Kimmo Vehkalahti

Kyselytutkimuksen mittarit ja menetelmät

Vuonna 2014 painettu kirja on Finn Lecturan julkaisema.

Tämä painetun kirjan kanssa saman sisältöinen pdf on
Helsingin yliopiston vuonna 2019 julkaisema.

ISBN painettu: 978-951-792-649-2

ISBN pdf: 978-951-51-4981-7

DOI: 10.31885/9789515149817

Typografia ja taitto: Kimmo Vehkalahti (Survo+LaTeX)

Sisällys

Alkusanat

Olet kirjoittanut uuden oppikirjan. Miten luonnehtisit sitä?

Kirja käsittelee käytännönläheisellä tavalla mittausta ja tilastollisten menetelmien soveltamista kyselytutkimuksessa.

Keitä olet ajatellut kirjasi lukijoiksi?

Esimerkiksi sosiologian, sosiaalipsykologian, kasvatustieteen, psykologian, viestinnän, markkinoinnin ja tilastotieteen opiskelijoita, opettajia, tutkijoita sekä muita asiantuntijoita, niin yliopistoista ja ammattikorkeakouluista kuin yrityksistä ja tutkimuslaitoksista.

Kerro hieman itsestäsi ja kirjan aihepiirin erityisosaamisestasi.

Olen yhteiskuntatieteelliseen mittaukseen erikoistunut ja menetelmistä kiinnostunut tilastotieteilijä. Kirjan aiheet kietoutuvat moniin sellaisiin aloihin ja asioihin, joista joutuisin kyselylomakkeessa valitsemaan kohdan *en osaa sanoa*. Toisaalta menetelmien soveltamisen kannalta osaan sanoa jotain myös useista muista kuin edellä mainituista aloista.

Olen opettanut näitä aiheita kursseilla ja koulutuspäivillä, ohjannut akateemisia opinnäytetöitä ja osallistunut tutkimusprojekteihin sekä neuvonut satoja opiskelijoita, tutkijoita ja opettajia. Myös oma tutkimustyöni tilastotieteen alalla koskee mittareita ja menetelmiä.

Soveltavana tilastotieteilijänä olen kiinnostunut eri alojen tutkimuskysymyksistä ja siitä, miten tilastotiedettä voidaan hyödyntää haettaessa vastauksia näihin kysymyksiin.

Perustele, miksi mainitsemasi aiheet kiinnostaisivat lukijoita. Valaise myös vähän, mitä kaikkea kirjassasi käsittelet. Voit tarvittaessa jatkaa kääntöpuolelle.

Kirjani kuviteltu kohderyhmä on laaja, samoin kuin kirjan aihepiiri. Kyselytutkimus, jonka juuret ovat selvimmin yhteiskuntatieteissä, on nykyään keskeinen tiedonkeruu- ja analysointiväline yhä useammalla alalla. En kata kaikkea, vaan nostan esiin asiantuntemukseni alueelle kuuluvia aiheita, joita ovat mittaus sekä aineiston keruu, hallinta, muokkaus ja analysointi tilastollisilla menetelmillä. Aiheet ovat paljolti alasta riippumattomia. Lisäksi tarkastelen dokumentoivaa työskentelytapaa, jota voisi tuoda enemmän esille pohdittaessa tutkimusprosessien ja -tulosten laatua ja luotettavuutta.

Näkökulmani painottuu siis tilastotieteeseen, mutta kohderyhmäni koostuu enimmäkseen muista kuin tilastotieteilijöistä. Ristiriitaa ei ole, sillä kaikilla aloilla tarvitaan tilastollisia menetelmiä mitattavissa olevan tiedon tiivistämiseen, kuvaamiseen ja mallintamiseen.

Menetelmäosaajista on jatkuvasti pulaa, mutta osaaminen ei saa olla vain tilastotieteilijöiden varassa. Tarvitaan enemmän eri alojen asiantuntijoita, jotka hallitsevat myös tilastollisten menetelmien soveltamisen oman alansa haasteissa. Kuvittelen kirjastani olevan hyötyä sen ymmärtämiseksi, mihin mitäkin menetelmää käytetään ja millaisiin tutkimuskysymyksiin menetelmillä voidaan saada vastauksia, mitä joudutaan olettamaan tai missä tilanteissa ja minkä vuoksi jotakin menetelmää ei pidä soveltaa.

Menetelmien ohella kirjaan punomani teema on mittaus, jota käsitellään usein liian vähän, liian pinnallisesti ja liian irrallaan menetelmistä. Omaksumani lähestymistapa korostaa mittauksen merkitystä ja vaikutuksia läpi kirjan.

Olen halunnut välttää teknisiä yksityiskohtia kuten laskukaavoja, joista menetelmien soveltajille ei yleensä ole paljoakaan hyötyä. Käsin ei ole enää vuosikausiin tarvinnut laskea: tietokoneet huolehtivat siitä rutiininomaisesti. Tutkimustyö ei sen sijaan ole rutiinia; sitä ei voi automatisoida. Pyrin korostamaan tilastollista ajattelua. