hyväksymispäivä arvosana

arvostelija

Simpsonin paradoksi

Toni Merivuori

Helsinki 16. helmikuuta 2003

Kausaaliset mallit seminaari, kevät 2003

HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

Simpsonin paradoksi

Toni Merivuori

Kausaaliset mallit seminaari, kevät 2003

Tietojenkäsittelytieteen laitos

Helsingin yliopisto

16. helmikuuta 2003, 9 sivua

Simpsonin paradoksi

Aiheluokat(Computing Reviews 1998): I.2.4, I.2.5, G.3, F.1.2

Avainsanat: kausaalisuus, kausaalinen mallintaminen, Simpsonin paradoksi

Sisältö

1 Johdanto

Simpsonin paradoksi perustuu havaintoon siitä, että mikä tahansa tilastollinen suhde kahden muuttujan välillä voi kääntyä päinvastaiseksi, jos löydetään kolmas muuttuja, joka korreloi molempien kanssa. Esimerkiksi tietyssä tilastollisessa analyysissä koskien koulun oppilaiden saamia arvosanoja voitaisiin havaita, että tupakoivat oppilaat saavat parempia arvosanoja kuin tupakoimattomat. Kuitenkin jos lisäämme tarkasteltavaksi tekijäksi oppilaan iän, tulos kääntyy päinvastaiseksi kaikissa erikseen tarkasteltavissa ikäryhmissä. Jos myöhemmin analyysiin lisätään tekijäksi oppilaiden vanhempien toimeentulo, voisi tulos jälleen kääntyä päinvastaiseksi, eli havaitaan tupakoinnin nostavan arvosanojen tasoa kussakin tarkasteltavassa ikä-toimeentulo-ryhmässä, ja niin edelleen.

Tilastollisen analyysin parissa paradoksia ei ole kyetty ratkaisemaan, sillä kukaan ei vielä ole kyennyt antamaan täsmällistä menetelmää paradoksin tuomien ongelmien välttämiseksi tai oikeiden tekijöiden valitsemiseksi.

Judea Pearl käsittelee artikkelissaan “Simpson’s Paradox: An Anatomy“ Simpsonin paradoksia kausaalisteoreettisista lähtökohdista ja analysoi syitä miksi ilmiö on käsitetty ja yhä käsitetään paradoksaaliseksi. Pearl osoittaa todennäköisyyslaskennan olevan riittämätön väline kausaalisten riippuvuuksien käsittelyyn ja tarjoaa vaihtoehdoksi edellisen laajennoksen, joka sallii mm. $do()$-operaattorin käytön.

2 Simpsonin Paradoksi: esimerkki

Ajatellaan tilannetta, missä tapahtuma C lisää tapahtuman E todennäköisyyttä annetussa populaatiossa p ja samaan aikaan vähentää tapahtuman E todennäköisyyttä jokaisessa annetun populaation p osapopulaatiossa. Olkoon esimerkiksi F ja $\neg F$ toisensa komplementoivat ominaisuudet, jotka kuvaavat populaation p osapopulaatioita. Tällöin voidaan hyvin kohdata seuraavat epäyhtälöt:

$$P(E \mid C) > P(E \mid \neg C), \tag{1}$$

$$P(E \mid C, F) < P(E \mid \neg C, F), \tag{2}$$

$$P(E \mid C, \neg F) < P(E \mid \neg C, \neg F). \tag{3}$$

Todennäköisyyslaskennan kannalta tässä ei ole mitään hämmästyttävää. Tilanne voi muuttua paradoksaaliseksi silloin kun sille on annettu kausaalinen esitysmuoto. Esimerkiksi liitetään tapahtumille C ja E seuraavat merkitykset: C kuvastaa tietyn lääkkeen käyttöä ja E kuvastaa toipumista. Lisäksi asetetaan ominaisuus F tarkoittamaan “naisena olemista“ (jolloin $\neg F$ tarkoittaa “miehenä olemista“). Nyt epäyhtälöistä $2-3$ nähdään lääkkeen käytön olevan haitallista miehille ja naisille erikseen, mutta epäyhtälön 1 mukaan lääke on hyödyllinen koko populaatiolle. Tämä tuntuu täysin järjen sekä intuition vastaiselta. (Alla olevassa taulukossa on annettu esimerkkiimme sopivat tilastolliset numeroarvot.)

