

## Puuttuvien muuttujanarvojen käsittelystä Terveys 2000 – aineiston analysoinnissa

### **Hallinnollisia asioita**

Terveys 2000 –tutkimuksessa onnistuttiin varsin hyvin osallistumisaktiivisuuden maksimointipyrkimyksessä, ja osallistuneilla henkilöillä muuttujakohtaisia puutteita on pääsääntöisesti hyvin vähän. Siitä huolimatta useimmissa muuttujissa on puuttuvia tietoja. Pulmia tästä voi aiheutua lähinnä silloin, kun tapauksia on melko vähän ja samassa mallissa halutaan tarkastella useita muuttujia, joissa on kohtalaisesti kysymyskohtaista katoa. Jos tällöin menetellään ”normaalisti”, eli hylätään kaikki tapaukset, joilta ainakin yhden tarkasteltavan muuttujan arvo puuttuu, voidaan menettää varsin suuri osa aineistosta, ja lisäksi jäljelle jäävien henkilöiden edustavuus tarkasteltavina olevien ilmiöiden ja riippuvuuksien suhteen saattaa olla tavoiteltua huonompi. Myös puuttuvien selittävien muuttujien arvojen asettamista omaksi luokakseen on käytetty, jolloin voidaan arvioida katotapausten luonnetta, mutta tämä ei välttämättä auta selittämään todellisten muuttujanarvojen vaikutusta vastemuuttujan arvoihin.

Puuttuvien tietojen korvaaminen arvioiduilla tai lasketuilla arvoilla (*imputointi*) on menettelytapa, joka on joissakin analyyseissä käyttökelpoinen. On kuitenkin monia käyttötarkoituksia, joissa puuttuvia tietoja ei pidä korvata: pääsääntöisesti Terveys 2000 –analyyseissä puuttuvia arvoja ei täydennetä imputoimalla vaan kadon vaikutuksia korjataan käyttämällä Tilastokeskuksessa muodostettuja painokertoimia.

Vakiintuneiden summa- ym. yhdistelmämuuttujien, kuten GHQ, Beck ja MMSE, sekä muiden laajempaan käyttöön soveltuvien yhdistelmämuuttujien (esim. päivittäinen tupakointi, paremman korvan kuulokynnysten keskiarvo, sosiaaliryhmä, mahd. jonkinlainen alkoholinkäytön yleiskuvausmuuttuja jne.) laatimisessa pyritään ensisijaisesti soveltamaan mittarin tekijöiden antamia sääntöjä ja toissijaisesti ko. aiheen asiantuntijan/vastuuhenkilön (esim. raportointiryhmän pj tai hänen nimeämänsä henkilö) kanssa laatimaan muuttujamuunnokset siten, että mahdollisimman suuri osa tutkituista saisi jonkin perustellun arvon, ja puuttuvien tietojen osuus jäisi mahdollisimman pieneksi. Tätä työtä - jossa puuttuvien tietojen käsittelyssä yleisimmin sovelletaan loogisia sääntöjä - koordinoi Sirkka Rinne (yhteistyössä muiden, mm. Paul Knektin, Pirkko Alhan ja työvaliokunnan kanssa).

Terveys 2000 –organisaatiolla ei ole riittäviä resursseja tilastollisen tuen tarjoamiseen muissa imputointiongelmassa, joten imputointi ja siihen liittyvät analyysit jäävät ainakin toistaiseksi ensisijaisesti tutkijoiden ja heidän tilastollisten tukihenkilöidensä tehtäväksi. Raportointiryhmiä pyydetään kertomaan puuttuvien tietojen käsittelyyn liittyvistä näkemyksistään, kokemuksistaan ja ehdotuksistaan Tommi Härkäselle. Imputointiin liittyviä ratkaisuja voidaan arvioida uudelleen Terveys 2000 –keskusorganisaatiossa kokemusten karttuessa. Aiheeseen liittyen on valmisteilla opinnäytetyö, jonka tuloksia voidaan odottaa syksyllä 2004.

### **Tilastotieteellisiä näkökohtia**

Tilanteissa, joissa tutkimukseen osallistuneen yksilön muuttujissa on yksittäisiä puuttuvia tietoja, voi olla hyödyllistä täydentää aineistoa puuttuvien arvojen osalta, koska yksikin puuttuva muuttujanarvo aiheuttaa yleensä yksilön jäämisen pois analyysistä. Ongelman kärjistyy esim. muodostettaessa uusi muunnosmuuttuja jonkin muuttujajoukon summana. Joissakin tapauksissa on olemassa standardi tapa käsitellä puuttuvia tietoja (esim. asettamalla puuttuvien havaintojen arvoiksi yksilön havaittujen muuttujanarvojen keskiarvo tai jättää korreloiva muuttuja pois summaindeksiä laskettaessa), mutta usein standardimenettelyä ei ole.

