to be 1)
19
20 Null deviance: 2593.8 on 3124 degrees of freedom
21 Residual deviance: 2593.8 on 3124 degrees of freedom
22 AIC: 2596
23
24 Number of Fisher Scoring iterations: 4

```

Estimatet for individer, der stopper med at ryge, er

$$\frac{\hat{P}(A = 1)}{\hat{P}(A = 1|L)}$$

og for dem, der ikke stopper

$$\frac{1 - \hat{P}(A = 1)}{1 - \hat{P}(A = 1|L)}.$$

```

1 pn.qsmk <- predict(numer.fit, type = "response")
2
3 dat1$sw <-
4   ifelse(dat1$qsmk == 0, ((1 - pn.qsmk) / (1 - pd.qsmk)),
5         (pn.qsmk / pd.qsmk))
6
7 summary(dat1$sw)
8 #Result
9   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
10  0.27  0.89   0.92   0.99  1.07   7.73

```

I teorien er middelværdien for de stabiliserede IP vægte lig med 1, og vores estimerede middelværdi er lig med 0.99, så den er tæt på. Intervallet på de stabiliserede IP vægte er (0.27, 7.73), hvor de tidligere IP vægte havde intervallet (1.001, 53.1). Det mindre interval kommer fra, at der i tælleren står  $f(A)$ , som er en stabiliserende faktor.

```

1 fitgee.sw <- geeglm(bmidef~ qsmk, data = dat1, weights = sw,
2                   id = RANDID, corstr = "independence" )
3 summary(fitgee.sw)
4 #Result
5 Call:
6 geeglm(formula = bmidef ~ qsmk, data = dat1, weights = sw,
7        id = RANDID, corstr = "independence")

```

```

8
9 Coefficients:
10             Estimate Std.err   Wald Pr(>|W|)
11 (Intercept)  0.0426  0.0453   0.88    0.35
12 qsmk        0.6598  0.1537  18.43  1.8e-05 ***
13 ---
14 Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
15
16 Correlation structure = independence
17 Estimated Scale Parameters:
18
19             Estimate Std.err
20 (Intercept)    5.45    0.224
21 Number of clusters: 3125 Maximum cluster size: 1

```

```

1 beta <- coef(fitgee.sw)
2 SE <- coef(summary(fitgee.sw))[, 2]
3 lcl <- beta - qnorm(0.975) * SE
4 ucl <- beta + qnorm(0.975) * SE
5 cbind(beta, lcl, ucl)
6 #Result
7             beta      lcl    ucl
8 (Intercept) 0.0426 -0.0462 0.131
9 qsmk        0.6598  0.3585 0.961

```

Estimatet for  $\beta_1$  med de stabiliserede IP vægte er  $\hat{\beta}_1 = 0.6598$ , og det 95% konfidensinterval er (0.3585, 0.961). Dette er lig med det tidligere resultat med ikke-stabiliserede IP vægte. Så nogen kan undre sig over, hvorfor man benytter sig af stabiliserede IP vægte, men estimerne er kun lig med hinanden, når der tilpasses en mættet lineær model på data, hvilket betyder, at der er observationer for alle mulige værdier af  $A$ , og i vores model er der kun to mulige værdier for  $A$ . Ofte har man ikke en mættet model, og så får man ofte et smallere konfidensinterval ved at benytte sig af stabiliserede IP vægte.

### Estimering af parameter med kontinuert behandlingsvariabel

I stedet for at anvende en binær variabel, prøver vi at estimere parametrene, når der haves en kontinuert behandlingsvariabel. Denne kontinuerte behandlingsvariabel er "cigdef", som er forskellen på antallet af cigaretter per dag, der blev røget i periode 3 og periode 1. Det undersøges nu, om "cigdef" har indflydelse på "bmidef".

