Puuttuva tieto ja vilppi

Juha Karvanen

Saako tutkija analyysivaiheessa lisätä tai poistaa havaintoja aineistostaan? Jokaisen eettisesti valveutuneen tutkijan tulisi ensi reaktionaan huudahtaa "Ei tietenkään saa!". Asiaa tarkemmin pohdittaessa vastaus ei enää olekaan yksiselitteinen. Aineiston tietoisen rajaamisen lisäksi havaintoja voi lähes huomaamattomasti jäädä pois käytettäessä tilastollisia ohjelmistoja. Aineiston tekaiseminen on selkeästi tuomittavaa, mutta toisaalta imputointimenetelmät luovat havaintoja silloin, kun niitä ei ole olemassa. Mikä siis on sallittua ja mikä ei?

Tämän kirjoituksen tavoitteena on kertoa puuttuvan tiedon oikeasta ja virheellisestä käsittelystä tilastotieteen ja hyvän tieteellisen käytännön näkökulmasta. Keskeisenä päämääränä on lisätä lukijan ymmärrystä puuttuvan tiedon vaikutuksesta tieteelliseen päättelyyn.

Puuttuvan tiedon ongelma on läsnä kaikkialla, missä aineistoja kerätään. Kyselytutkimuksissa suuri joukko otokseen valituista ei halua osallistua ja osallistujistakin moni jättää vastaamatta joihinkin kysymyksiin. Kokeellisessakin tutkimuksessa puuttuvaa tietoa voi syntyä esimerkiksi mittalaitteiden toimintahäiriöiden, teknisten rajoitusten tai mittaajan huolimattomuuden takia. Puuttuvaksi tiedoksi voidaan ymmärtää myös tiedon osittainen puuttuminen. Havaitusta arvosta voidaan tietää esimerkiksi vain, että se on suurempi kuin tietty tunnettu arvo. Tämä on tyypillistä seurantatutkimuksissa, joissa henkilöiden iät tunnetaan, mutta elinikää ei voida määrittää elossa oleville. Tässä kirjoituksessa puuttuvuudella tarkoitetaan kuitenkin tiedon puuttumista kokonaan.

Tutkijan tulisi ensisijaisesti aina pyrkiä siihen, että kaikki kerättäväksi suunniteltu tieto saadaan kerättyä. Tutkimuseettisen neuvottelukunnan ohjeen (Tutkimuseettinen neuvottelukunta 2013) mukaan tulosten ja tutkimusaineistojen puutteellinen kirjaaminen ja säilyttäminen katsotaan piittaamattomuudeksi hyvästä tieteellisestä käytännöstä. Jos jotakin tietoa ei ole saatu kerättyä, tulisi pohtia keinoja, joilla tietoa kuitenkin voitaisiin saada. Käytännössä kysymykseen tulevat esimerkiksi mittauksen suorittaminen uudelleen, vastaamatta jättäneiden tavoittelu toistamiseen ja vaihtoehtoisten tietolähteiden käyttö. On hyödyllistä, että tiedonkeruuprosessiin liittyvä tieto tallennetaan yksityiskohtaisesti. Tällaista tietoa kutsutaan joskus paradataksi (Couper 1998; Kreuter, Couper ja Lyberg 2010), ja sitä voidaan käyttää tutkittaessa tiedonkeruumenetelmän vaikutusta tiedon kattavuuteen ja laatuun.

Puuttuvan tiedon syntymistä ei kuitenkaan yleensä voi estää ja tiedonkeruuvaiheen korjaustoimenpiteiden vaikutukset ovat rajalliset. Tällöin tutkijan tulee ottaa kantaa siihen, kuinka puuttuvaa tietoa käsitellään aineistoa analysoitaessa. Usein käytetty näennäisen yksinkertainen vaihtoehto on rajata puutteellisesti havaitut tilastoyksiköt analyysin ulkopuolelle. Tätä menettelyä kutsutaan täydellisten havaintorivien analyysiksi ja englanniksi siihen viitataan esimerkiksi termeillä *complete-case analysis* tai *listwise deletion* (Little ja Rubin 2002).

Yksinkertaisuudestaan huolimatta täydellisten havaintorivien analyysi ei ole yleispätevä ratkaisu puuttuvan tiedon ongelmaan. Vaikka otos alunperin edustaisikin kiinnostuksen kohteena olevaa populaatiota, voi tiedon puuttuminen muuttaa tilannetta. Esimerkiksi, jos miehet ovat vastanneet tiettyyn kysymykseen naisia harvemmin, naisten vastaukset ylikorostuvat täydellis-

ten havaintorivien analyysissa, jota ei ole painotettu sukupuolten vastausosuuksien mukaan. Painottamaton täydellisten havaintorivien analyysi antaa oikeita tuloksia vain mikäli puuttuvuus on täysin satunnaista (*missing completely at random*) eli ei riipu mistään mitatuista tai mittaamattomista tekijöistä (Little ja Rubin 2002).

