

Kustannustehokkuus kokeellisessa ja havainnoivassa tutkimuksessa

■ Juha Karvanen

Tilastotieteelliset periaatteet tehokkaiden tutkimusasetelmien suunnitteluun ja analysointiin ovat käytettävissä kaikilla tieteenaloilla. Vaaditaan kuitenkin työtä, että tutkimusongelma saadaan käännettyä tilastotieteen kielelle ja että tilastotieteen antama vastaus saadaan käännettyä takaisin sovellusalan kielelle.

Tilastotiedettä voisi kutsua ”epävarmuustieteeksi”. Epävarmuutta on sekä ilmiöiden luontainen satunnaisuus että tiedon puute. Epävarmuuden vähentäminen eli tiedon lisääminen on tieteen yleinen tehtävä. Tilastotiede taas pyrkii epävarmuuden mittaamiseen, ymmärtämiseen ja hallitsemiseen. Tilastotieteen tehtävä on siis tiedon lisääminen epävarmuudesta eli epävarmuuden vähentäminen epävarmuudesta.

Tilastotieteilijän näkökulmasta tieteelliseen päättelyyn tarvitaan kolme elementtiä: kausaaliioletukset, tutkimusasetelma ja aineisto. Kausaaliioletuksilla tarkoitetaan tunnettuja ja oletettuja syy-seuraussuhteita tutkittavien ilmiöiden välillä. Syy-seuraussuhteet on määriteltävä aina jossakin perusjoukossa, ja ne eivät välttämättä päde perusjoukon ulkopuolella. Tutkimusasetelmalla tarkoitetaan yksityiskohtaista suunnitelmaa aineiston keräämiseen käytettävästä menettelystä. Tutkimusasetelma pitää sisällään myös ymmärryksen siitä, mitkä tekijät vaikuttavat aineiston valikoitumiseen. Aineisto eli data on tutkimusasetelman mukaisesti kerättyjä havaintoja tutkittavien ilmiöiden välisistä syy-seuraussuhteista.

Tilastollinen malli esittää kausaaliioletukset ja tutkimusasetelman täsmällisessä matemaattisessa muodossa. Malli kuvaa sekä muuttujien väliset systemaattiset yhteydet että muuttujiin liittyvän epävarmuuden. Tuntemattomat tekijät kuvataan mallin parametrien avulla. Datan perusteella tuntemattomista tekijöistä voidaan

oppia ja mallin parametrit voidaan estimoida. Parametreihin sisältyy aina epävarmuutta, mutta yleensä epävarmuus vähenee, kun aineistoa kerätään enemmän. Tieteellisessä päättelyssä kausaaliivaikutusten voimakkuudesta tehdään päätelmiä aineiston perusteella ottaen huomioon aineiston tuottamiseen käytetty tutkimusasetelma. Esimerkiksi, jos halutaan tehdä päätelmiä koko väestöstä mutta tutkimusasetelmassa nuorista ikäluokista on otettu suhteellisesti pienempi otos kuin vanhoista ikäluokista, on aineistoa painotettava vastaamaan kunkin ikäryhmän väestöosuutta.

Tutkimusasetelman valinta

Tutkija ei voi luoda aineistoa eikä päättää syy-seuraussuhteista. Sen sijaan tutkija voi valita tutkimusasetelman. Tutkimusasetelma ratkaisee, saadaanko kerättävän aineiston pohjalta tietoa tutkittavasta syy-seuraussuhteesta. Hyvin määritellyt tavoitteet ovat välttämättömiä, mutta eivät riitä takaamaan tutkimuksen onnistumista, jos käytetyllä tutkimusasetelmalla ei ole mahdollista vastata tutkimuskysymyksiin. Tyypillisesti ongelmia aiheuttavat liian pieni otoskoko, valikoituminen otokseen ja puuttuvat mittaukset. Joskus suunnittelussa tehty virheet on mahdollista korjata aineistoa analysoidessa, mutta usein näin ei ole. Tällöin tutkimuksen kustannustehokkuus on erittäin huono: käytettiin resursseja, mutta ei saavutettu tuloksia.

