Tampereen yliopisto

Tuomas Piirainen

Graafit kausaalipäättelyssä

Informaatioteknologian ja viestinnän tiedekunta
Tilastollisen data-analyysin kandidaattitutkielma
Toukokuu 2021

Tiivistelmä

Tuomas Piirainen: Graafit kausaalipäättelyssä
Kandidaattitutkielma
Tampereen yliopisto
Matematiikan ja tilastollisen data-analyysin tutkinto-ohjelma
Toukokuu 2021

Tilastollista päättelyä tehdessä ollaan usein kiinnostuneita muuttujien välisistä suhteista. Muuttujien välistä riippuvuutta voidaan mitata monella tapaa, joista yksi yleisimmistä lienee korrelaatiokerroin. Ongelmaksi perinteisessä tilastollisessa päättelyssä muodostuu kuitenkin korrelaation erottaminen kausaliteetista eli syy-seuraussuhteesta. Perinteisen tilastollisen päättelyn avulla ei kyetä toteamaan tai mittaamaan kausaliteettia, joten avuksi tarvitaan kausaalipäättelyn työkaluja.

Tässä tutkielmassa tehdään kirjallisuuskatsaus graafeja hyödyntävään Judea Pearlin kausaaliteoriaan. Tutkielman keskiössä ovat Pearlin kausaaliteorian avulla tehtävä kausaalivaikutuksen löytäminen ja kausaalivaikutuksen estimointi. Graafit ovat olennainen osa tutkielmaa, sillä Pearlin kausaaliteorian työkalut perustuvat graafien avulla tehtäviin päätelmiin.

Tutkielmassa havainnollistetaan esimerkkien kautta, kuinka kausaalivaikutuksen löytämiseksi aineiston ja oletusten pohjalta rakennetaan kausaalimalli, jota sitten visualisoidaan graafin avulla. Kun kausaalimalli on rakennettu, tutkitaan graafin avulla muuttujien riippumattomuutta. Jos muuttujat todetaan riippumattomiksi, ei niiden välillä tällöin ole kausaliteettia. Jos muuttujat todetaankin riippuviksi, voidaan päättelyssä edetä kausaalivaikutuksen estimointiin.

Kausaaliteorian lisäksi tutkielmassa tehdään katsaus kausaalisuuden historiaan. Huomataan, että yksi kausaalipäättelyn tärkeimmistä askelista oli tieteen erotus filosofiasta. Kausaalisuus ei enää siis ollut vain abstrakti käsite, vaan tieteellisen lähestymistavan avulla kausaalisuudesta alettiin tehdä kokeellista tutkimusta.

Avainsanat: syy-seuraussuhde, riippumattomuus, korrelaatio, d-erottelu, interventio

Sisällys

1 Johdanto

Kausaliteetti eli syy-seuraussuhde on ilmiö, jossa tapahtuma (syy) aiheuttaa toisen tapahtuman (seuraus). Monilla tilastotieteen sovellusaloilla tutkimuksen kohteena ovat usein muuttujien väliset yhteydet. Useimmiten muuttujien välistä yhteyttä mitataan korrelaatiokertoimella, joka kertoo, kuinka paljon muuttujien välillä on riippuvuutta. Jos korrelaatiota huomataan olevan, ollaan usein myös kiinnostuneita mahdollisesta kausaliteetista muuttujien välillä.

Kausaliteetin selvittäminen voi kuitenkin aiheuttaa ongelmia niin satunnaiskokeissa kuin havainnoivassa tutkimuksessa. Esimerkiksi syöpäriskin ja tupakoinnin yhteyttä tutkiessa satunnaiskoe olisi epäeettinen, sillä ketään ei voida pakottaa tupakoimaan. Jos sama tutkimus päätettäsiin suorittaa havainnoivana tutkimuksena, voisi ongelmaksi muodostua korrelaation ja kausaliteetin erottaminen: syöpäriskin ja tupakoinnin välistä korrelaatiota voi mahdollisesti selittää jokin sekoittava (*confounding*) tekijä, jota tutkimuksessa ei ole huomioitu. Tällaisten ongelmien ratkaisuun voidaan käyttää graafien avulla tehtävää kausaalipäättelyä, jonka tarjoamat työkalut toimivat tutkimusmenetelmästä riippumatta. (Pearl, Glymour ja Jewell 2016, s. 53.)

Tässä tutkielmassa tehdään kirjallisuuskatsaus kausaalipäättelyyn, jonka erityisenä mielenkiinnon kohteena on Judea Pearlin kausaaliteoria. Pearlin kausaaliteoriaa on sanottu mullistavaksi tavaksi tutkia kausaalisuutta ja yksi Pearlin tavoitteista onkin ollut tuoda kausaalipäättely osaksi eri tieteenalojen ongelmanratkaisuprosessia. (Pearl 2009a, s. ii, xix.)

Toisessa luvussa käsitellään kausaalisuuden historiaa. Tarkoituksena on tuoda ilmi sitä, kuinka kausaalisuuden käsite on ajan saatossa muuttunut.

Kolmas luku esittää johdattelevan esimerkin kausaalipäättelyn tärkeydestä. Kappaleessa tutustutaan Simpsonin paradoksiin, joka on tuottanut tilastollisessa tutkimuksessa paljon ongelmia.

