Kausaalipäättelyn uudet menetelmät

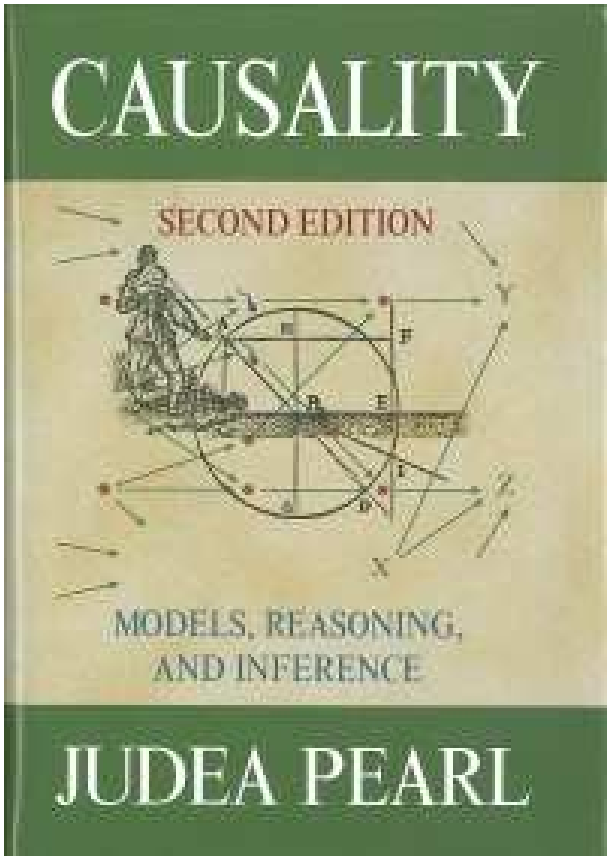
Juha Karvanen
Matematiikan ja tilastotieteen laitos
Jyväskylän yliopisto

UNIVERSITY OF JYVÄSKYLÄ

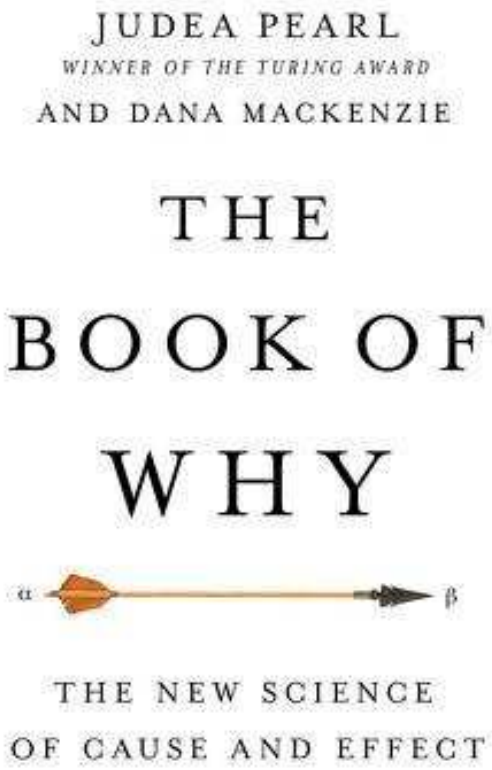
DEMO

Profiloitumisalue: Decision analytics utilizing causal models and multiobjective optimization (DEMO)

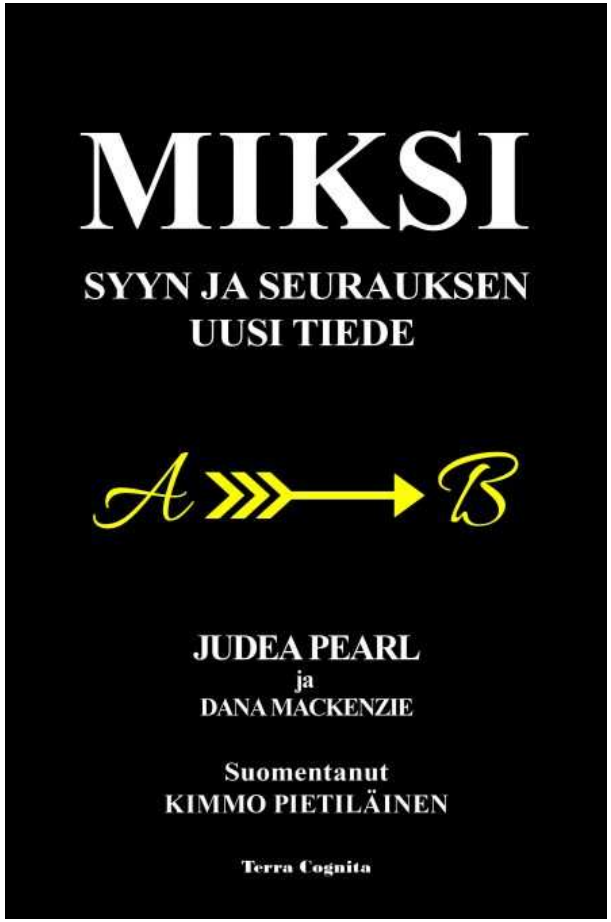
Kausaalisuus on tieteen ja päätöksenteon keskiössä.



2009



2018

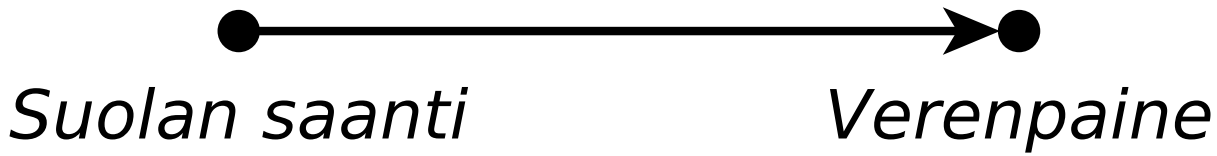


2018

Sisältö

1. Johdanto: kausaalisuuden portaat
2. Seitsemän teesiä kausaalisuudesta
3. Kausaalioletusten esittäminen graafeilla
4. Kausaalivaikutusten identifioituvuus
5. Työvälineitä identifioituvuuden selvittämiseen
6. Kausaalivaikutusten estimointi
7. Tutkimusasetelman esittäminen graafeilla
8. Meta-analyysi 3.0

Esimerkki



Kausaalisuuden portaat (Pearl: ladder of causation)