Tutkimuksia voi luokitella monella eri tavalla, mutta tilastotieteen kannalta tärkeä luokittelu on jako kokeelliseen ja havainnoivaan tutkimukseen. Kokeellisessa tutkimuksessa tutkija voi itse valita jonkin muuttujan arvon. Esimerkiksi fysiikassa tutkija voi vaikkapa säädellä sähkövirran suuruutta ja mitata jännitteen vaihtelua. Ihmisiä tutkittaessa vaaditaan kokeellisessa tutkimuksessa usein satunnaistuksen

käyttöä. Esimerkiksi kliinisessä lääketutkimuksessa päätös siitä, annetaanko potilaalle lääkettä vai plasebovalmistetta on satunnaistettava eli tehtävä huolellisesti määritellyllä arvontamenetelyllä. Havainnoivassa tutkimuksessa tutkija ei voi valita muuttujan arvoja mutta voi tehdä päätöksen siitä, mitkä yksilöt tutkimukseen valitaan. Valinta voi olla yksinkertainen satunnaistotos populaatiosta tai hyvinkin monimutkainen valintajärjestely, jossa valintatodennäköisyydet riippuvat aiemmista mittauksista.

Kokeellisessa tutkimuksessa

Tehokas koeasetelma hyödyntää kaikkea saatavilla olevaa tietoa, myös sellaista tietoa, jota kerätään tutkimuksen aikana. Esimerkkinä tällaisesta mukautuvasta asetelmasta mainitsen kuulokynnyksen määrittämisen. Kuulokynnyksen voi määritellä esimerkiksi sellaiseksi äänenvoimakkuudeksi, jonka ihminen pystyy kuulemaan keskimäärin joka toinen kerta. Tämä äänenvoimakkuus on luonnollisesti erilainen eri ihmisillä. Kuulokokeessa testattava henkilö painaa nappia joka kerta, kun kuulee äänen. Jos henkilö kuulee äänen, äänenvoimakkuutta pienennetään, ja jos henkilö ei kuule ääntä, äänenvoimakkuutta kasvatetaan. Ärsyke siis riippuu kokeen aikana mitatusta vasteesta. Tämänäyttypiset mukautuvat koeasetelmat ovat yleensä paljon tehokkaampia kuin etukäteen kiinnitetty koeasetelmat.

Samantapaista mukautuvaa koeasetelmaa voi hyödyntää myös aivotutkimuksessa ja fysiikassa. Aivosähkökäyrän (elektroenkefalografia, EEG) tai aivomagneettikäyrän (magnetoenkefalografia, MEG) mittaamien vaatii kaikkine etukäteisvalmisteluineen runsaasti tutkijan työaika. Koehenkilöitä on yleensä vähän, eikä yhtä koehenkilöä voi rasittaa mittauksella kohtuuttoman kauan. On siis tärkeää, että mittauksiin käytävissä oleva aika hyödynnetään mahdollisimman tehokkaasti. Mukautuvassa koeasetelmassa tämä tarkoittaa sitä, että koehenkilölle esitettävä ärsyke riippuu aiemmin mitatuista vasteista.

Kokeellisessa fysiikassa mittausten tekeminen on yleensä nopeaa, mutta tutkittava ilmiö saattaa pysyä stabiilina vain lyhyen ajan. Esimerkkinä tilastollisen koesuunnittelun käytöstä kvanttify-