Neljäs luku aloitetaan käymällä läpi graafiteorian perusteet, josta sitten siirrytään kausaalimallin määrittelyyn. Luvun lopussa huomataan, kuinka kausaalimalli voidaan esittää graafina.

Viides luku keskittyy muuttujien välisten suhteiden tutkimiseen.

Kuudennessa luvussa esitellään graafien manipulointia ja kausaalivaikutuksen estimointia. Luvussa palataan takaisin Simpsonin paradoksin tuottamaan ongelmaan ja esitetään sille ratkaisu laskennallisen esimerkin avulla.

Lopuksi tehdään yhteenveto työn tärkeimmistä tuloksista.

2 Kausaalisuus ja tieteen historia

Kausaalisuuden tieteellinen historia alkaa antiikin Kreikasta ja erityisesti filosofi Aristoteleen (384–322 eaa.) teoksesta *Metafysiikka*, jossa hän määritteli kausaalisuuden neljä eri olemusta: materiaalinen, formaalinen, aikaansaava ja päämääräsyy. Näillä kategorioilla Aristoteles pyrki luokittelemaan kaikki mahdolliset syyt sille, miksi jotain tapahtuu tai miksi jokin asia on olemassa. (Hopkins 2004, s. 2–3.) Klassinen esimerkki Aristoteleen neljästä syystä on päämääräsyy kiven putoamiselle maahan: kivi putoaa maahan, koska sen tarkoitus on olla maassa.

Aristoteleen jalanjäljissä Francis Bacon (1561–1626) ehdotti, että kausaalisuutta olisi mahdollista tutkia empiirisesti. Jos A aiheuttaa B:n, silloin näiden tapahtumien välillä on looginen suhde, joka voidaan löytää selvittämällä A:n ja B:n olemus. Tämä lähestymistapa johti Baconin suurimpaan saavutukseen: tieteen erottamisen filosofiasta. (Hopkins 2004, s. 3.)

Vuonna 1638 Galileo Galilei (1564–1642) julkaisi teoksen *Kaksi uutta tiedettä*, joka tuli muuttamaan tiedettä merkittävästi ja samalla synnyttämään lisää eriäviä mielipiteitä. Galilein ensimmäinen sääntö kuvasi sitä, kuinka kuvaus tapahtumasta tehdään ensin ja selitys jälkeen. Tämä sääntö vastasi siis kysymyksiin *miten* ja *miksi*. Toinen sääntö koski tapahtuman kuvauksen esittämistä matematiikan kielellä, etenkin yhtälöillä. Galilein ja muiden aikalaistensa töiden kautta tiede siis muuttui merkittävästi. Esimerkiksi tiedonhankinnassa kokeiden suorittaminen yleistyi, Aristoteleen neljän syyn oppi unohdettiin ja matematiikka nousi ensijaiseksi tavaksi kuvata ilmiöitä. Näitä muutoksia eivät etenkään filosofit hyväksyneet, sillä esimerkiksi René Descartes (1596–1650) ajatteli, että Jumala on kaiken takana. Tästä johtuen esimerkiksi kausaalisuuden kysymyksiin tavallinen ihminen ei kykenisi vastaamaan. (Pearl 2009a, s. 404–406.)

Noin sata vuotta myöhemmin skotlantilainen filosofi David Hume (1711–1776) haastoi vallitsevan teorian kausaalisuuden tulkinnasta. Erityisesti hän kyseenalaisti Galilein ensimmäisen säännön ja väitti, että kausaaliset assosiaatiot ovat subjektiivisia ja rinnastettavissa optisiin illuusioihin. Hume perusteli väitettään sillä, että kukaan ei todellisuudessa voi havaita kausaliteetin olemassaoloa, vaan sen olemassaolo kuvitellaan tapauksissa, joissa toinen tapahtuma seuraa toista. (Pearl 2009a, s. 406; Hopkins 2004, s. 3.)

Kiivas keskustelu kausaalisuudesta ei jäänyt huomiotta myöskään tilastotieteen puolella. Oli kuitenkin sattumaa, miten kausaalisuuden tutkiminen sai alkunsa. Vuonna 1888 Sir Francis Galton (1822–1911) mittasi ihmisten kyynärvarren pituutta sekä pään kokoa. Hän yritti tutkia, voisiko toisella suureista ennustaa toista. Galton huomasi, että jos suureista saadut pisteparit sovitetaan xy-koordinaatistoon, omaa pisteisiin parhaiten sopiva suora mielenkiintoisia matemaattisia ominaisuuksia. Tästä havainnosta syntyi muun muassa korrelaation käsite.

Kausaalisuuden tutkiminen ei kuitenkaan edennyt korrelaation käsitettä pidemmälle, sillä Galtonin oppipoika Karl Pearson (1857–1936) piti koko kausaalisuuden käsitettä turhana. Vuoden 1911 teoksessaan Pearson esitti, että kaikki tarvittava tieto löytyi kontingenssitauluista. Pearson ei tulevissa teoksissaan maininnut kausaalisuu-

desta sanallakaan.

Kului 25 vuotta ennen kuin kausaalipäättely otti askeleen eteenpäin, jolloin suureksi osaksi Sir Ronald Fisherin (1890–1962) ansiosta satunnaiskokeet yleistyivät. Satunnaiskoetta pidettiinkin pitkään ainoana hyväksyttynä tapana tutkia kausaalisuutta. (Pearl 2009a, s. 409–410.)