Porras	Yksilökysymys	Väestökysymys
Ennustaminen	Millä todennäköisyydellä sairastun verenpainetautiin seuraavan viiden vuoden aikana?	Kuinka moni tiettyyn väestöryhmään kuuluva sairastuu verenpainetautiin seuraavan viiden vuoden aikana?
Interventio	Millä todennäköisyydellä sairastun verenpainetautiin, jos päätän noudattaa vähäsuolaista ruokavaliota?	Kuinka moni tiettyyn väestöryhmään kuuluva sairastuu verenpainetautiin, jos kaikki päättävät noudattaa vähäsuolaista ruokavaliota?
Kontrafaktuaali	Olisinko välttänyt verenpainetaudin, jos olisin noudattanut vähäsuolaista ruokavaliota?	Kuinka monta tautitapausta väestöryhmässä olisi vältetty, jos kaikki olisivat noudattaneet vähäsuolaista ruokavaliota?

Kausaalipäättelyn perusongelma (Fundamental problem of causal inference): vastetta ei voi havaita (samalle henkilölle samaan aikaan) sekä ilman interventiota että sen kanssa.

Interventiot ja do-operaattori

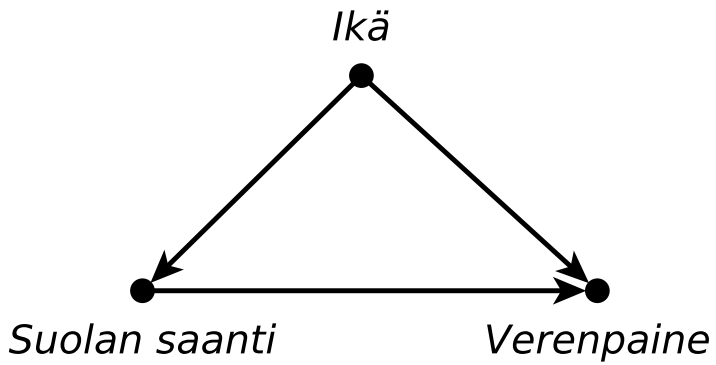
Millainen muuttujan Y jakauma on, kun muuttujaan X kohdistetaan interventio (toiminta)? Esimerkiksi kuinka verenpaine muuttuu, kun henkilö päättää vähentää suolan käyttöään.

Tämä voidaan kirjoittaa $P(Y = y \,|\, \mathrm{do}(X = x))$ tai lyhyemmin $P(Y \,|\, \mathrm{do}(X))$ tai $P(y \,|\, \mathrm{do}(x))$.

Yleisesti $P(Y \,|\, \mathrm{do}(X)) \neq P(Y \,|\, (X))$ eli intervention vaikutus eroaa ehdollisesta jakaumasta.

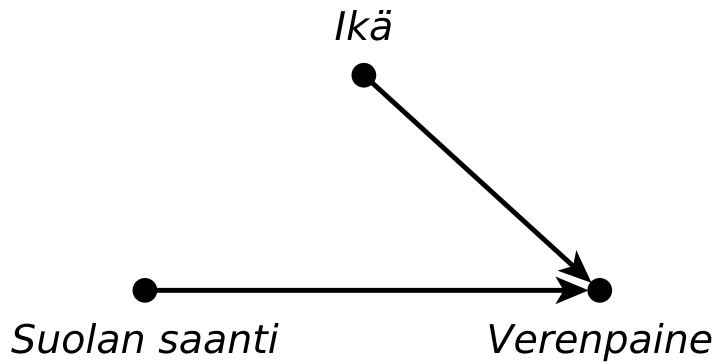
Sekoittavat tekijät (confounders)

Havainnoiva tutkimus



Ikä on sekoittava tekijä suolan saannin vaikutukselle verenpaineeseen.

Kokeellinen tutkimus



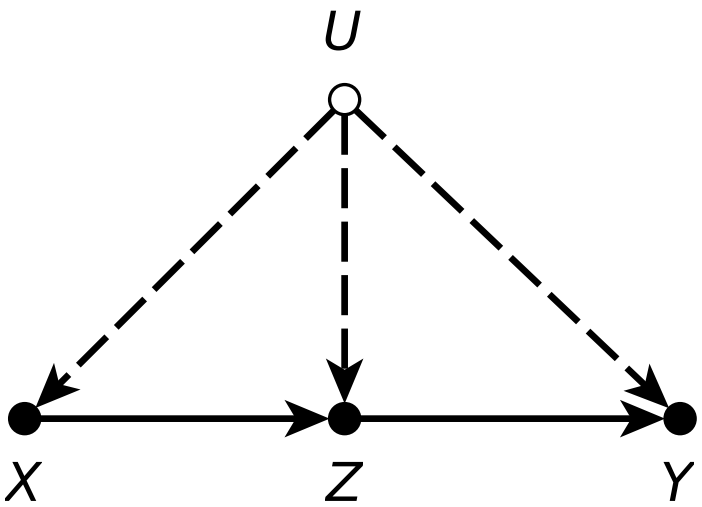
Suolan saantiin kohdistuva interventio katkaisee iän vaikutuksen suolan saantiin. Sekoittumista ei tapahdu.

Teesit

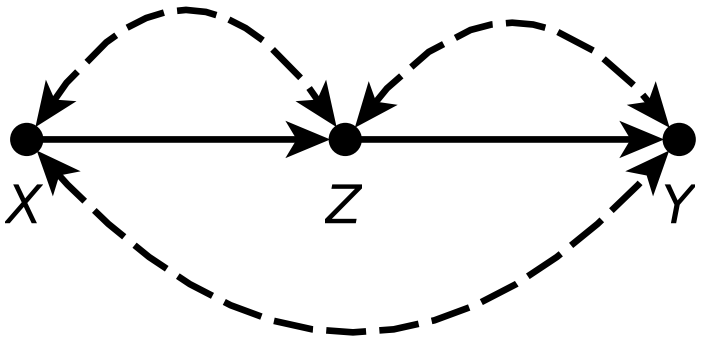
1. Kausaalipäättely nojautuu aina joihinkin oletuksiin.
2. Kausaalimallia ei voi rakentaa pelkästään datan perusteella vaan tarvitaan asiantuntemusta.
3. Oletukset on esitettävä eksplisiittisesti, jotta niitä voidaan arvioida.
4. Havaitsemattomia muuttujia koskevat oletukset ovat ratkaisevia silloin, kun kausaalipäättelyä tehdään havainnoivan tutkimuksen pohjalta.
5. Kausaalipäättelyyn on olemassa työvälineitä.
6. On tärkeää tietää, kuinka data on kerätty.
7. Näyttö (evidenssi) voi rakentua useasta erityyppisestä tutkimuksesta.