Usein käytetyissä *single imputation* –menetelmissä, joissa muuttujan puuttuvat arvot korvataan esim. keskiarvolla, ongelmana on tulosten tarkkuuden yliarvioiminen, eli p-arvot ovat yleensä liian pieniä ja luottamusvälit kapeita.

Eräs käyttökelpoinen ja yleisesti tunnettu menetelmä on *moni-imputointi* (MI; Rubin 1987, Schafer 1999, ...), jossa analyysi suoritetaan kolmessa vaiheessa:

1. **Muodostetaan 'uusia' havaintoaineistoja imputoimalla**, joissa puuttuvien havaintojen paikalle generoidaan satunnaisesti arvoja ennustejakaumasta, jossa hyödynnetään usein muuttujien välisiä korrelaatioita. Imputointimalliin pitäisi valita periaatteessa kaikki muuttujat, jotka korreloivat puuttuvia havaintoja sisältävien tutkimusmuuttujien kanssa. On turvallisempaa valita imputointimalliin liikaa selittäjiä kuin liian vähän. Jos puuttuvien arvojen lukumäärä on pieni (esim. alle 20 %), usein esim.  $m=5$  aineistoa riittää. Dokumentaatioksi imputoinnin toistamista varten käy ohjelmakoodi ja siinä oleva satunnaislukugeneraattorin siemenluku (SEED)
2. **Suoritetaan halutut analyysit** erikseen jokaisella imputoidulla aineistolla, jolloin tulokseksi saadaan  $m$  kappaletta parametri- ja niiden varianssiestimaattia.
3. **Yhdistetään tulokset**. Esim. regressiomallien tapauksessa regressiokertoimien estimaatti on imputointikohtaisten estimaattien keskiarvo, ja varianssiestimaatti on yksinkertaistettuna imputointikohtaisten varianssiestimaattien keskiarvon ja imputointikohtaisten piste-estimaattien varianssien summa.

MI:n etuina ovat mm. helppo toteutettavuus, eli standardit tilastolliset menetelmät soveltuvat sellaisenaan. Esimerkiksi SAS-ohjelmiston MI- ja MIANALYZE-proseduurit soveltuvat moni-imputoinnin apuvälineiksi. Puuttuviin havaintoihin liittyvä epävarmuus tulee huomioitua estimaattien lisääntyneenä epätarkkuutena. Moni-imputoinnilla voidaan arvioida myös kadon mahdollisen informatiivisuuden (eli kadon todennäköisyys riippuu tutkittavana olevasta ilmiöstä, esim. sairaiden jääminen pois tutkimuksesta voi aiheuttaa merkittävää harhaa tutkittaessa sairauden vallitsevuutta väestössä) vaikutusta tuloksiin herkkyyksianalyysin avulla käyttämällä imputointistrategioina esim. ”katotapaukset ovat yleensä sairaita” tai ”katotapaukset ovat yleensä terveitä”. Ero single- ja multiple imputation –menetelmien välillä riippuu mm. puuttuvan tiedon osuudesta, muuttujien keskinäisten riippuvuuksien voimakkuudesta (vahvat riippuvuudet tuottavat kapeita ennustejakaumia ja siten vähäisempää vaihtelua estimaateissa) ja käytetystä mallista.

Huonona puolena on realistisen imputointimallin valinta, joka yleensä vaatii vahvaa sisältö- ja menetelmäasiantuntemusta, koska moni-imputointi on parhaimmillaan analyysikohtaisena. Kokemusta imputointimenetelmien käytöstä on vain harvoilla tilastotieteilijöillä ja monilla tutkimusaloilla niiden käyttäminen ei ole yleistynyt, joten näiden menetelmien käyttäminen perusteluineen voi vaatia huomattavasti työtä ja aikaa. Satunnainen imputointi aiheuttaa vaihtelua tuloksiin. Vaihtelun merkitys on pieni, mutta raportoitaessa tuloksia useamman numeron tarkkuudella eri analyysiajot voivat antaa eri tuloksia. Tällaisen satunnaisen vaihtelun pienentämiseksi voi ajaa enemmän imputointeja, esim.  $m=50$ .

Eräs mahdollisuus välttää em. satunnaisia eroja tuloksissa ja tutkijoille koituvaa lisävaivaa sopivan imputointimallin valinnassa voisi olla valmiiksi imputoitujen aineistojen muodostaminen, mutta Terveys 2000 –aineiston laajuus tekee keskitetyn imputoinnin mahdottomaksi. Tutkijoiden tulisi konsultoida tilastotieteilijää imputoinnin ja siihen liittyvien toimenpiteiden suorittamiseksi mieluiten erikseen jokaisessa tutkimusongelmassa. Tutkimusryhmiä pyydetään tiedottamaan Tommi Härkäselle mahdollisista imputointiin liittyvistä ratkaisuisistaan.

## **Viitteitä**

Rubin, DB (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley.

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, 8, 3-15.