3 Kausaalipäättelyn tärkeys: Simpsonin paradoksi

Edward Simpsonin (1922–2019) mukaan nimetty Simpsonin paradoksi on ilmiö, jossa koko aineistolle huomataan pätevän tietty assosiaatio, mutta tämä assosiaatio ei päde, kun aineisto jaetaan osiin. Pearl, Glymour ja Jewell (2016) kuvaavat kausaalipäättelyn tärkeyttä klassisen esimerkin kautta, jonka Simpson esitti jo vuonna 1951.

Joukolle sairaita potilaita annettiin mahdollisuus kokeilla uutta lääkettä. Lääkkeen ottaneiden joukossa oli vähemmän parantuneita kuin joukossa, jossa lääkettä ei otettu. Näyttäisi siis siltä, että lääkkeellä ei ole parantavaa vaikutusta. Kuitenkin, jos potilaat jaetaan sukupuolen mukaan, huomataan, että suurempi osa lääkkeen ottaneista miehistä parantuu verrattuna miehiin, jotka eivät ottaneet lääkettä. Sama ilmiö huomataan naisilla.

Saavutaan siis ristiriitaan, jossa lääkkeellä ei ole parantavaa vaikutusta, kun potilaita tutkitaan yhtenä ryhmänä, mutta parantava vaikutus kuitenkin huomataan erikseen miehillä ja naisilla. Jotta ristiriita saadaan ratkaistua, tarvitaan lisätietoa aineiston taustalla vaikuttavasta kausaalirakenteesta. (Pearl, Glymour ja Jewell 2016, s. 1–5.) Simpsonin paradoksi voidaan tiivistää seuraavasti: Kahden muuttujan välinen korrelaatio voi kääntyä *päinvastaiseksi*, kun analyysiin lisätään faktoreita. (Pearl 2009a, s. 424.)

Esimerkki 3.1. Tutkitaan Pearlin, Glymourin ja Jewellin esimerkkiaineistoa (2016, s. 2), jossa 700 potilasta jaetaan kahteen 350 hengen ryhmään. Molemmat ryhmät jaetaan lisäksi potilaan sukupuolen mukaan miehiin ja naisiin. Toiselle ryhmälle annetaan lääkettä ja toiselle taas ei. Kun potilaan sukupuolta ei oteta huomioon, tutkimustulokset osoittavat, että lääkkeen ottaneita potilaita parantui vähemmän (78 %) kuin potilaita, ketkä eivät ottaneet lääkettä (83 %). Kun sukupuoli otetaan huomioon, huomataan lääkkeen ottaneita parantuvan enemmän niin miehissä (93 % vastaan 87 %) kuin naisissa (73 % vastaan 69 %). Pelkän aineiston pohjalta ei siis voida tehdä johtopäätöstä lääkkeen toimivuudesta.

Taulukko 3.1. Lääkkeen parantamisasteen tulokset potilaan sukupuoli huomioituna.

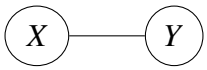
	Otti lääkettä	Ei ottanut lääkettä
Mies	81/87 parantui (93 %)	234/270 parantui (87 %)
Nainen	192/263 parantui (73 %)	55/80 parantui (69 %)
Yhteensä	273/350 parantui (78 %)	289/350 parantui (83 %)

4 Graafien ja kausaalimallien yhteys

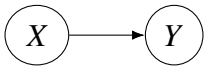
4.1 Graafiteorian peruskäsitteitä

Simpsonin paradoksin kaltaisissa ongelmatilanteissa päätöksentekoa ei voida enää perustaa pelkän aineiston varaan. Kausaalisuuden kysymyksiin tarvitaan formalisointia, joka onnistuu Pearlin kausaaliteorian avulla. Koska Pealin kausaaliteoria pohjautuu vahvasti graafeihin, käydään seuraavaksi läpi graafien yleisiä ominaisuuksia.

Graafi koostuu solmujen joukosta V ja solmuja yhdistävistä kaarien joukosta E. Jokainen graafin solmu voi olla joko suunnattu tai suuntaamaton. Suunnatun kaaren tunnistaa sen päässä olevasta nuolesta, kun taas suuntaamattomalla kaarella ei puolestaan ole nuolta. Solmua, josta suunnattu kaari alkaa kutsutaan vanhemmaksi. Vastaavasti vanhemman lapsi on solmu, johon suunnattu kaari osoittaa. (Pearl, Glymour ja Jewell 2016, s. 24–25; Pearl 2009a, s. 12–13.)



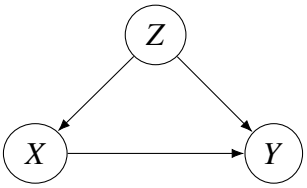
Kuva 4.1. Suuntaamaton graafi.



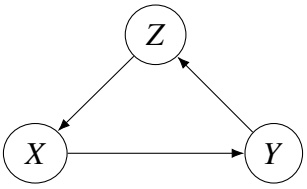
Kuva 4.2. Suunnattu graafi, jossa X on Y:n vanhempi.

Polku on jono peräkkäisiä kaaria, joita pitkin graafissa päästään kulkemaan solmusta toiseen. Jos polun kaikki kaaret ovat samansuuntaisia, on polku tällöin suunnattu. Jos yhdestäkin solmusta lähtevä kaari tai suunnattu polku osoittaa takaisin itseensä, on kyseessä syklinen graafi. (Pearl, Glymour ja Jewell 2016, s. 25.)