Kausaalioletusten esittäminen graafeilla

Havaitsemattomat muuttujat mukana graafissa



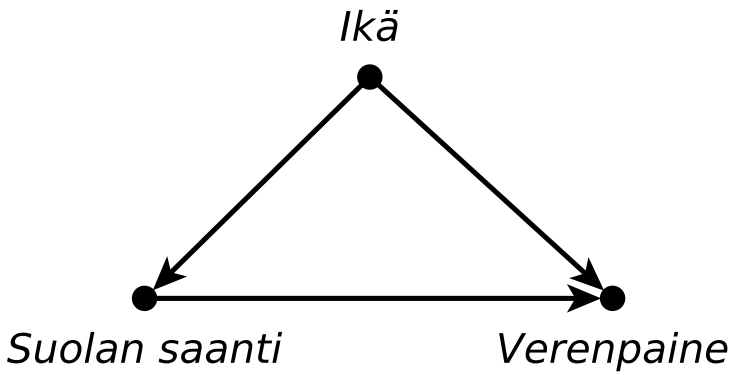
Semi-Markov-graafi: käytössä kaksisuuntaiset särmät



Kausaalivaikutusten identifioituvuus

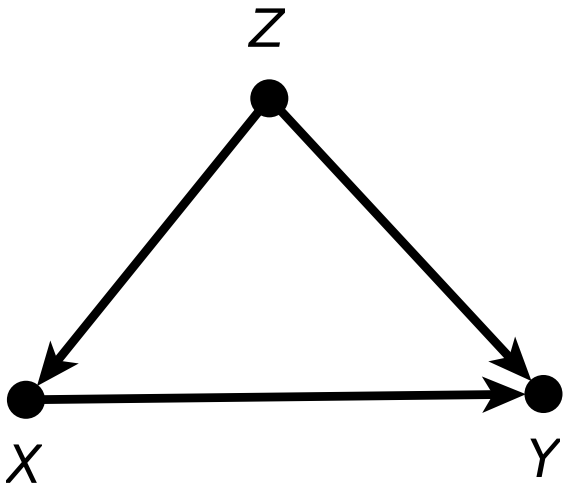
Kausaalivaikutus on identifioituva, jos se on mahdollista estimoida annetun kausaalimallin (kausaalirakenteen) ja käytössä olevien datalähteiden perusteella.

- Dataa ei itsessään ei tarvita identifioituvuuden määrittämiseen – tieto siitä mitä dataa on käytettävissä riittää.
- Esimerkki: Suolan kausaalivaikutus verenpaineeseen



  - identifioituva, jos suolan saanti, verenpaine ja ikä on mitattu havainnoivassa tutkimuksessa
  - ei identifioituva, jos ainoastaan suolan saanti ja verenpaine on mitattu havainnoivassa tutkimuksessa
  - identifioituva, jos verenpaine on mitattu kokeessa, jossa interventio kohdistuu suolan saantiin

Takaovikorjaus (back-door adjustment)



- Z ikä
- X suolan saanti
- Y verenpaine

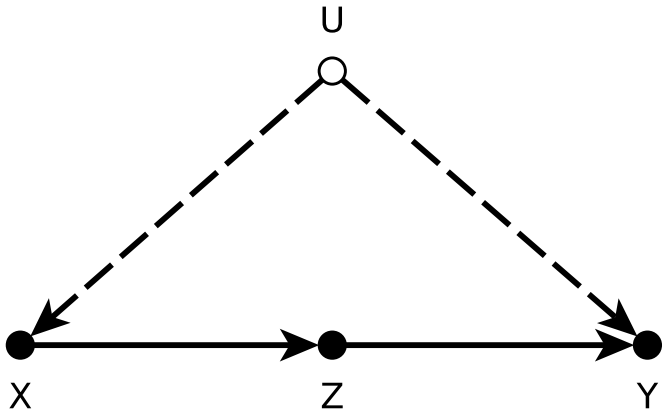
Voiko suolan kausaalivaikutuksen verenpaineeseen $P(Y \mid \mathrm{do}(X))$ identifioida havainnoista $P(X, Y, Z)$?

Do-laskenta (kausaalilaskenta) (Pearl 1995) antaa tuloksen

$$P(Y \mid \mathrm{do}(X)) = \sum_Z P(Y \mid X, Z) P(Z),$$

joka tunnetaan takaovikorjauksena.

Etuovikorjaus (front-door adjustment)