```

1 dat1$cigdef <- dat1.1$frmg_per3.CIGPDAY-dat1$CIGPDAY
2 cig.denom.fit <- lm(cigdef~ as.factor(SEX)+AGE+ I(AGE^2)
3                   +as.factor(educ)+ CIGPDAY+ I(CIGPDAY^2)+ BMI
4                   + I(BMI^2), data = dat1)
5
6 p.den <- predict(cig.denom.fit, type = "response")
7 dens.den <- dnorm(dat1$cigdef, p.den, summary(cig.denom.fit)$sigma)

```

Når de stabiliserede IP vægte skal estimeres, kan der ikke gøres som tidligere, da der haves en kontinuert behandlingsvariabel. Den kontinuerte behandlingsvariabel betyder, at  $f(A|L)$  i stedet for at være en sandsynlighed nu er en pdf, som antages at være tætheden for en normalfordeling, som bruges til at beregne middelværdierne for individerne. Disse beregnede middelværdier gemmes i middelværdivektoren "p.den", og variansen kommer fra den tilpassede model "cig.denom.fit". I næste del findes tælleren ved at tilpasse en lineær model.

```

1 cig.numer.fit <- lm(cigdef ~ 1, data = dat1)
2 p.num <- predict(cig.numer.fit, type = "response")
3 dens.num <- dnorm(dat1$cigdef, p.num, summary(cig.numer.fit)$sigma)

```

```

1 dat1$sw.a <- dens.num / dens.den
2 summary(dat1$sw.a)
3 #Result
4   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
5     0      1      1    1418      1 4163111

```

Det kan ses, at vi får en minimumsværdi på 0 og en maksimumsværdi på 4163111, hvilket umiddelbart ikke giver mening i denne sammenhæng, men for at afprøve teori og koden fortsættes der med at arbejde med det. Man skal også være opmærksom på, at IP vægte for kontinuerte behandlingsvariabler er følsomme for valget af modellen, der tilpasses den betingede tæthed  $f(A|L)$ .

```

1 fitgee.sw.a <- geeglm(bmidef~ cigdef+ I(cigdef*cigdef), data = dat1,
2                   weights = sw.a, id =frmg_per3.RANDID,
3                   corstr = "independence" )
4 summary(fitgee.sw.a)
5 #Result
6 Call:
7 geeglm(formula = bmidef ~ cigdef + I(cigdef * cigdef), data = dat1,
8       weights = sw.a, id = RANDID, corstr = "independence")
9
10

```

```

11 Coefficients:
12             Estimate   Std.err   Wald Pr(>|W|)
13 (Intercept)   -0.880570  0.426415  4.26   0.039 *
14 cigdef        -0.075042  0.032733  5.26   0.022 *
15 I(cigdef * cigdef)  0.004052  0.000571 50.36  1.3e-12 ***
16 ---
17 Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
18
19 Correlation structure = independence
20 Estimated Scale Parameters:
21
22             Estimate Std.err
23 (Intercept)   0.0138   0.015
24 Number of clusters: 3125 Maximum cluster size: 1

```

```

1 beta <- coef(fitgee.sw.a)
2 SE <- coef(summary(fitgee.sw.a))[, 2]
3 lcl <- beta - qnorm(0.975) * SE
4 ucl <- beta + qnorm(0.975) * SE
5 cbind(beta, lcl, ucl)
6 #Result
7             beta      lcl      ucl
8 (Intercept)  -0.88057 -1.71633 -0.04481
9 cigdef       -0.07504 -0.13920 -0.01089
10 I(cigdef * cigdef)  0.00405  0.00293  0.00517

```

Ud fra disse 95% konfidensintervaller kan det ses, at nulhypotesen om at "cigdef" ikke har nogen indflydelse på "bmidef", kan forkastes, da konfidensintervallerne ikke indeholder 0. Hvis et konfidensinterval indeholder 0, betyder det, at estimatet på parameteren ikke er forskellig nok fra 0 til at sige, om parameteren har nogen indflydelse på responsvariablen "bmidef". Da vores konfidensintervaller ikke indeholder 0, har "cigdef" indflydelse på "bmidef". Denne signifikans kan muligvis komme fra, at vores vægte er mærkelige, og at den største vægt har en værdi over 4 millioner.

### Estimering af parametre i en logistisk model

I denne model kigges der på, om rygestop har en indflydelse på dødfald. Det er en logistisk model, da variabelen "DEATH" er binær og ikke kontinuert. Der laves først en tabel for at få et indblik i, hvordan individerne er fordelt blandt dødsfald og rygestop.