Huomautus. Yleensä suunnatussa graafissa kaaria voi kulkea vain niiden suunnan mukaisesti (Koivisto ja Niemistö 2018, s. 31). Tässä tutkielmassa tehdään poikkeus tähän sääntöön ja kaaria voidaan kulkea myös niiden vastaiseen suuntaan (Pearl 2009a, s. 16).



Kuva 4.3. Graafi on asyklinen, koska se ei sisällä syklejä.



Kuva 4.4. Graafi on syklinen, koska esimerkiksi solmusta Z lähtevää polkua voidaan kulkea takaisin solmuun Z.

4.2 Kausaalimallin määritelmä

Jotta kausaalipäättelyn kysymyksiin voidaan vastata täsmällisesti, tarvitsee aineiston ja mahdollisten oletusten pohjalta rakentaa *kausaalimalli*. Kausaalimallia ei kuitenkaan voida perustaa esimerkiksi logiikan lauseiden tai todennäköisyyslaskennan varaan. Esimerkiksi kysymyksen ”Jos otan lääkkeen, parantaako se minut?” kausaliteettia ei logiikan tai todennäköisyyslaskennan avulla voida todeta taikka mitata. (Pearl, Glymour ja Jewell 2016, s. 26; Pearl 2009a, s. 202–203.)

Pearlin (2009b, s. 102) mukaan minkä tahansa kausaaliteorian pitäisi pystyä (1) esittämään kausaalisuuden kysymyksiä matemaattisella kielellä, (2) tarjoamaan tarkka kieli oletusten ilmaisemiselle, joita kysymysten vastaamiseen tarvitaan, (3) tarjoamaan systemaattinen tapa vastata ainakin osaan kysymyksistä ja merkitsemään loput kysymyksistä ”vastaamattomiksi” sekä (4) tarjoamaan menetelmä, jolla päättää, mitä oletuksia tai uusia mittauksia tarvittaisiin vastaamaan ”vastaamattomiin” kysymyksiin.

”Yleisen kausaaliteorian” pitäisi kuitenkin tarjota enemmän. Sen lisäksi, että kaikki kysymykset tulkittaisiin kausaalisiksi, pitäisi yleisen teorian sisällyttää mikä tahansa muu teoria, joka on todettu hyödylliseksi kausaalisuutta tutkittaessa. Vaihtoehtoisten teorioiden pitäisi siis kehittyä yleisen teorian erikoistapauksina, kun rajoituksia asetetaan mallin, oletusten tai kielen suhteen. Tästä syystä Pearlin kausaalimalli yhdistääkin kausaalipäättelyssä paljolti käytettyjen rakenneyhtälömallien (*structural equation model, SEM*) ja potentiaalisten lopputulosten mallin (*potential outcomes model, POM*) periaatteita (Pearl 2009b, s. 102.).

Koska SEM ja POM eivät ole tämän tutkielman keskiössä, sivuutetaan niiden lähempi tarkastelu.

Määritelmä 4.1. *Kausaalimalli* (*structural causal model, SCM*) on *kolmikko* $M = \langle U, V, F \rangle$, jossa

1. U on taustamuuttujien joukko, joita kutsutaan myös eksogeenisiksi muuttujiksi. Eksogeeniset muuttujat määrittyvät mallin ulkopuolella.

2. V on muuttujien joukko $\{V_1, V_2, \ldots, V_n\}$, joita kutsutaan myös endogeenisiksi muuttujiksi. Endogeeniset muuttujat määrittyvät mallin muuttujien mukaan eli joukon $U \cup V$ alkioista.

3. F on funktioiden joukko $\{f_1, f_2, \ldots, f_n\}$, jossa jokainen f_i on kuvaus joukolta $U_i \cup PA_i$ (*parent, vanhempi*) joukolle V_i, missä $U_i \subseteq U$ ja $PA_i \subseteq V \setminus V_i$.

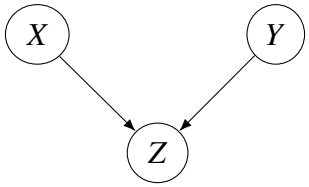
Tällöin joukko F muodostaa kuvauksen joukolta U joukolle V ja f_i antaa arvon $v_i = f_i(pa_i, u_i)$ muuttujalle V_i, kun $i = 1, \ldots, n$. (Pearl 2009a, s. 203.)

Mallin jokainen endogeeninen muuttuja on vähintään yhden eksogeenisen muuttujan jälkeläinen. Eksogeeniset muuttujat esitetään graafeissa *juurisolmuina*, sillä ne eivät voi olla muiden muuttujien jälkeläisiä. Jos jokaisen eksogeenisen muuttujan arvo tiedetään, tällöin joukon F funktioita käyttämällä voidaan täydellä varmuudella määrittää jokaisen endogeenisen muuttujan arvo.

Kausaalimalli voidaan nyt esittää suunnattuna asyklisena graafina (*directed acyclic graph, DAG*), jossa graafin solmut vastaavat joukkojen U ja V muuttujia sekä graafin kaaret vastaavat joukon F funktioita eli solmujen välisiä kausaalisuhteita. Tästä seuraa, että kausaliteetille voidaan antaa kuvallinen määritelmä:

Määritelmä 4.2. Jos graafissa solmu Y on solmun X lapsi, on X silloin välitön Y:n syy (Pearl, Glymour ja Jewell 2016, s. 26–28).