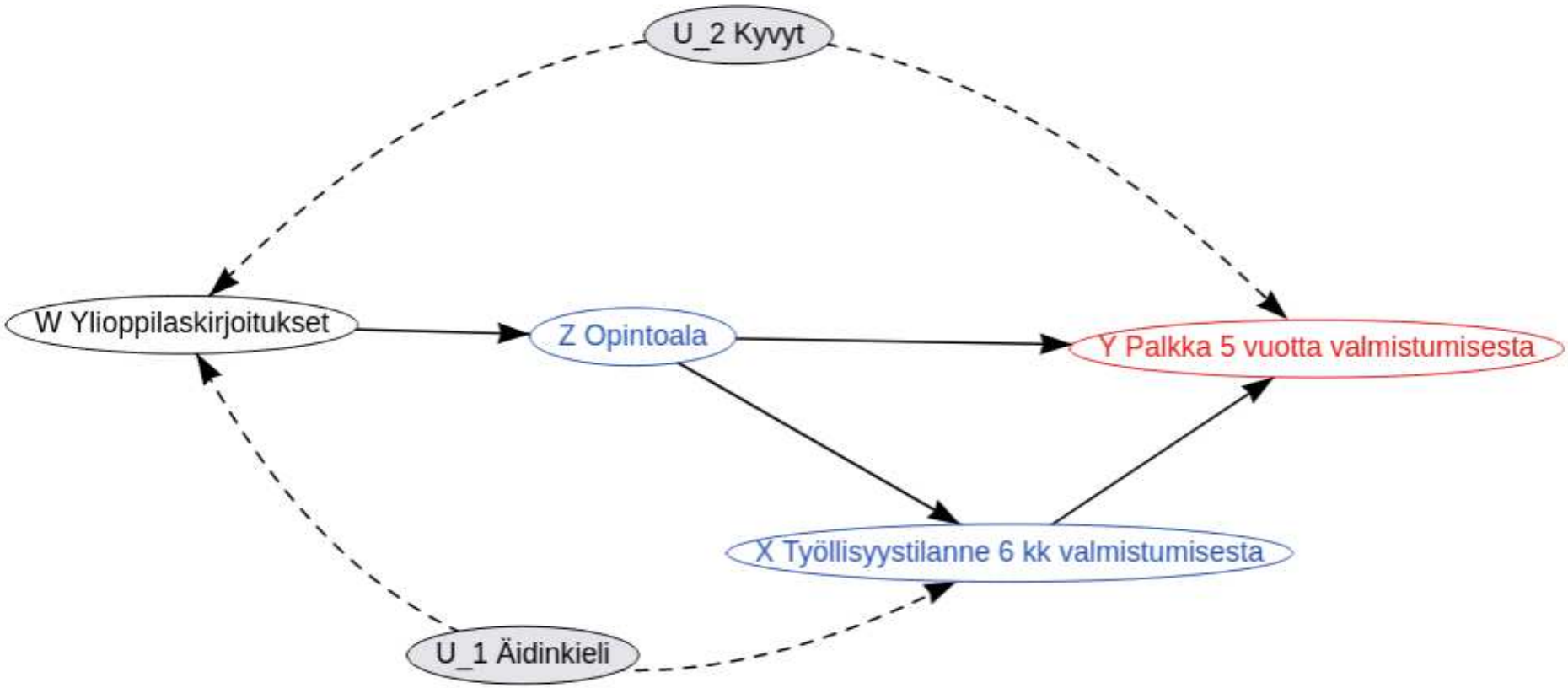


- ● Y verenpaine
- ● X vähäsuolaisten tuotteiden suosiminen
- ● Z suolan saanti
- ○ U havaitsemattomat sekoittavat tekijät

- Voiko vähäsuolaisten valintojen kausaalivaikutuksen verenpaineeseen selvittää havaintojen perusteella?
- Havaitsematon sekoittava tekijä U estää takaovikorjauksen käytön.
- Etuovikorjauksen (Pearl 1995) avulla saadaan tulos

$$P(Y \mid \mathrm{do}(X)) = \sum_{Z} P(Z \mid X) \sum_{X'} P(Y \mid Z, X') P(X').$$

Useita sekoittavia tekijöitä



$$P(Y \mid \mathrm{do}(X), \mathrm{do}(Z)) = \frac{\sum_W P(Y \mid X, Z, W) P(X \mid W, Z) P(W)}{\sum_{W,Y'} P(Y' \mid X, Z, W) P(X \mid W, Z) P(W)}$$

Työvälineitä kausaalivaikutusten identifiointiin

Do-laskenta (kausaalilaskenta) (Pearl 1995)

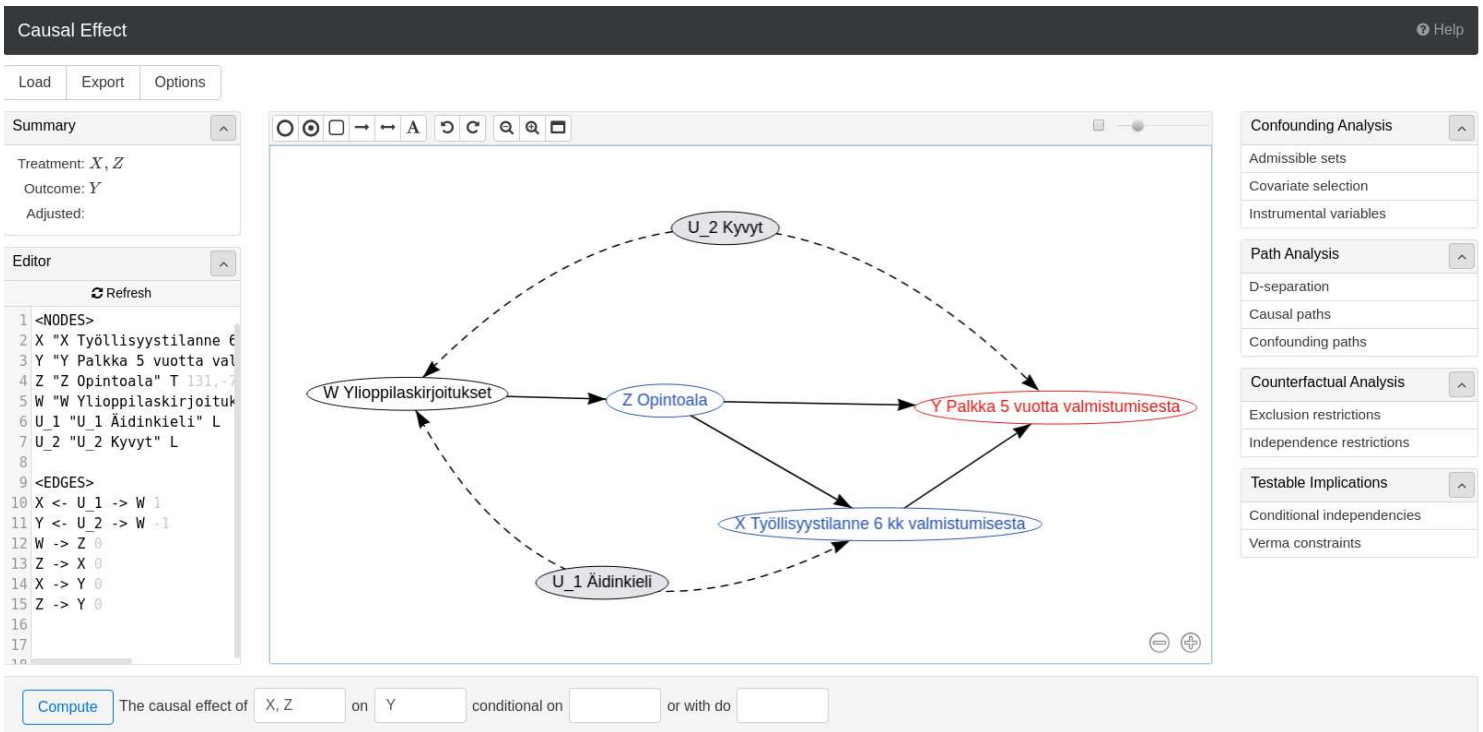
- Syöte: käytettävissä olevat jakaumat (symbolisessa muodossa)
- Kolme sääntöä jakaumien muokkaamiseen:
  1. Havaintojen lisääminen ja poistaminen
  2. Toiminnan ja havainnon vaihtaminen
  3. Toiminnan lisääminen ja poistaminen
- Tavoite: lauseke, joka ei sisällä do-operaattoreita (tai yleisemmin tuntemattomia osia).
- Heikkous: ei kerro missä järjestetyksessä sääntöjä (ja tavanomaista todennäköisyyslaskentaa) tulisi käyttää.