```

1 table(dat1$qsmk, dat1$DEATH)

```

```

2 #Result
3     0    1
4  0 2060  610
5  1   334  121

```

Den følgende kode er den samme opskrift som i de tidligere modeller. De vægte, der bliver anvendt i modellen, er de tidligere estimerede stabiliserede IP vægte.

```

1 fitgee.sw.log <- geeglm(DEATH~ qsmk, data = dat1,
2                       weights = sw, id =RANDID,
3                       corstr = "independence" )
4 summary(fitgee.sw.log)
5 #Result
6 Call:
7 geeglm(formula = DEATH ~ qsmk, data = dat1, weights = sw,
8       id = RANDID, corstr = "independence")
9
10 Coefficients:
11             Estimate Std.err   Wald Pr(>|W|)
12 (Intercept)  0.23756  0.00848 784.05  <2e-16 ***
13 qsmk        -0.03166  0.02410   1.73   0.19
14 ---
15 Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
16
17 Correlation structure = independence
18 Estimated Scale Parameters:
19
20             Estimate Std.err
21 (Intercept)    0.179 0.00427
22 Number of clusters:   3125 Maximum cluster size: 1

```

```

1 beta <- coef(fitgee.sw.log)
2 SE <- coef(summary(fitgee.sw.log))[, 2]
3 lcl <- beta - qnorm(0.975) * SE
4 ucl <- beta + qnorm(0.975) * SE
5 cbind(beta, lcl, ucl)
6 #Result
7             beta    lcl    ucl
8 (Intercept)  0.2376  0.2209 0.2542
9 qsmk        -0.0317 -0.0789 0.0156

```



Ud fra 95% konfidensintervallet for "qsmk" kan nulhypotesen ikke forkastes, da den indeholder 0. Dette betyder, at rygestop ikke har nogen indflydelse på, om man dør ifølge denne model. Intuitivt giver det mening, at individer har mindre sandsynlighed for at dø, hvis de stopper med at ryge. Vores resultat kan komme fra, at datasættet omhandler individer, der har været med i et studie om hjerte-kar-sygdomme, og derfra kan der være bias.

### Indflydelse af køn på rygestop

Der ses nu på, om køn har en indflydelse på, om man stopper med at ryge. Dette gøres ved at inkludere en kovariat, som muligvis ikke er en confounder, for at kunne bedømme om kovariaten har nogen effekt. Denne model ser ud på følgende måde

$$E[Y^a|V] = \beta_0 + \beta_1 a + \beta_2 V + \beta_3 Va,$$

hvor kovariaten har en effekt, hvis  $\beta_3 \neq 0$ . Til dette skal vi finde nogle stabiliserede IP vægte. Der benyttes den tidligere udregnede "pd.qsmk" som nævneren, og herefter udregnes tælleren med samme opskrift som tidligere.

```

1 numer.fit <-
2   glm(qsmk ~ as.factor(SEX), family = binomial(),
3     data = dat1)
4 summary(numer.fit)
5 #Result
6 Call:
7 glm(formula = qsmk ~ as.factor(SEX), family = binomial(), data = dat1)
8
9 Deviance Residuals:
10  Min       1Q   Median       3Q      Max
11 -0.683  -0.683  -0.459  -0.459   2.147
12
13 Coefficients:
14             Estimate Std. Error z value Pr(>|z|)
15 (Intercept)    -1.3379     0.0678  -19.8   <2e-16 ***
16 as.factor(SEX)2 -0.8606     0.1037   -8.3   <2e-16 ***
17 ---
18 Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
19
20 (Dispersion parameter for binomial family taken to be 1)
21
22 Null deviance: 2593.8  on 3124  degrees of freedom
23 Residual deviance: 2523.1  on 3123  degrees of freedom
24 AIC: 2527

```

```
25 |
26 | Number of Fisher Scoring iterations: 4
```

```
1 | pn.qsmk <- predict(numer.fit, type = "response")
2 |
3 | dat1$sw.sex <-
4 |   ifelse(dat1$qsmk == 0, ((1 - pn.qsmk) / (1 - pd.qsmk)),
5 |         (pn.qsmk / pd.qsmk))
6 |
7 | summary(dat1$sw.sex)
8 | #Result
9 |   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
10 |  0.24   0.87   0.95   0.99   1.08   11.04
```

Disse vægte ser fornuftige ud, da middelværdien, 0.99, er tæt på hvad den teoretiske middelværdi skal være for de stabiliserede IP vægte, nemlig 1.

Nu tilpasses der en GEE model med variableerne "qsmk", "SEX" og interaktionsvariablen mellem disse. Dette gøres for at se, om der er interaktion mellem rygestop og køn.