Esimerkki 4.1. Tutkitaan Pearlin, Glymourin ja Jewellin esimerkkiä (2016, s. 27) koulutuksen X ja kokemuksen Y vaikutuksesta palkkaan Z. Olkoon joukko $U = \{X, Y\}$, $V = \{Z\}$ ja $F = \{f_1 : Z = 2X + 3Y\}$. Nyt määritelmän 4.1 mukaan on määritelty kausaalimalli, joka voidaan esittää graafina. Tällaisen kausaalimallin määrittelystä kävisi ilmi, että koulutus ja kokemus ovat eksogeenisiä muuttujia, sillä niiden taustoja ei pyritä selvittämään, vaan ollaan ainoastaan kiinnostuneita niiden vaikutuksesta palkkaan. Siksi palkka on mallin endogeeninen muuttuja. Joukkoon F puolestaan kuuluu yhtälö, joka kuvastaa sitä, miten koulutuksen ja kokemuksen perusteella palkkaa maksetaan. Olettaen, että kuvaus f_1 on tosi, voidaan määritelmän 4.2 perusteella voidaan sanoa, että X ja Y ovat Z:n syitä.



Kuva 4.5. Graafi, jossa esitetään koulutuksen ja kokemuksen vaikutus palkkaan.

5 Muuttujien riippuvuussuhteen selvittäminen graafien avulla

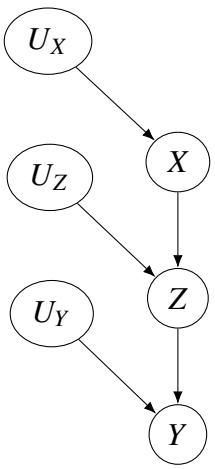
Kausaalimallien avulla saadaan selitettyä aineiston takana piilevää *mekanismia*, joka vaikuttaa aineistosta saatuihin tuloksiin. Kun kausaalimalli esitetään graafina, saadaan jo pelkän visuaalisen esityksen avulla ennustettua mahdollisia riippumattomuuksia muuttujien välillä. Riippumattomuuksia tutkimalla saadaan uutta tietoa muuttujien välisestä korrelaatiosta ja myöhemmin graafeja manipuloimalla voidaan korrelaatio erottaa kausaalisuudesta. (Pearl, Glymour ja Jewell 2016, s. 35.)

5.1 Ketjut, haarukat ja törmäykset graafeissa

Kuvitellaan tilanne, jossa kausaalimallia kuvaava graafi koostuu kolmesta solmusta ja kahdesta kaaresta. Koska kausaalimalli on asyklinen suunnattu graafi, seuraa siitä, että kolmen muuttujan tilanteessa muuttujien väliset kaaret voivat olla kolmessa eri järjestyksessä. Näitä eri järjestyksiä kutsutaan *ketjuksi* (*chain*), *haarukaksi* (*fork*) ja *törmäykseksi* (*collider*). Ketjussa ($A \rightarrow B \rightarrow C$) graafin molemmat kaaret kulkevat samansuuntaisesti, eli keskimmäiseen muuttujaan osoittaa ja siitä myöskin lähtee kaari. Haarukassa ($A \leftarrow B \rightarrow C$) graafin yhdestä muuttujasta lähtee kaksi kaarta kahteen eri muuttujaan. Törmäyksessä ($A \rightarrow B \leftarrow C$) kahden muuttujan kaaret osoittavat kolmanteen muuttujaan.

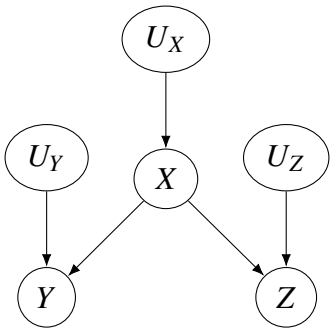
Edellä mainittuja graafeja katsomalla voidaan tutkia muuttujien välisiä riippuvuus- tai riippumattomuussuhteita. Ketjuille, haarukoille ja törmäyksille pätevät seuraavat säännöt:

1. Ehdollinen riippumattomuus ketjuissa: Muuttujat X ja Y ovat riippumattomia ehdolla Z, jos X:n ja Y:n välillä on täsmälleen yksi yksisuuntainen polku ja Z on mikä tahansa on muuttujajoukko, joka katkaisee tämän polun.



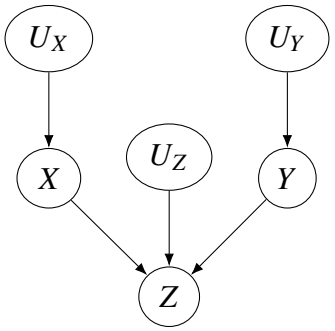
Kuva 5.1. Graafi, jonka solmut muodostavat ketjun.

2. Ehdollinen riippumattomuus haarukoissa: Jos muuttuja X on muuttujien Y ja Z yhteinen syy sekä Y:n ja Z:n välillä on täsmälleen yksi polku, ovat Y ja Z riippumattomia ehdolla X.



Kuva 5.2. Graafi, jonka solmut muodostavat haarukan.