ID-algoritmi (Tian & Pearl 2002, Shpitser & Pearl 2006)

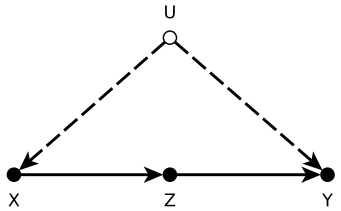
- Syöte: muuttujien yhteisjakauma (symbolisessa muodossa).
- Joko löytää rakenteen, joka estää identifioituvuuden tai palauttaa lausekkeen, joka ei sisällä do-operaatteita.
- Palauttaa aina oikean vastauksen (Huang and Valtorta 2006;Shpitser & Pearl 2006).
- Käyttökohteiltaan do-laskentaa suppeampi.

`causaleffect` R-paketti

- Santtu Tikan `causaleffect`-paketti sisältää ID-algoritmin ja muita kausaalipäättelyn algoritmeja.
- Avoimesti saatavilla: `https://cran.r-project.org/package=causaleffect`
- Graafinen käyttöliittymä Jyväskylän yliopiston sisäverkossa `http://shiny-test.maths.jyu.fi:3838/causaleffect/`



`causaleffect` R package: an example



- Y verenpaine
- X vähäsuolaisten tuotteiden suosiminen
- Z suolan saanti
- U havaitsemattomat sekoittavat tekijät

```
> library(causaleffect)
> library(igraph)

#Option 1: Reading the graph from a file
> fig1 <- parse.graphml("fig1.graphml", format = "standard")

#Option 2: Creating the graph in R
> fig1 <- graph.formula(X -+ Z, Z -+ Y, X -+ Y, Y -+ X, simplify = FALSE)
> fig1 <- set.edge.attribute(graph = fig1, name = "description",
                             index = c(3,4), value = "U")

#Calling causal.effect()
> ce1 <- causal.effect(y = "Y", x = "X", z = NULL, G = fig1, expr = TRUE)
> cat(ce1)

\sum_{Z}[P(Z\vert X)\sum_{X}[P(Y\vert X,Z)P(X)]]
```

Kausaalivaikutusten estimointi

Jos do-laskennan avulla päädytään takaovikorjaukseen

$$P(Y \mid \mathrm{do}(X)) = \sum_Z P(Y \mid X, Z) P(Z),$$

on jakaumat $P(Y \mid X, Z)$ ja $P(Z)$ on mallinnettava ja summattava muuttujan Z arvojen ylitse. Usein estimaattori voidaan esittää muodossa

$$\hat{P}(Y \mid \mathrm{do}(x)) = \frac{1}{n} \sum_{i=1}^{n} \hat{P}(Y \mid x, z_i),$$

missä n on otoskoko ja merkintä $\hat{P}$ viittaa jakauman estimaattiin. Muuttujan Z jakaumaa ei siis tarvitse estimoida, vaan sen estimaatti on aineisto itsessään. Oletetaan esimerkiksi, että sovitettu regressiomalli on muotoa

$$\mathrm{E}(Y_i \mid x_i, z_i) = \hat{a} x_i^2 + \hat{b}\sqrt{z_i} + \hat{c},$$

missä $\hat{a}$, $\hat{b}$ ja $\hat{c}$ ovat datasta estimoituja malliparametreja. Estimaatti odotetulle kausaalivaikutukselle saadaan laskemalla

$$\hat{\mathrm{E}}(Y \mid \mathrm{do}(x)) = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{a} x^2 + \hat{b}\sqrt{z_i} + \hat{c} \right).$$

Arvo x on siis vakio, joka määräytyy interventiosta.

Kausaalipäättelyn osa-alueet

Kausaalivaikutusten löytäminen (causal discovery)

- Pyrkii rakentamaan kausaalimallin havaitun datan pohjalta.
- Perustuu ehdollisten riippumattomisien testaamiseen tai oletuksiin lineaarista yhteyksistä ja normaalijakautuneista virhetermeistä.
- Toimii joissakin tapauksissa.

Kausaalivaikutusten identifiointi

- Pyrkii esittämään kausaalivaikutuksen havaittujen todennäköisyysjakaumien avulla.
- Yleinen (parametriton) tilanne on ratkaistu ja siihen olemassa algoritmeja (ID). Joitakin laajennuksia on esitetty.
- Parametristen mallien kohdalla tutkimus on keskittynyt lineaarisiin malleihin (esim. instrumenttimuuttujat).

Kausaalivaikutusten estimointi

- Pyrkii estimoimaan identifioituvan kausaalivaikutuksen datasta.
- Monia lähestymistapoja: parametriset mallit (erityisesti lineaariset rakenneyhtälömallit (SEM)), g-estimointi, propensiteettipainotus (propensity score weighting), targeted learning, jne.