(a)	Yhdistetty taulu	E	$\neg E$	Yht.	Toipumisaste
	lääke(C)	20	20	40	50%
	ei lääkettä($\neg C$)	16	24	40	40%
	Yht.	36	44	80	
(b)	Miehet taulu	E	$\neg E$	Yht.	Toipumisaste
	lääke(C)	18	12	30	60%
	ei lääkettä($\neg C$)	7	3	10	70%
	Yht.	25	15	40	
(c)	Naiset taulu	E	$\neg E$	Yht.	Toipumisaste
	lääke(C)	2	8	10	20%
	ei lääkettä($\neg C$)	9	21	30	30%
	Yht.	11	29	40	

Selitys löytyy todennäköisyyslaskennan tarjoamien välineiden puutteellisuudesta kausaalisten riippuvuussuhteiden käsittelyyn: ehdollinen todennäköisyys merkitsee ilmeistä riippuvuutta "given that we see", kun taas $do()$-operaattori on laadittu esittämään kausaalista riippuvuutta "given that we do". Epäyhtälössä $P(E \mid C) > P(E \mid \neg C)$ tapahtuma C ei kuvaa tapahtumaan E vaikuttavaa kausaalitekijää vaan se on paremminkin tapahtumaan E vaikuttava evidenssi. Paradoksaalisiin tuloksiin voidaan päätyä, koska tapahtumaan C voi liittyä sekoittuneita(confounding) tekijöitä, jotka vaikuttavat molempiin tapahtumiin C ja E. Esimerkissämme lääke tulee hyödylliseksi koko populaatiolle, koska miehet, jotka toipuvat naisia useammin, ovat myös alttiinpia käyttämään lääkettä.

Yleinen menetelmä käsitellä potentiaalisia sekoittumisalttiita tekijöitä on määrätä todennäköisyydet kaikille mahdollisille tekijöille, jotka voivat olla syy sekä tapahtumalle C että E: jos esimerkissämme "miehenä oleminen" ($\neg F$) nähdään olevan syy sekä toipumiselle että lääkkeen käytölle, niin lääkkeen vaikutukset täytyy arvioida erikseen miehille ja naisille (epäyhtälöt $2-3$). Jos siis oletamme että F on ainoa sekoittuva tekijä, epäyhtälöt $2-3$ asianmukaisesti esittävät lääkkeen tehokkuutta omissa populaatioissaan, kun epäyhtälö 1 esittää pelkästään suuntaa antavaa informaatiota sukupuolitiedon (F) puuttuessa.

3 Tilastollinen ahdistus

Tilastotieteiden parissa ollaan kautta historian oltu innottomia käsittelemään Simpsonin paradoksia kausaalisista lähtökohdista. Paradoksi myönnetään todelliseksi ja häiritseväksi ongelmaksi, joka esiintyy tilastoissa ja voi johtaa tilastotarkastelijan vääriin johtopäätöksiin. Kausaalisuus on yleensä sivuutettu ongelman ratkaisemisen yhteydessä, koska sen on ajateltu olevan mentaalinen konstruktio, joka ei ole hyvin määritelty. Paradoksi on haluttu nähdä tilastollisena ilmiönä, joka voidaan havaita, ymmärtää ja välttää käyttämällä tilastollisia menetelmiä. Näin tehdessään tilastotieteilijät ovat kuitenkin sivuuttaneet ongelman perusluonteen, jonka juuret lepäävät kausaalisissa suhteissa.

Pearson käsitti ilmiön johtuvan vääristyneistä kausaaliesityksistä, jotka hän oikaisi tilastollisten korrelaatioiden käsittelyllä sekä sattuma(contingency) taulujen avulla. Hänen seuraajansa jopa väittivät vakavissaan ettei kausaatio ole muuta kuin korrelaation laji. Kieltämällä kaikki viittaukset kausaaliseen intuitioon, päädyttiin paradoksi selittämään vain datan "pahana" ominaisuutena, joka voidaan välttää tarkkaavaisten tutkijoiden toimesta.