3. Ehdollinen riippumattomuus törmäyksissä: Jos muuttuja Z on törmäyssolmu muuttujien X ja Y välillä sekä X:n ja Y:n välillä on täsmälleen yksi polku, ovat X ja Y riippumattomia, mutta riippuvia ehdolla Z tai millä tahansa Z:n jälkeläisellä.



Kuva 5.3. Graafi, jonka solmut muodostavat törmäyksen.

(Pearl, Glymour ja Jewell 2016, s. 35–44.)

5.2 d-erottelu polkujen tutkimisessa

Kausaalimallit ovat harvoin yhtä yksinkertaisia kuin esimerkiksi edellisessä kappaleessa esitetyt mallit. On hyvin yleistä, että kahden muuttujan välillä on useita kaaria tai polkuja ja jokin polku voi kulkea esimerkiksi ketjun, haarukan tai törmäyksen läpi. Tämä luonnollisesti vaikeuttaa riippuvuussuhteen tutkimista moninkertaisesti. (Pearl, Glymour ja Jewell 2016, s. 46.)

Monimutkaisiin graafeihin voidaan käyttää *d-erottelua* (*directional separation*), jonka avulla pystytään selvittämään, mikä riippuvuussuhde kahden muuttujan välillä vallitsee. *D*-erottelulla voidaan myös testata graafia. Jos *d*-erottelu tuottaa oletusten vastaisten tuloksen, on syytä pohtia, rakennettiinko graafi aineiston pohjalta oikein

vai kertooko graafi mahdollisesti uutta tietoa aineistosta (Pearl, Glymour ja Jewell 2016, s. 48-50).

Olkoon muuttujat X ja Y sekä muuttujajoukon Z alkiot graafin G solmuja. Muuttujien X ja Y sanotaan olevan *d-erotettuja* ehdolla Z, jos Z estää kaikki X:n ja Y:n väliset polut. Polun estämistä voidaan luonnehtia tiedonkulun tai riippuvuuden katkeamisena kahden solmun välillä. Jos polulla yksikin *kaari* on estetty, ei tieto tällöin pääse kulkemaan sen läpi, josta seuraa, että koko polusta tulee estetty. Tällöin X ja Y ovat riippumattomia ehdolla Z. Vastaavasti, jos yksikin *polku* X:n ja Y:n välillä on estämätön, ovat X ja Y *d-yhdistettyjä*. Estämätön polku ei siis sisällä yhtäkään joukon Z alkiota. Tällöin X ja Y toisistaan riippuvia.

Määritelmä 5.1. Polku p on *d-erotettu* solmujen joukolla Z, jos ja vain jos

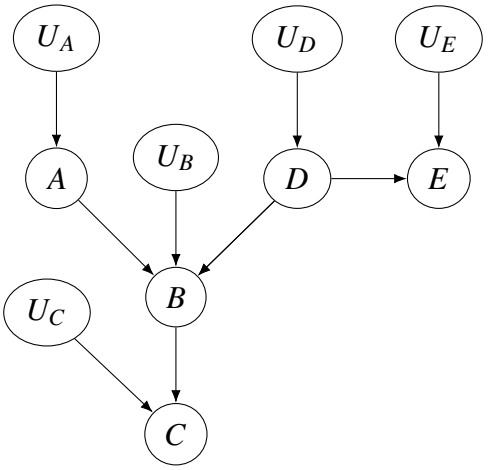
1. p sisältää ketjun $A \rightarrow B \rightarrow C$ tai haarukan $A \leftarrow B \rightarrow C$ siten että keskimmäinen solmu $B \in Z$ tai

2. p sisältää törmäyksen $A \rightarrow B \leftarrow C$ siten että törmäyssolmu B tai mikään sen jälkeläisistä ei kuulu joukkoon Z.

(Pearl, Glymour ja Jewell 2016, s. 46; Pearl 2009a, s. 16–17.)

Kuten edellä mainittiin, polun estämistä eli muuttujalla ehdollistamista voidaan luonnehtia riippuvuuden katkaisemisena kahden muuttujan väliltä. Jos esimerkiksi ketju $A \rightarrow B \rightarrow C$ ehdollistetaan muuttujalla B, tarkoittaa se, että muuttujan B arvo tiedetään. (Pearl, Glymour ja Jewell 2016, s. 46; Pearl 2009a, s. 16–17.) Tällöin muuttujalla A ei ole enää vaikutusta muuttujaan C, koska A ei enää muuta B:n arvoa, josta seuraa, että ainoastaan B muuttaa C:n arvoa. Muuttujien A ja C välinen polku on siis estetty.

Esimerkki 5.1. Sovelletaan nyt määritelmää 5.1 Pearlin, Glymourin ja Jewellin (2016, s. 47) esimerkkigraafiin, joka on kuvattu kuvassa 5.4. Valitaan tutkimuksen kohteeksi muuttujat A ja E. Polku muuttujasta A muuttujaan E sisältää törmäyksen ja haarukan. Jos joukko Z on tyhjä joukko, määritelmän 5.1 kohdan 2 mukaan A ja E *d-erotettuja*, koska muuttuja B estää polun. Jos B tai sen jälkeläinen C kuuluisi joukkoon Z, olisivat A ja E *d-yhdistettyjä*, koska nyt määritelmän 5.1 kohtaa 2 rikotaan ja kohdan 1 mukaisesti haarukkaa $B \leftarrow D \rightarrow E$ ei ole estetty.