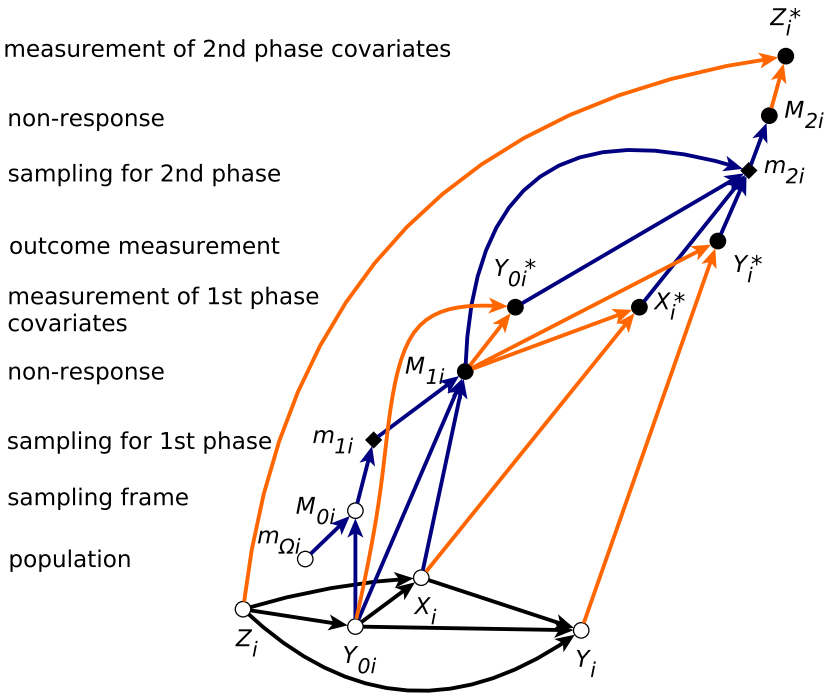
Tutkimusasetelman esittäminen graafeilla

On tärkeää tietää, kuinka data on kerätty!

Asetelmakausaalimallit

1. yhdistävät kausaalioletukset ja tutkimusasetelman ja avaavat tietä uusille identifioituvuusalgoritmeille
2. visualisoivat tutkimuksen keskeiset piirteet ja selkeyttävät kommunikointia

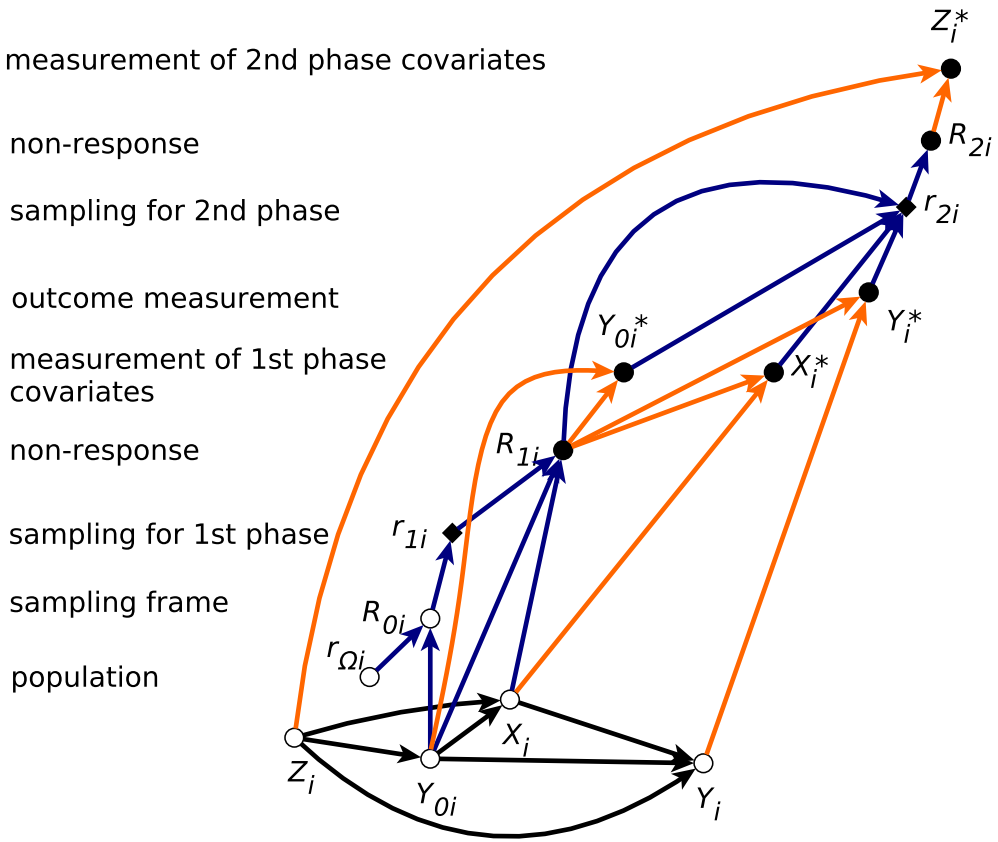
J. Karvanen (2015), Study design in causal models. *Scandinavian Journal of Statistics*, 42(2):361–377



● observed ○ not observed ◆ determined and known

- ○ Z_i genetic covariates
- ○ X_i classic covariates
- ○ Y_i time-to-event outcome
- ○ Y_{0i} outcome status at baseline
- ● Z_i^* measured genetic covariate
- ● X_i^* measured classic covariate
- ● Y_i^* measured time-to-event outcome
- ● Y_{0i}^* measured outcome status at baseline
- ○ $m_{\Omega i}$ population indicator
- ○ M_{0i} left truncation due to early death
- ◆ m_{1i} sampling indicator for the cohort
- ● M_{1i} non-participation for the cohort
- ◆ m_{2i} sampling indicator for the case-cohort set
- ● M_{2i} indicator for failure to measure genetic covariate

Asetelmakausaalimalli eräälle epidemiologiselle tutkimukselle



● observed ○ not observed ◆ determined and known

- ○ Z_i genetic covariates
- ○ X_i classic covariates
- ○ Y_i time-to-event outcome
- ○ Y_{0i} outcome status at baseline
- ● Z_i^* measured genetic covariate
- ● X_i^* measured classic covariate
- ● Y_i^* measured time-to-event outcome
- ● Y_{0i}^* measured outcome status at baseline
- ○ $r_{\Omega i}$ population indicator
- ○ R_{0i} left truncation due to early death
- ◆ r_{1i} sampling indicator for the cohort
- ● R_{1i} non-participation for the cohort
- ◆ r_{2i} sampling indicator for the case-cohort set
- ● R_{2i} indicator for failure to measure genetic covariate