```
1 | fitgee.sw.sex <- geeglm(bmidef~ as.factor(qsmk)+ as.factor(SEX)+
2 |                       as.factor(qsmk):as.factor(SEX)
3 |                       , data = dat1, weights = sw.sex,
4 |                       id =RANDID, corstr = "independence" )
5 | summary(fitgee.sw.sex)
6 | #Result
7 | Call:
8 | geeglm(formula = bmidef ~ as.factor(qsmk) + as.factor(SEX) +
9 |        as.factor(qsmk):as.factor(SEX), data = dat1, weights = sw.sex,
10 |        id = RANDID, corstr = "independence")
11 |
12 | Coefficients:
13 |
14 |                Estimate Std.err  Wald Pr(>|W|)
15 | (Intercept)          -0.1867  0.0635   8.63  0.0033 **
16 | as.factor(qsmk)1           0.6714  0.2293   8.58  0.0034 **
17 | as.factor(SEX)2           0.3967  0.0893  19.73  8.9e-06 ***
18 | as.factor(qsmk)1:as.factor(SEX)2 -0.0345  0.3045   0.01  0.9099
19 | ---
20 | Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
21 | Correlation structure = independence
```

```

22 Estimated Scale Parameters:
23
24           Estimate Std.err
25 (Intercept)      5.47   0.227
26 Number of clusters: 3125 Maximum cluster size: 1

```

```

1 beta <- coef(fitgee.sw.sex)
2 SE <- coef(summary(fitgee.sw.sex))[, 2]
3 lcl <- beta - qnorm(0.975) * SE
4 ucl <- beta + qnorm(0.975) * SE
5 cbind(beta, lcl, ucl)
6 #Result
7           beta      lcl      ucl
8 (Intercept) -0.1867 -0.311 -0.0621
9 as.factor(qsmk) 1      0.6714  0.222  1.1207
10 as.factor(SEX) 2      0.3967  0.222  0.5718
11 as.factor(qsmk) 1:as.factor(SEX) 2 -0.0345 -0.631  0.5624

```

Ud fra disse 95% konfidensintervaller kan det ses, at interaktionsvariablen har konfidensintervallet  $(-0.631, 0.5624)$ . Dette betyder at nulhypotesen, om at der ikke er interaktion mellem rygestop og køn, ikke kan forkastes. Dette tolkes som om, at der ikke er interaktion, og variabelen ikke er vigtig at have med i en model. Derudover kan det ses, at nulhypoteserne for variableerne ”qsmk” og ”SEX” kan forkastes, hvilket betyder, at de har indflydelse på ens vægt.

Der er nu blevet set på, hvordan IP vægtning kan anvendes, og i næste afsnit ses der på standardisering.

## 3.2 Standardisering

I dette afsnit estimeres den gennemsnitlige kausale effekt ved anvendelse af standardisering, se Definition 2.2.7. I den første del estimeres der kun et groft bud på, hvad den gennemsnitlige BMI forskel er for alle værdier af ”qsmk” ved anvendelse af en glm. Den tilpassede glm anvendes til at prædiktere BMI forskel for alle individer i datasættet, hvilket gemmes i en variabel, som kaldes ”predicted.meanY”, som anvendes til at finde det grove estimat ved at tage gennemsnittet af de prædikterede værdier.