Kuva 5.4. Graafi, joka sisältää törmäyksen ja haarukan.

6 Graafien manipulointi ja kausaalivaikutuksen estimointi

6.1 Interventiot ja *do*-laskenta

Interventiot ovat olennainen osa kausaalivaikutuksen tutkimista. Kun muuttujan X vaikutusta halutaan tutkia vasteeseen Y, ollaan kiinnostuneita siitä, voidaanko X:n arvoa muuttumalla vaikuttaa Y:n arvoon. Muuttujaan X täytyy tällöin kohdistaa *interventio*, eli antaa X:lle mielivaltainen arvo.

Klassinen tilastollinen menetelmä interventoiden kohdistamiselle on satunnaiskoe. Oikein toteutetussa satunnaiskokeessa ainoastaan intervention kohteena oleva muuttuja vaikuttaa vastemuuttujaan, sillä muiden muuttujien arvot joko vakioidaan tai ne vaihtelevat satunnaisesti. Tällaisen satunnaiskokeen toteutus on kuitenkin monissa tilanteissa mahdotonta, sillä esimerkiksi luonnonilmöihin on mahdotonta vaikuttaa. Ratkaisuksi Pearl on kehittänyt *do-laskennan*, jossa graafeja manipuloimalla ja *do*-operaattoria käyttämällä interventiot ovat täysin hypoteettisia, eli esimerkiksi satunnaiskokeissa tarvittavia konkreettisia interventioita ei tarvita.

On tärkeää huomata, että interventio ei tarkoita samaa asiaa kuin ehdollistaminen. Kun muuttujaan kohdistetaan interventio, muutetaan kausaalimallin rakennetta. Tällöin määritelmän 4.1 kohdassa 3 esitetty funktio $v_i = f_i(pa_i, u_i)$ poistetaan kausaalimallista ja funktion korvaa mielivaltainen arvo. Tuloksena saadaan siis uusi kausaalimalli ja siten myös graafi, jossa intervention kohteena olevaan muuttujaan ei enää kohdistu kaarta. Kaikki muut funktiot eli kaaret pysyvät koskemattomina.

Kun muuttujalle X annetaan mielivaltainen arvo $X = x$, merkitään sitä $do(X = x)$. Nyt siis $P(Y = y \mid do(X = x))$ on todennäköisyys $Y = y$, kun muuttujaan X on kohdistettu interventio. (Pearl, Glymour ja Jewell 2016, s. 53–56; Pearl 2009a, s. 68–70.)

Esimerkki 6.1. Käytetään Pearlin, Glymourin ja Jewellin esimerkkiä (2016, s. 55–58). Kappaleessa 3.1 kuvattu Simpsonin paradoksi voidaan nyt kuvata kuvan 6.1 mukaisesti, jossa solmu Z on sukupuoli, solmu X on lääke ja solmu Y on paranemisaste. Jotta lääkkeen kausaalivaikutusta saadaan estimoitua, käytetään apuvälineinä hypoteettista interventiota ja *do*-laskentaa. Kohdistetaan muuttujaan X kaksi interventiota, joista ensimmäisessä kaikille potilaille annetaan lääkettä ja toisessa kenellekään potilaista ei anneta lääkettä. Merkitään näitä interventioita $do(X = 1)$ ja $do(X = 0)$, joiden avulla estimoidaan todennäköisyyksien erotusta (*average causal effect, ACE*)

$$P(Y = 1 \mid do(X = 1)) - P(Y = 1 \mid do(X = 0)).$$

Laskuissa täytyy kuitenkin huomioida solmu Z, sillä Z:n arvo määrittää ryhmän, jolle kausaalivaikutus lasketaan. Käytetään tähän kaavaa,

$$P(Y = y \mid do(X = x)) = \sum P(Y = y \mid X = x, Z = z)P(Z = z),$$

joka huomioi kaikki Z:n määrittämät ryhmät.

Olkoon $X = 1$, kun potilas ottaa lääkkeen, $Z = 1$, kun potilas on mies ja $Y = 1$, kun potilas kokee parantavan vaikutuksen. Saadaan kaava

$$P(Y = 1 \mid do(X = 1)) = P(Y = 1 \mid X = 1, Z = 1)P(Z = 1) + P(Y = 1 \mid X = 1, Z = 0)P(Z = 0),$$

johon taulukon 3.1 arvot sijoittamalla saadaan todennäköisyydet

$$P(Y = 1 \mid do(X = 1)) = \frac{0.93(87 + 270)}{700} + \frac{0.73(263 + 80)}{700} = 0.832$$

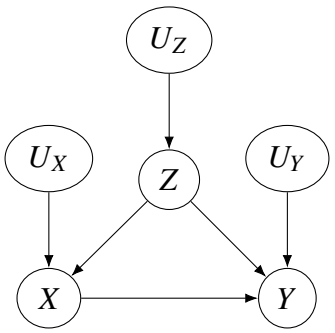
ja

$$P(Y = 1 \mid do(X = 0)) = \frac{0.87(87 + 270)}{700} + \frac{0.69(263 + 80)}{700} = 0.7818.$$

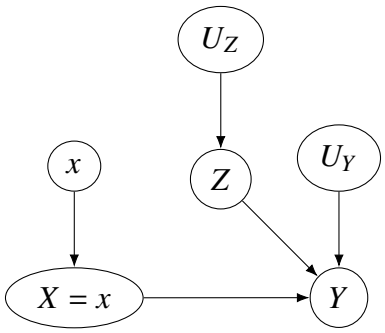
Nyt siis $ACE =$

$$P(Y = 1 \mid do(X = 1)) - P(Y = 1 \mid do(X = 0)) = 0.832 - 0.7818 = 0.0502.$$

Todennäköisyys puoltaa lääkkeenottamista. Voidaan siis todeta, että sukupuolen huomiotta jättäminen olisi johtanut virhepäätelmään.