Meta-analyysi 3.0

Erityyppisten tutkimusten yhdistäminen

Vaikutusketjujen koostaminen erilaisista osista

Meta-analyysi 2.0

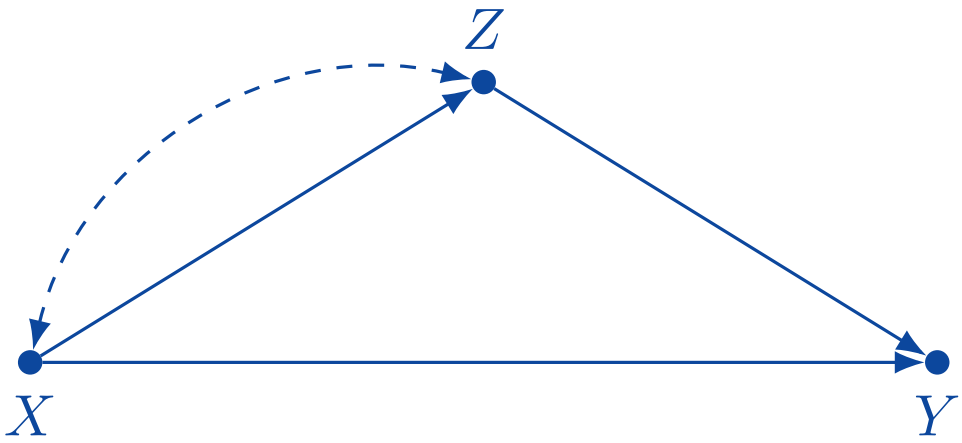
Tutkimusten välisten erojen mallintaminen:
satunnaisvaikutusmalli, metaregressio

Network meta-analysis:
käsittelyjen epäsuorat vertailut tutkimusten ylitse

Meta-analyysi 1.0

Samankaltaisten tutkimusten yhdistäminen

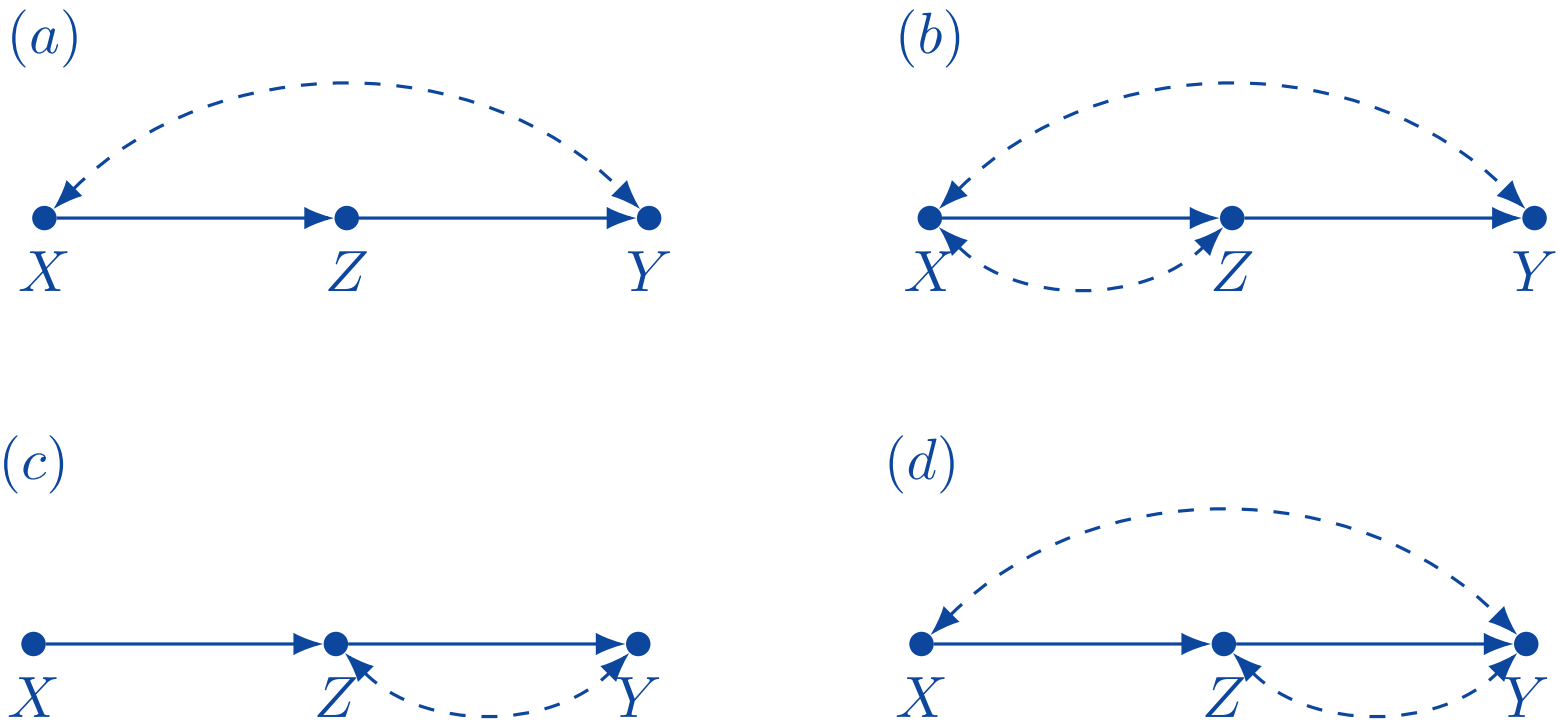
Kokeet ja havainnot



Kausaalivaikutus $P(Y \mid \mathrm{do}(X))$ ei ole identifioituva havaintojen $P(X, Y, Z)$ perusteella, mutta on identifioituva, kun sekä havainnot $P(X, Y, Z)$ että koe $P(Z \mid \mathrm{do}(X))$ ovat käytettävissä

$$P(Y \mid \mathrm{do}(X)) = \sum_Z P(Y \mid X, Z) P(Z \mid \mathrm{do}(X)).$$

Lisää kokeita ja havaintoja



Voiko kausaalivaikutuksen $P(Y \mid \mathrm{do}(X))$ identifioida näissä graafeissa, kun käytettävissä on

1. kyselytutkimus $P(X, Z, Y)$,
2. rajoitettu kyselytutkimus $P(X, Z)$ ja kokeellinen tutkimus $P(Y \mid \mathrm{do}(Z))$,tai
3. kokeelliset tutkimukset $P(Z \mid \mathrm{do}(X))$ ja $P(Y \mid \mathrm{do}(Z))$?