```

1 fit <- glm(bmidef~ qsmk + SEX+ AGE+ I(AGE*AGE)+ as.factor(educ)
2           + CIGPDAY+ I(CIGPDAY*CIGPDAY)+ BMI + I(BMI*BMI)
3           + qsmk*CIGPDAY, data = dat1)
4 summary(fit)

```

```

5 #Result
6 Call:
7 glm(formula = bmidef ~ qsmk + SEX + AGE + I(AGE * AGE) +
8     as.factor(educ) + CIGPDAY + I(CIGPDAY * CIGPDAY) + BMI +
9     I(BMI * BMI) + qsmk * CIGPDAY, data = dat1)
10
11 Deviance Residuals:
12     Min       1Q   Median       3Q      Max
13 -14.239   -1.229   -0.021    1.160   11.796
14
15 Coefficients:
16             Estimate Std. Error t value Pr(>|t|)
17 (Intercept)      3.853357    1.726619    2.23  0.0257 *
18 qsmk              0.090237    0.194238    0.46  0.6423
19 SEX              0.268741    0.086650    3.10  0.0019 **
20 AGE             -0.004172    0.057587   -0.07  0.9423
21 I(AGE * AGE)    -0.000734    0.000575   -1.28  0.2019
22 as.factor(educ)2 -0.008065    0.097445   -0.08  0.9340
23 as.factor(educ)3 -0.323371    0.115386   -2.80  0.0051 **
24 as.factor(educ)4 -0.283174    0.129939   -2.18  0.0294 *
25 CIGPDAY         -0.028182    0.009129   -3.09  0.0020 **
26 I(CIGPDAY * CIGPDAY) 0.000381    0.000258    1.48  0.1397
27 BMI             -0.022914    0.071987   -0.32  0.7503
28 I(BMI * BMI)    -0.002079    0.001279   -1.63  0.1041
29 qsmk:CIGPDAY     0.046222    0.010158    4.55  5.6e-06 ***
30 ---
31 Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
32
33 (Dispersion parameter for gaussian family taken to be 4.79)
34
35 Null deviance: 17572 on 3124 degrees of freedom
36 Residual deviance: 14917 on 3112 degrees of freedom
37 AIC: 13781
38
39 Number of Fisher Scoring iterations: 2

```

```

1 dat1$predicted.meanY <- predict(fit, dat1)
2
3 summary(dat1$predicted.meanY)
4 #Result

```

```

5   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
6  -6.07  -0.45   0.23   0.16   0.80   2.65

```

```

1  summary(dat1$bmidef)
2  #Result
3   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
4  -17.05  -1.10   0.18   0.16   1.48   12.08

```

Det ses at estimatet på BMI forskellen har en middelværdi på 0.16, og en minimumsværdi på  $-6.07$  og en maksimumsværdi på  $2.65$ . Det kan ses, at middelværdien passer godt på de observerede BMI forskelle, da middelværdien af de observerede værdier er  $0.16$ , hvilket er det, der blev estimeret.

I afsnittet om standardisering blev der anvendt data fra Tabel 2.4, hvor det ikke var et problem at udregne  $P(L = l)$  for alle værdier af  $L$ , men når der arbejdes med data med mange variabler, kan det blive kompliceret. Så for at undgå at udregne  $P(L = l)$  benyttes egenskaben, at den vægtede middelværdi

$$\sum_l E[Y|A = a, L = l]P(L = l)$$

kan skrives som

$$E[E[Y|A = a, L]].$$

For at benytte dette anvendes en 4-trins metode, som først er udvidelse af datasættet, derefter modellering af udfald, prædiktion og til sidst standardisering ved gennemsnit. Udvidelse af datasættet sker på den måde, at der laves 3 kopier af datasættet, hvor der ikke ændres noget i den første kopi. I den anden kopi sættes behandling  $A$ , her "qsmk", til 0 altså ikke behandlet, og responsvariablen sættes til NA. Analogt for den tredje kopi dog sættes "qsmk" til 1, og herefter kombineres alle tre kopier i et samlet datasæt. Derefter tilpasses en glm, der bliver anvendt til at estimere middelværdien for de tre kopier af datasættet.

```

1  dat1$interv <- -1 #Første kopi af data lig med det originale datasæt.
2
3  interv0 <- dat1 #Anden kopi: behandling sat til 0, respons missing.
4  interv0$interv <- 0
5  interv0$qsmk <- 0
6  interv0$bmidef <- NA
7
8  interv1 <- dat1 #Tredje kopi: behandling sat til 1, respons missing.
9  interv1$interv <- 1

```

```

10 interv1$qsmk <- 1
11 interv1$bmiddef <- NA
12
13 onesample <- rbind(dat1, interv0, interv1)

```