siikassa voin mainita niin sanottuihin Josephsonin liitoksiin liittyvän kokeen, jossa tutkija kontrolloi virtapulssin suuruutta ja tarkkailee vasteena jännitepulssin ilmaantumista. Jos virtapulssi on riittävän suuri, jännitepulssi havaitaan aina. Jos virtapulssi on riittävän pieni, jännitepulssia ei havaita koskaan. Tilanne on siis samankaltainen kuulokynnyksen määrittämisen kanssa. Tavoitteena on löytää sellainen kapea virta-alue, jolla jännitepulssin ilmaantuminen on aidosti satunnainen ilmiö. Mukautuva koeasetelma tämän ongelman ratkaisemiseksi on kaksivaiheinen (Karvanen, Vartiainen, Timofeev & Pekola 2007). Ensimmäisessä vaiheessa käytetään puolitus-hakua tutkittavan virta-alueen rajaamiseksi. Toisessa vaiheessa jännitepulssin todennäköisyyttä mallinnetaan tilastollisella mallilla ja virtapulssien suuruus määrätään adaptiivisesti koesuunnittelun teorian mukaisella optimaalisella asetelmalla. Käytännön kokeessa mukautuva koeasetelma vaatii vain kymmenesosan aiemmin käytössä olleen koeasetelman vaatimasta ajasta.

Havainnoivassa tutkimuksessa

Tehokkaita mukautuvia tutkimusasetelmia voidaan käyttää myös havainnoivassa tutkimuksessa. Aina ei ole kustannustehokasta mitata kaikkia muuttujia kaikilta tutkimuksen osallistujilta. Käytän esimerkkinä terveystieteellisiä seuranta-tutkimuksia, joissa tavoitteena on selvittää jonkin taudin ilmaantumiseen vaikuttavat riskitekijät. Tutkimuksen alussa osallistujilta otetaan verinäytteitä ja heiltä mitataan riskitekijöitä, kuten kolesterolitaso, verenpaine ja painoindeksi. Verinäytteitä myös pakastetaan myöhempää käyttöä varten. Seurantavaihe kestää esimerkiksi kymmenen vuotta, ja sen aikana rekisteröidään osallistujien tautitapahtumat.

Oletetaan, että tutkimuksessa haluttaisiin selvittää myös geenien vaikutusta tautiriskiin. Geenitiedon määrittäminen on edelleen kallista, ja kustannussyistä sitä ei ole mahdollista tehdä kaikille osallistujille. Niinpä niin sanotussa tapauskohorttiasetelmassa kallis muuttuja, tässä tapauksessa geenitieto, mitataankin vain osajoukosta. Tämä tapauskohorttijoukko koostuu sairastuneista ja pienestä satunnaisotoksesta tutkimuk-

seen osallistujista. Geenitiedon määrittämiseen käytetään tutkimuksen alussa pakastettuja verinäytteitä. Valintatodennäköisyys tapaus-kohorttijoukkoon riippuu siis tutkimuksen aikana tehdyistä havainnoista. Koska tällä tavalla valittu osajoukko ei ole yksinkertainen satunnaisotos, sen analysointiin tarvitaan tilastotieteen menetelmiä, jotka ottavat huomioon käytetyn tutkimusasetelman (Kulathinal, Karvanen, Saarela & Kuulasmaa 2007). Tutkimusasetelmaa on edelleen mahdollista kehittää siten, että myös tutkimuksen alussa tehdyt riskitekijämittaukset vaikuttavat siihen, keneltä uudet mittaukset tehdään (Karvanen, Kulathinal & Gasbarra 2009).

Suunnitellusti puuttuva tieto

Empiirisessä tutkimuksessa tutkijan tulee olla valmistautunut puuttuvan tiedon käsittelyyn. Kun kyse on tietoisesta päätöksestä jättää kustannussyistä mittauksia tekemättä joiltakin yksilöiltä, puhutaan suunnitellusti puuttuvasta tiedosta (*data missing by design*). Tällainen aineisto edellyttää kehittyneitä analyysimenetelmiä, joissa otetaan huomioon tutkimusasetelman aiheuttama tiedon puuttuminen. Keskeiset periaatteet ja menetelmät puuttuvan tiedon käsittelyyn ovat olleet tunnettuja vuosikymmeniä (Rubin 1976; Little & Rubin 2002), mutta niiden soveltaminen vaatii aina syventymistä käsillä olevaan tutkimusongelmaan. Usein tilannetta monimutkais-taa se, että suunnitellusti puuttuvan tiedon lisäksi osa havainnoista puuttuu vastauskadon tai jonkin muun suunnittelemattoman syyn vuoksi. Monimutkaisissa tilanteissa kausaaliioletusten ja tutkimusasetelman täsmällinen kuvaaminen on edellytys tilastollisen analyysin onnistumiselle.