Do-search – työväline kausaalipäättelyyn, tutkimusten suunnitteluun ja meta-analyysiin

Syöte: graafi, tavoitejakauma ja datalähteet symbolisessa muodossa

Tulos: tavoitejakauman identifioituvuus ja lauseke

Toimitaperiaate: kattava haku do-laskennan ja todennäköisyyslaskennan sääntöjen ylitse

Rajoitukset: Muuttujien määrä ei voi olla suuri (noin 10).

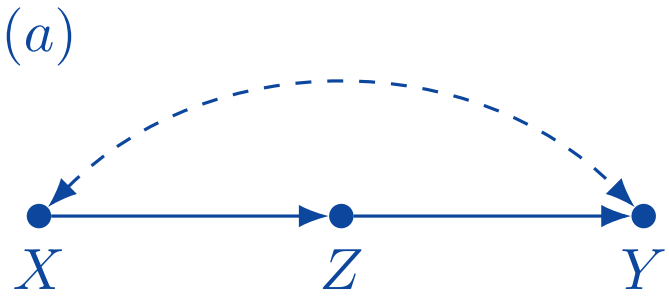
Lataa: `https://cran.r-project.org/package=dosearch`

Lue lisää: S. Tikka, A. Hyttinen, J. Karvanen (2019), Causal Effect Identification from Multiple Incomplete Data Sources: A General Search-based Approach, Submitted, arXiv:1902.01073.

Do-search: esimerkki

```
library(dosearch)

graafi <- "
    X -> Z
    Z -> Y
    Z -> Y
    X -- Y"

data <- "
    P(Y | do(Z))
    P(X,Z)"

tavoite <- "P(Y | do(X))"

dosearch(data, tavoite, graafi)
```

Do-search vastaa kysymyksiin (a)



Kausaalivaikutus $P(Y \mid \mathrm{do}(X))$ voidaan identifioida, kun datalähteet ovat

1. Kyselytutkimus $P(X, Z, Y)$,

$$P(Y \mid \mathrm{do}(X)) = \sum_Z P(Z \mid X) \sum_{X'} P(X')P(Y \mid X', Z)$$

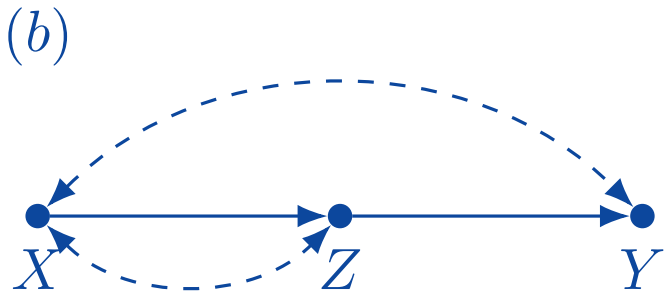
2. Rajoitettu kyselytutkimus $P(X, Z)$ ja koe $P(Y \mid \mathrm{do}(Z))$,

$$P(Y \mid \mathrm{do}(X)) = \sum_Z P(Z \mid X)P(Y \mid \mathrm{do}(Z))$$

3. Kokeiden ketju $P(Z \mid \mathrm{do}(X))$ ja $P(Y \mid \mathrm{do}(Z))$

$$P(Y \mid \mathrm{do}(X)) = \sum_Z P(Z \mid \mathrm{do}(X))P(Y \mid \mathrm{do}(Z)).$$

Do-search vastaa kysymyksiin (b)



1. Kyselytutkimus $P(X, Z, Y)$ ei riitä identifiointiin,
2. Rajoitettu kyselytutkimus $P(X, Z)$ ja koe $P(Y \mid \mathrm{do}(Z))$ eivät riitä identifiointiin,
3. Kokeiden ketju $P(Z \mid \mathrm{do}(X))$ ja $P(Y \mid \mathrm{do}(Z))$ riittää identifiointiin

$$P(Y \mid \mathrm{do}(X)) = \sum_{Z} P(Z \mid \mathrm{do}(X)) P(Y \mid \mathrm{do}(Z)).$$

Do-search vastaa kysymyksiin (c)

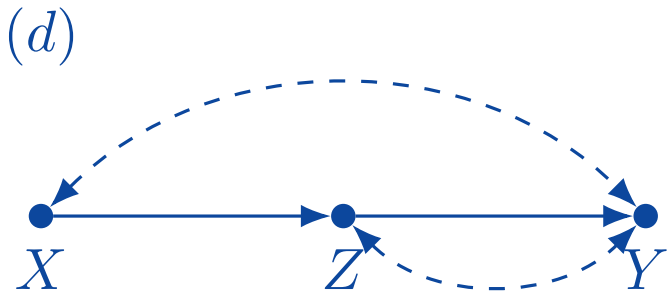
(c)

1. Kyselytutkimus $P(X, Z, Y)$ riittää identifiointiin

$$P(Y \mid \mathrm{do}(X)) = P(Y \mid X),$$

2. Rajoitettu kyselytutkimus $P(X, Z)$ ja koe $P(Y \mid \mathrm{do}(Z))$ eivät riitä identifiointiin,
3. Kokeiden ketju $P(Z \mid \mathrm{do}(X))$ ja $P(Y \mid \mathrm{do}(Z))$ ei riitä identifiointiin.

Do-search vastaa kysymyksiin (d)



$P(Y \mid \mathrm{do}(X))$ ei ole identifioituva

1. kyselytutkimuksen $P(X, Z, Y)$,
2. rajoitetun kyselytutkimuksen $P(X, Z)$ ja kokeen $P(Y \mid \mathrm{do}(Z))$,
3. kokeiden ketjun $P(Z \mid \mathrm{do}(X))$ ja $P(Y \mid \mathrm{do}(Z))$,

eikä minkään näiden datalähteiden yhdistelmän avulla.

Yhteenveto

- Kausaalivaikutusten estimointi vaatii datan ja asiantuntemuksen yhdistämistä.
- Pearlin lähestymistapa: ”Määrittele ensin, identifioi seuraavaksi, estimoi lopuksi”
  - Graafit soveltuvat kausaalirakenteen määrittelyyn.
  - Identifioituvuuden selvittämiseen on työvälineitä (`causaleffect`, `dosearch`).
  - Identifioituva kausaalivaikutus estimoidaan tilastotieteen menetelmillä.
- Meta-analyysi 3.0: kokeellisten ja havainnoivien tutkimusten yhdistäminen
  - Uusi tutkimussuunta

Kurssi ”Causal models” alkaa Jyväskylän yliopistossa 4.9.2019.