```

1 std <- glm(bmiddef~ qsmk + SEX+ AGE+ I(AGE*AGE)+ as.factor(educ)
2           + CIGPDAY+ I(CIGPDAY*CIGPDAY)+ BMI +
3           I(BMI*BMI) + qsmk*CIGPDAY, data = onesample)
4 summary(std)
5 #Result
6 Call:
7 glm(formula = bmiddef ~ qsmk + SEX + AGE + I(AGE * AGE) +
8     as.factor(educ) + CIGPDAY + I(CIGPDAY * CIGPDAY) + BMI +
9     I(BMI * BMI) + qsmk * CIGPDAY, data = onesample)
10
11 Deviance Residuals:
12     Min       1Q   Median       3Q      Max
13 -14.239   -1.229   -0.021    1.160   11.796
14
15 Coefficients:
16             Estimate Std. Error t value Pr(>|t|)
17 (Intercept)    3.853357    1.726619    2.23  0.0257 *
18 qsmk           0.090237    0.194238    0.46  0.6423
19 SEX           0.268741    0.086650    3.10  0.0019 **
20 AGE          -0.004172    0.057587   -0.07  0.9423
21 I(AGE * AGE)  -0.000734    0.000575   -1.28  0.2019
22 as.factor(educ)2 -0.008065    0.097445   -0.08  0.9340
23 as.factor(educ)3 -0.323371    0.115386   -2.80  0.0051 **
24 as.factor(educ)4 -0.283174    0.129939   -2.18  0.0294 *
25 CIGPDAY      -0.028182    0.009129   -3.09  0.0020 **
26 I(CIGPDAY * CIGPDAY) 0.000381    0.000258    1.48  0.1397
27 BMI          -0.022914    0.071987   -0.32  0.7503
28 I(BMI * BMI)  -0.002079    0.001279   -1.63  0.1041
29 qsmk:CIGPDAY    0.046222    0.010158    4.55  5.6e-06 ***
30 ---
31 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
32
33 (Dispersion parameter for gaussian family taken to be 4.79)
34
35 Null deviance: 17572 on 3124 degrees of freedom

```

```
36 Residual deviance: 14917 on 3112 degrees of freedom
37 (6250 observations deleted due to missingness)
38 AIC: 13781
39
40 Number of Fisher Scoring iterations: 2
```

```
1 onesample$predicted.meanY <- predict(std, onesample)
2
3 mean(onesample[which(onesample$interv== -1), ]$predicted.meanY)
4 #Result
5 [1] 0.158
```

```
1 mean(onesample[which(onesample$interv== 0), ]$predicted.meanY)
2 #Result
3 [1] 0.0357
```

```
1 mean(onesample[which(onesample$interv== 1), ]$predicted.meanY)
2 #Result
3 [1] 0.522
```

Det første estimat fra den uændrede kopi, er lig med det tidligere grove estimat, hvilket giver god mening, da det er det samme data, der er blevet anvendt. Estimatet på den anden kopi er 0.0357, som er vores estimerede middelværdi, hvis alle individer i datasættet ikke stoppede med at ryge. Da dette er positivt, betyder det, at selv hvis man ikke stoppede med at ryge, tager man stadig på i vægt, dog ikke særlig meget. Den estimerede middelværdi, hvis alle stoppede med at ryge, er 0.522, hvilket er markant større, end hvis de ikke stoppede med at ryge. Dette betyder, at hvis man stopper med at ryge, tager man mere på i vægt, end hvis man ikke stopper med at ryge. Den estimerede gennemsnitlige kausale effekt bliver derfor  $0.522 - 0.0357 = 0.4863$ , hvilket giver, at når man stopper med at ryge, stiger ens BMI med 0.4863 ifølge modellen.

For at få et 95% konfidensinterval anvendes en metode kaldet ”bootstrap”. Denne metode går ud på, at man har et datasæt og en funktion, der udregner de interessante estimater. Når bootstrap begynder, laver den et nyt datasæt ud fra det givne ved at udvælge tilfældige individer i datasættet, hvor det samme individ godt kan udvælges flere gange. Derefter anvendes funktionen, og estimaterne findes. Dette gentages et vist antal gange, i vores tilfælde 1000 gange, og til sidst har man en vektor bestående af de interessante estimater. Derefter kan et 95% konfidensinterval udregnes på sædvanligvis.