Ajan myötä kuitenkin havaittiin puutteita kehitellyissä tilastollisissa menetelmissä paradoksin välttämiseen. Tuolloin käytössä olevat kielet eivät olleet riittäviä ilmaisemaan kausaalisia suhteita, joita tutkijat yrittivät ymmärtää. Lindlay ja Novick olivat ensimmäisinä osoittamassa, ettei ole tilastollista kriteeriä, jonka avulla voidaan välttää vetämästä vääriä johtopäätöksiä tai joka osoittaisi oikean vastauksen saamiseen käytettävän taulun.

Lindlay ja Novick suuntasivat huomion ensin käytännön kysymyksiin, kuten käytämmekö lääkettä vai emmekö käytä uuden asiakkaan saapuessa vastaanotolle: ekvivalentisti luvun 1 esimerkkiin viitaten, käytämmekö yhdistettyä vai sukupuoliriippuvaista taulua. Kysymyksessä päädyttiin pitkien pohdintojen jälkeen vastaukseen, missä pitäisi käyttää sukupuoliriippuvaista taulua, jonka nojalla lääkettä ei pidä käyttää (vrt. luku 1). Seuraavaksi kysyttiin voisiko jokin muu tilastollinen tieto yleisesti osoittaa oikean taulun. Tähän jouduttiin vastaamaan kielteisesti, sillä tietyllä samalla aineistolla voitiin päätyä ristiriitaisiin tuloksiin: kun käytämme luvun 1 taulukon arvoja, mutta vaihdamme taustatilanteen siten että C vaikuttaa ominaisuuteen F (eikä toisinpäin, ks. kuva 1 b-kohta), niin voimme heti päätellä että vastaus löytyy yhdistetystä taulusta.

tähän tulee kuva

Ajattele kahta kausaalista mallia esittävän samaa tilastollista dataa (kuva 1, a- ja b-kohta ovat havainnollisesti (observationally) ekvivalentit) ja toinen niistä johtaa päätökseen käyttää lääkettä mutta toinen ei, niin on selvää, että päätös muodostuu kausaalisten tietojen eikä tilastollisten näkökohtien perusteella. Kuvan 1 c-kohta osoittaa, että ajallinen tieto ei vaikuta päätökseen, sillä nyt F voi tapahtua ennen tai jälkeen C:n ja silti päädytään yhdistetyn-taulun käyttöön.

Nämä esimerkit osoittavat, että tilastollinen tieto yksinään on riittämätön sanomaan mitään varmaa tapahtumien seuraussuhteista ja niiden oikeellisuudesta. Meidän pitää siis käsitellä niitä kausaalisten menetelmien avulla.

4 Kausaalinen käsittely

Käsitellessämme Simpsonin paradoxia olemme joutuneet umpikujaan joko

i) olettamalla että kausaaliset suhteet toimivat todennäköisyyslaskennan lakien alaisena,

tai että

ii) tekemämme epäsuorien oletusten joukko kausaalisista suhteista on väärä.

Ensimmäisen luvun esimerkkimme nojalla epäyhtälöt $(1-3)$ ovat i-kohdan mukaan samaan aikaan voimassa tarjoten meille todennäköisyysmallin väitteen tueksi, kun taas ii-kohdan mukaan ei voi olla olemassa lääkettä, joka on haitallinen sekä miehille että naisille, mutta samaan aikaan on edullinen koko populaatiolle.

Ratkaistaksemme paradoksin meidän on joko näytettävä että kausaalinen intuitiomme on harhaanjohtava (ii) tai kiellettävä että kausaaliset suhteet ovat todennäköisyyslaskennan lakien alaiset (i). Käymme kiinni i-kohtaan ja yritämme näyttää, että kausaaliset suhteet omaavat oman logiikkansa niiden käsittelemiseen ja että tämä logiikka vaatii toimiakseen tärkeän laajennuksen todennäköisyyslaskentaan.