Vaikka aineisto olisikin edustava ja puuttuvuus täysin satunnaista, voi täydellisten havaintorivien analyysi johtaa datan tehottomaan käyttöön. Jos aineistossa on vaikkapa 20 muuttujaa ja kustakin muuttujasta puuttuu 10 % havainnoista riippumattomasti muista muuttujista, analyysiaineistoon jää jäljelle vain 12 % alkuperäisen aineiston havaintoriveistä. Tässä vaiheessa tutkimuseettisten hälytyskellojen tulisi soida. Suurella vaivalla kerättyä tutkimusaineistoa käytetään hyvin tehottomasti, mikä tarkoittaa tutkimusmäärärahojen tuhlaamista. Tätä voidaan pitää hyvän tieteellisen käytännön loukkauksena, koska Tutkimuseettisen neuvottelukunnan ohjeen mukaan vääristelyksi (*falsification*) luetaan myös johtopäätösten kannalta olennaisten tulosten tai tietojen esittämättä jättäminen.

Imputointi eli puuttuvien tietojen korvaaminen keinotekoisesti luoduilla havainnoilla on yksi mahdollinen tapa parantaa aineiston käytön tehokkuutta. On kuitenkin ilmeistä, että mielivaltaisella tavalla toteutettu imputointi ei ole hyvän tieteellisen käytännön mukaista, vaan kyseessä on tietojen tekaiseminen (*fabrication*). Lukija saattaa perustellusti kysyä, voiko imputointi koskaan olla muuta kuin tietojen tekaisemista. Tilastotiede antaa vastauksen, että tiettyjen ehtojen täyttyessä moni-imputointi (*multiple imputation*, Rubin 1987; van Buuren 2012) on luvallinen ja tieteellisesti pätevä lähestymistapa puuttuvan tiedon käsittelyyn. Moni-imputoinnissa puuttuva tieto korvataan usealla mahdollisella havainnolla, jolloin puuttuvaan havaintoon liittyvä epävarmuus tulee otettua huomioon. Tarvittavien imputointien määrä on yllättävän pieni (Rubin 1987). Esimerkiksi kymmenen imputointia riittää monessa tapauksessa. Moni-imputointi tuottaa tällöin kymmenen aineistoa, jotka ovat samanlaisia todellisten havaintojen osalta, mutta poikkeavat imputoitujen havaintojen osalta. Tilastollinen analyysi toistetaan jokaiselle imputoidulle aineistolle ja tulokset yhdistetään tavalla, joka ottaa huomioon sekä puuttuvaan tietoon että imputointiprosessiin liittyvän epävarmuuden (Rubin 1987; van Buuren 2012).

Moni-imputointi on nykyään teknisesti suoritettavissa useilla tilastollisilla ohjelmistoilla. Valitettavasti ohjelmistojen käyttö ei takaa moni-imputoinnin onnistumista vaan voi johtaa tuloksiin, jotka ovat täysin virheellisiä. Tutkijan tulee huolellisesti tarkastella moni-imputoinnissa vaadittavien oletusten voimassaoloa ja tutkia imputoitujen arvojen yhteensopivuutta havaittuihin arvoihin. Keskeinen oletus tunnetaan hieman harhaanjohtavasti nimellä satunnainen puuttuvuus (*missing at random*, Rubin 1976) ja konseptin täsmällinen määritelmä herättää edelleen keskustelua (Seaman, Galati, Jackson ja Carlin 2013). Perusajatus on, että puuttuvuus voi riippua kaikesta havaitusta datasta, mutta ei itse arvosta, jota ei ole havaittu. Esimerkiksi pitkittäistutkimuksessa oletus tarkoittaa, että puuttuvuus voi riippua kaikista aiemmista mittauksista, mutta oletuksen voimassaoloa ei voi päätellä aineistosta, vaan siihen vaaditaan tutkittavan ilmiön ja tiedonkeruumenetelmän tuntemusta. Lisäksi moni-imputoinnissa edellytetään, että imputointimallin tuottamat imputoinnit uskottavasti kuvaavat puuttuvia arvoja. Useiden ohjelmistojen vakioasetuksena käyttämä lineaarinen malli ei läheskään aina ole sopiva imputointimalliksi.