```
1 library(boot)
```

```

1 standardization <- function(data, indices) {
2   # Lav de tre datasæet og kombiner dem som tidligere.
3   d <- data[indices, ]
4   d$interv <- -1
5   d0 <- d
6   d0$interv <- 0
7   d0$qsmk <- 0
8   d0$bmidef <- NA
9   d1 <- d
10  d1$interv <- 1
11  d1$qsmk <- 1
12  d1$bmidef <- NA
13  d.onesample <- rbind(d, d0, d1)
14
15  fit <- glm(bmidef~ qsmk + SEX+ AGE+ I(AGE*AGE)+ as.factor(educ)
16            + CIGPDAY+ I(CIGPDAY*CIGPDAY)+ BMI + I(BMI*BMI)
17            + qsmk*CIGPDAY, data = d.onesample)
18
19  d.onesample$predicted_meanY <- predict(fit, d.onesample)
20
21  # Estimer middelværdi for hver gruppe i interv og forskellen
22  # mellem interv=1 og interv=0.
23  return(c(
24    mean(d.onesample$predicted_meanY[d.onesample$interv == -1]),
25    mean(d.onesample$predicted_meanY[d.onesample$interv == 0]),
26    mean(d.onesample$predicted_meanY[d.onesample$interv == 1]),
27    mean(d.onesample$predicted_meanY[d.onesample$interv == 1]) -
28      mean(d.onesample$predicted_meanY[d.onesample$interv == 0])
29  ))
30 }

```

```

1 results <- boot(data = dat1,
2               statistic = standardization,
3               R=1000)
4
5 se <- c(sd(results$t[, 1]),
6        sd(results$t[, 2]),
7        sd(results$t[, 3]),
8        sd(results$t[, 4]))

```