Tiedon tarve yhteiskunnassa on kasvanut. Esimerkiksi ympäristön tilaa halutaan seurata entistä tarkemmin, mikä edellyttää paljon kalliita ja aikaavieviä mittauksia. Resurssija tiedon hankintaan ei kuitenkaan ole käytettävissä aiempaa enemmän.

Tilastotieteellä on paljon annettavaa mittauksen ja tutkimusasetelmien suunnitteluun. Tieteellisessä tutkimuksessa hyvin suunniteltu tutkimusasetelma voi olla ratkaiseva edistysaskel pyrittäessä keskinkertaisesta tutkimuksesta

kohti huippututkimusta. Yhteiskunnan kannalta tehokkaat tutkimusasetelmat voivat johtaa merkittäviin kustannussäästöihin.

Monissa tapauksissa mittaamisen ja tiedonkeruun tehostaminen on mahdollista mutta edellyttää kahta muutosta ajattelutapaan: Ensiksi, päätöksentekoa ja mittausten suunnittelua on tarkasteltava yhtenä kokonaisuutena siten, että päätöksenteon tarpeet määritellään täsmällisesti ja tutkimusasetelma johdetaan suoraan näistä tarpeista. Toiseksi, on hyväksyttävä se tosiasia, että tehokkaita tutkimusasetelmiä ei aina voi analysoida yksinkertaisilla menetelmillä vaan analyysimenetelmät vaativat tilastotieteellistä osaamista.

Uuden sukupolven tilastotiede on rohkeaa ja päättäväistä epävarmuustiedettä: tilastotiede käsitetään laaja-alaisesti, ja tutkimuksessa ja opetuksessa ajattelu perustuu uusimpaan eturivin tutkimukseen. Tilastotieteen tutkimusongelmat saavat motivaationsa sovellusalojen ongelmista ja tilastotieteen merkitystä tuodaan aktiivisesti esille tiedeyhteisössä ja julkisessa keskustelussa. Uuden sukupolven tilastotieteilijänä esitän kaikille niin kokeellista, havainnoivaa kuin teoreettistakin tutkimusta tekeville avoimen kutsun kustannustehokkaaseen yhteistyöhön.

Kirjallisuus

- Karvanen, Juha, Vartiainen, Juha J., Timofeev, Andrey & Pekola, Jukka 2007. Experimental designs for binary data in switching measurements on superconducting Josephson junctions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 56 (2), s. 167–181.
- Karvanen, Juha, Kulathinal, Sangita & Gasbarra, Dario 2009. Optimal designs to select individuals for genotyping conditional on observed binary or survival outcomes and non-genetic covariates. *Computational Statistics & Data Analysis*, 53, s. 1782–1793.
- Kulathinal, Sangita, Karvanen, Juha, Saarela, Olli & Kuulasmaa, Kari, for the MORGAM Project 2007. Case-cohort design in practice – experiences from The MORGAM Project. *Epidemiologic Perspectives & Innovations* 4:15.
- Little, Roderick J.A. & Rubin, Donald B. 2002. *Statistical Analysis with Missing Data*. Toinen painos. New York: John Wiley.
- Rubin, Donald B. 1976. Inference and missing data. *Biometrika*, 63, s. 581–592.

Kirjoittaja on tilastotieteen professori Jyväskylän yliopistossa. Artikkelin perustuu hänen juhlaluentoonsa 12. joulukuuta 2012.