Käytämme $Do()$-operaattoria ja sen logiikkaa, joka soveltuu tilanteeseen mainiosti. Näytämme, että kyseinen logiikka mahdollistaa halutunlaisen lääkkeen olemassa olon. Kausalisen formalismin mukaan lääke C on haitallinen miehille ja naisille, jos

$$P(E \mid do(C), F) < P(E \mid do(\neg C), F), \tag{4}$$

$$P(E \mid do(C), \neg F) < P(E \mid do(\neg C), \neg F). \tag{5}$$

Demonstroidaan, että tapahtuman C on oltava haitallinen populaatiolle kokonaisuudessaan. Tällöin epäyhtälö

$$P(E \mid do(C)) > P(E \mid do(\neg C)). \tag{6}$$

on osoitettava ristiriitaiseksi niiden tietojen kanssa mitä tiedämme lääkkeistä ja sukupuolesta.

Teoreema 1 (*The Sure-Thing Principle*) *Toimenpide C, joka nostaa todennäköisyyttä tapahtumalle E jokaisessa osapopulaatiossa, täytyy myös nostaa tapahtuman E todennäköisyyttä kokonaispopulaatiossa, sillä ehdolla että toimenpide ei muuta osapopulaatioiden todennäköisyysjakaumia.*

Todistetaan teoreema luvun 1 esimerkin valossa, missä populaatio on jaettu miehiin ja naisiin. Nyt on todistettava, että epäyhtälöjen $4-6$ nuolien suunnat eivät päde samanaikaisesti oletuksen kanssa, missä lääkkeellä ei ole vaikutusta sukupuoleen:

$$P(F \mid do(C)) = P(F \mid do(\neg C)) = P(F). \tag{7}$$

Laajennetaan yhtälöä $P(E \mid do(C))$ ja käytetään yhtälöä (7) saamme

$$P(E \mid do(C)) = P(E \mid do(C), F)P(F \mid do(C)) + P(E \mid do(C)), \neg F)P(\neg F|do(C)) = P(E \mid do(C)), F). \tag{8}$$

Tekemällä samoin operaatiolle $do(\neg C)$ saamme:

$$P(E \mid do(\neg C)) = P(E \mid do(\neg C), F)P(F) + P(E \mid do(\neg C))P(\neg F) \tag{9}$$

Koska kaikki termit yhtälön 8 oikealla puolella ovat pienempiä kuin vastaavat yhtälössä 9, voimme päätellä, että $P(E \mid do(C)) < P(E \mid do(\neg C))$, josta väitteemme seuraa. Näin ollen lääke C on haitallinen myös populaatiolle kokonaisuudessaan, jos lääke C ei vaikuta sukupuoleen F.

Nyt voimme hahmottaa mistä kausaalinen intuitiomme on kotoisin: ilmeinen, mutta ratkaiseva oletus meidän intuitiivisessa logiikassamme on ollut, että lääkkeet eivät vaikuta sukupuoleen. Tämä selittää miksi intuitiomme muuttuu rajusti, kun F esitetään välitapahtumana, johon lääke (C) vaikuttaa (ks. kuvan 1 b-kohta). Tällöin intuitiivisen logiikkamme mukaan on löydettävissä lääke, joka täyttää ehdot epäyhtälöille $(4-6)$ ja lisäksi jonka on sopimatonta vaikuttaa F:ään. Jos C vaikuttaa F:ään, niin yhtälöä (8) ei voida johtaa ja erotus $P(E \mid do(C)) - P(E \mid do(\neg C))$ voi olla positiivinen tai negatiivinen riippuen yhtälöiden $P(F \mid do(C))$ ja $P(F \mid do(\neg C))$ suhteellisista voimakkuuksista. Kun tiedetään ettei C:llä ja E:llä ole yhteistä tekijää, C:n tehokkuutta on arvioitava suoraan yhdistetty-taulusta, eikä sukupuoliriippuvaisista tauluista.

Huomaa, että kausaalinen logiikkamme sallii meidän käsitellä luvun 1 esimerkin kaltaisia tapauksia, mutta kuitenkin torjuu hyväksymästä epäyhtälöiden $4-6$ samanaikaisen voimassaolon. Tässä mielessä $do()$-operaattorin ja sen käyttämä logiikka poikkeaa aikaisemmista menetelmistä ja näin ollen voimme välttää analyyseissämme ajautuvan Simpsonin paradoksin kaltaisiin tuloksiin.