```

9 mean <- results$t0
10 ll <- mean - qnorm(0.975)*se
11 ul <- mean + qnorm(0.975)*se
12
13 bootstrap <- data.frame(cbind(
14   c("Observed", "No_Treatment", "Treatment",
15     "Treatment_-_No_Treatment"),
16   mean,
17   se,
18   ll,
19   ul
20 ))
21 bootstrap
22 #Result
23
24           V1                mean                se
25 1      Observed  0.158160000000003  0.0420634136902096
26 2   No Treatment  0.0357019893213362  0.0445892298013628
27 3      Treatment  0.521713591505134    0.12490555109426
28 4 Treatment - No Treatment  0.486011602183798    0.13191191632042
29
30           ll                ul
31 1  0.0757172241003832  0.240602775899623
32 2 -0.051691295187715  0.123095273830387
33 3  0.276903209891257  0.766523973119011
34 4  0.227468997064113  0.744554207303483

```

Det fundne 95% konfidensinterval for den gennemsnitlige kausale effekt er (0.23, 0.74).

### 3.3 Konklusion af databehandling

For IP vægte estimerede vi, at den gennemsnitlige kausale effekt er 0.6598 med 95% konfidensinterval (0.3585, 0.961), og for standardisering er estimeret på den gennemsnitlige kausale effekt 0.486 med 95% konfidensinterval (0.23, 0.74). Der er forskel på estimererne på 0.1738, hvilket ikke er en stor forskel. Der er også overlap mellem deres 95% konfidensintervaller, hvilket kan betyde, at begge estimater tager fejl, men med 95% sikkerhed er den sande værdi i konfidensintervallerne. Vores første grove estimat på den gennemsnitlige kausale effekt var associationsforskellen på 0.61. Dette estimat ligger mellem estimererne for IP vægtning og standardisering, så det har været et godt bud, men ikke særlig præcist. Selvom de to estimater er forskellige, kan vi dog konkludere ud fra, at de begge to er positive, at hvis man stopper med at ryge, stiger ens BMI.

Disse estimater skal tages med skepticisme, da man kun er så sikker på estimaterne, som man er sikker på, at datasættet opfylder de følgende krav:

- Ombyttelighed,
- Positivitet,
- Konsistens,
- Ingen målingsfejl,
- Ingen model fejlspecifikation.

Disse krav kan være svære at opfylde, så når man laver kausal inferens, skal man være kritisk og passe på med, hvad man konkluderer ud fra observationsdata.

# Konklusion 4

---

Kausalitet kan estimeres ved at anvende begrebet gennemsnitlig kausale effekt, som for binære behandlingsvariabler og udfaldsvariabler estimeres ved  $E[Y|A = 1] - E[Y|A = 0]$ . Hvis ikke der tages højde for indflydelse fra andre variabler, er dette et groft estimat. Det er heller ikke nødvendigvis den gennemsnitlige kausale effekt, der er blevet estimeret. Det kan også være den gennemsnitlige associationseffekt.

Den måde man tager højde for indflydelse fra andre variabler til estimation af den gennemsnitlige kausale effekt, kan gøres ved at anvende specielle metoder. I dette projekt kiggede vi på to forskellige metoder kaldet IP vægtning og standardisering. IP vægtning er en parametrisk metode, som anvender variabler til at give et individ en specifik vægt baseret på disse variabler. Standardisering er en ikke-parametrisk metode, som udregnes ved at behandle et datasæt på en speciel måde og derefter udregne middelværdier. Man skal selv bestemme, hvilken metode man vil anvende til estimation, da det ikke er forkert at anvende den ene eller den anden.

Ud fra databehandlingen har rygestop en kausal effekt på ændring i BMI. Med IP vægtning estimerede vi den gennemsnitlige kausale effekt til at være 0.6598 med et 95% konfidensinterval på (0.3585, 0.961), hvilket betyder, at ens BMI stiger efter rygestop, hvilket vil sige, at man tager på i vægt. Standardisering gav et estimat på 0.486 med et 95% konfidensinterval på (0.23, 0.74), hvilket betyder, at standardisering også siger, at man tager på i vægt efter rygestop. Man skal dog ikke være helt sikre på disse resultater, da der kan være nogle krav, som ikke er opfyldt til at kunne stole på den kausale inferens. Da begge estimater er positive, går vi ud fra, at man tager på i vægt efter rygestop.

- [1] J. Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2:47–53, 1946.
- [2] D. R. Cox. *Planning of Experiments*. Wiley, New York, NY, 1958.
- [3] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2000.
- [4] J. Y. Halpern. Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research*, 12:317–337, 2000.
- [5] Ernst Hansen. *Kapitel 5 Asymptotisk Estimationsteori*, 2003. Tilgængelig på [http://web.math.ku.dk/~erhansen/stat1TS\\_03/doku/noter/kap5.pdf](http://web.math.ku.dk/~erhansen/stat1TS_03/doku/noter/kap5.pdf). Sidst set d. 16/3-2022.
- [6] P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81:968–970, 1986.
- [7] A. Nicholson, K. Korb, L. Hope and K. Axnick. Varieties of causal intervention. *Proceedings of the Pacific Rim Conference on Artificial Intelligence*, pages 322–331, 2004.
- [8] S. L. Lauritzen. *Graphical Models*. Oxford University Press, New York, NY, 1996.
- [9] J. Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9 (translated). *Statistical Science*, 5:255–270, 1991.
- [10] Jonas Peters, Dominik Janzing, og Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, Massachusetts : The MIT Press, 2017. ISBN 978-0-262-03731-9.
- [11] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, 1988.
- [12] J. Pearl. Belief networks revisited. *Artificial Intelligence*, 59:49–56, 1993.
- [13] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, 2nd edition, 2009.
- [14] H. Pishro-Nik. *Introduction to probability, statistics, and random processes*. Kappa Research LLC, 2014. Tilgængelig på <https://www.probabilitycourse.com>. Sidst set d. 16/3-2022.
- [15] T. S. Richardson and J. M. Robins. *Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality*. Working Paper Number 128, Center for Statistics and the Social Sciences, University of Washington, 2013.

- [16] J. M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods — applications to control of the healthy worker survivor effect. *Mathematical Modeling*, 7:1393–1512, 1986.
- [17] Miguel A. Hernán & James M. Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.
- [18] D. B. Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331, 2005.
- [19] I. Shpitser, T. J. VanderWeele, and J. M. Robins. On the validity of covariate adjustment for estimating causal effects. *Proceedings of the 26th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 527–536, 2010.
- [20] T. Verma and J. Pearl. Equivalence and synthesis of causal models. *In Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 255–270, 1991.