Kaikkein vaikeimmin hallittavia ovat tilanteet, jossa oletus satunnaisesta puuttuvuudesta ei ole voimassa. Esimerkiksi kysyttäessä henkilön vuosituloista on mahdollista, että kaikkein suurituloisimmat ja kaikkein pienituloisimmat jättävät vastaamatta muita herkemmin. Tällöin puuttuvuusmekanismia pitää mallintaa eksplisiittisesti substanssitietämyksen perusteella. Tämä on usein vaikea tehtävä, ja tulosten herkkyyttä tehdyille mallioletuksille onkin syytä tutkia systemaattisesti. Malliparametrien estimointi edellyttää yleensä kehittyneiden tilastotieteellisten menetelmien, kuten Bayes-päättelyn (Gelman, Carlin, Stern, Dunson, Veh-

tari ja Rubin 2013) tai EM-algoritmin (Dempster, Laird ja Rubin 1977; McLachlan ja Krishnan 1997) käyttöä.

Moni-imputointi, sen paremmin kuin täydellisten havaintorivien analyysikaan, ei siis takaa tulosten luotettavuutta puuttuvan tiedon tapauksessa. Hyvään tieteelliseen käytäntöön ei ole olemassa oikopolkuja, vaan tutkijan tulee syvällisesti pohtia tiedon puuttuvuuteen johtavia syitä ja mekanismeja. Tämän tietämyksen pohjalta on mahdollista valita sopivat mallit ja menetelmät puuttuvan tiedon käsittelyyn. Helppoa tämä ei kuitenkaan ole. Usein analyysi vie aikaa vähintään kymmenkertaisesti verrattuna tilanteeseen, jossa puuttuvaa tietoa ei olisi. Jos vaihtoehtona on se, että aineistoa ei voi puuttuvan tiedon takia käyttää lainkaan, investointi puuttuvan tiedon käsittelyyn kannattaa.

Kokemukseni mukaan puutteet epätäydellisesti havaittujen aineistojen analysoinnissa johtuvat yleensä tietämättömyydestä ja vain harvoin piittaamattomuudesta. Varsinainen vilppi on nähdäkseni kyseessä, jos tutkija tietoisesti pyrkii esittämään todellista parempia tuloksia esimerkiksi käyttämällä täydellisten havaintorivien analyysia ilman mainintaa puuttuvan tiedon osuudesta tai käyttämällä yksinkertaista imputointia ja esittämällä aineiston ikään kuin se ei sisältäisi lainkaan puuttuvia havaintoja.

Puuttuvan tiedon kanssa kamppailevaa tutkijaa saattaa lohduttaa se, että puuttuvaan tietoon liittyvät kysymykset ovat usein erittäin haastavia tilastotieteilijällekin. Tätä ei kuitenkaan saa käyttää tekosyynä sopimattomien menetelmien käytölle. Hyvä tieteellinen käytäntö ei salli kompromisseja puuttuvan tiedon käsittelyssä.

Kirjallisuus

Couper, Mick P. 1998. Measuring survey quality in a CASIC environment. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 41–49.

Dempster, Arthur P., Laird, Nan M., ja Rubin, Donald B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38.

Gelman, Andrew, Carlin, John B., Stern, Hal S., Dunson, David B., Vehtari, Aki ja Rubin, Donald B. 2013. *Bayesian data analysis*, Kolmas painos. Boca Raton, FL: Chapman & Hall/CRC Press.

Kreuter, Frauke, Couper, Mick ja Lyberg, Lars 2010. The use of paradata to monitor and manage survey data collection. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 282–296.

Little, Roderick J.A. ja Rubin, Donald B. 2002. *Statistical Analysis with Missing Data*, Toinen painos, New York: John Wile & Sons.

McLachlan, Geoffrey J. ja Krishnan, Thriyambakam 1997. *The EM algorithm and extensions. New York:* John Wiley & Sons.

Rubin, Donald B. 1976. Inference and missing data. *Biometrika*, 63: 581–592.

Rubin, Donald B. 1987. *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons.

Seaman, Shaun, Galati, John, Jackson, Dan ja Carlin, John 2013. What is meant by "missing at random"? *Statistical Science*, 28(2): 257–268.

Tutkimuseettinen neuvottelukunta 2013. *Hyvä tieteellinen käytäntö ja sen loukkausepäilyjen käsitteleminen Suomessa, Tutkimuseettisen neuvottelukunnan ohje 2012*. Helsinki: Tutkimuseettinen neuvottelukunta.

Van Buuren, Stef 2012. *Flexible imputation of missing data*. Boca Raton, FL: CRC Press.

Kirjoittaja on tilastotieteen professori Jyväskylän yliopistossa.