Kuva 6.1. Graafi, jossa sukupuoli Z on sekoittava tekijä. Lääkkeeseen X ei ole vielä kohdistettu interventiota.



Kuva 6.2. Kuvan 6.1 graafi, kun muuttujaan X on kohdistettu interventio.

7 Yhteenveto

Tässä tutkielmassa esitettiin graafeihin perustuvaa Judea Pearlin kausaaliteoriaa. Tutkielman päätavoite oli esittää Pearlin kausaaliteoriasta tiivistetty kokonaisuus, jonka avulla lukija johdatellaan kausaalipäättelyn ongelmiin. Tutkielman edetessä lukija samalla huomaa, kuinka esitetyt työkalut auttavat ongelmanratkaisussa. Tutkielman toinen tavoite oli sisällön rakentaminen niin, ettei lukija tekstiä ymmärtääkseen tarvitse ennakkotietoa graafeista tai kausaalipäättelystä. Mielestäni näissä tavoitteissa onnistuttiin.

Kausaalisuus on käsitteenä laaja-alainen, joten ennen kausaalipäättelyyn perehtymistä tutkielman alussa kerrottiin kausaalisuuden historiasta. Kausaalisuuden sijasta tieteen historiassa huomataan, kuinka eri aikakaudet ovat muokanneet ihmisen käsitystä ympäröivästä maailmasta.

Historiaosuuden jälkeen käytiin läpi Simpsonin paradoksi, joka kuvastaa kausaalipäättelyn tärkeyttä. Perinteisillä tilastollisilla menetelmillä sekoittavien tekijöiden tutkiminen on ollut haastavaa, mutta kausaalipäättelyn avulla paradoksin tuottamaan ongelmaan on saatu selvyyttä.

Kun graafin kätevyys kausaalimallin esittämiseen huomattiin, päästiin käsiksi tutkielman tärkeimpiin aiheisiin: d-erotteluun, interventiohin ja do-laskentaan. D-erottelu on tärkeä työkalu kausaliteetin tutkimisessa sekä pääteksenteossa kausaalimallin oikeellisuudesta. Jos d-erottelun nojalla muuttujat ovat riippumattomia, ei niiden välillä myöskään ole kausaliteettia. Jos muuttujien ajateltiinkin olevan riippuvia, on syytä palata oletuksiin, joihin perustuen kausaalimalli rakennettiin. D-erottelun voidaan siis ajatella toimivan molempiin suuntiin: sen avulla voidaan päättää, mallintaako aineisto graafia vai toisinpäin.

Interventiot ovat olennainen osa kausaalipäättelyä. Pearlin kausaaliteoriassa interventioiden tärkeä ominaisuus on, että graafien avulla ne voidaan suorittaa täysin hypoteettisesti, eli konkreettisia interventioita ei tarvita. Intervention avulla muuttujalle voidaan määrittää mielivaltainen arvo, joka tarkoittaa, että esimerkiksi sekoittavan tekijän aiheuttama vaikutus voidaan katkaista. Interventio ilmaistaan käyttämällä do-laskentaa, jonka do-operaattori formalisoi intervention. Kun do-laskentaa sovelletaan todennäköisyyslaskennan lauseisiin, saadaan kausaalivaikutus estimoitua.

Kausaalipäättelyn ongelmana voidaan kuitenkin pitää sen toistaiseksi vakiintumatonta kieltä. Kausaalipäättelyä varten tehtävien oletusten ja yhtälöiden ilmaisu ei ole vielä täysin yhtenäistä, joten joissain tapauksissa aiempia tutkimuksia voi olla vaikea soveltaa. Vaikka Pearl on tietoinen tästä ongelmasta, on hän silti rohkaissut myös opiskelijoita sisällyttämään kausaalipäättelyä ongelmien ratkaisemiseen.

Lähteet

Hopkins, B. (2004). ”Causality and development”. Teoksessa: *Mind and Causality*. Toim. Peruzzi, A. Philadelphia, PA, USA : John Benjamins Publishing Company.

Koivisto, P., Niemistö, R. (2018). ”Graafiteoriaa”. 2. painos. Luentomoniste. Tampereen yliopisto, Informaatiotieteiden yksikkö.

Pearl, J. (2009a). *Causality : Models, Reasoning, and Inference*. 2. painos. Cambridge : Cambridge University Press.

— (2009b). ”Causal inference in statistics: An overview”. *Statistics Surveys*, Vol. 3, 96–146, DOI 10.1214/09-SS057.

Pearl, J., Glymour, M., Jewell, N. (2016). *Causal Inference in Statistics: A Primer*. Chicester: John Wiley & Sons, Incorporated.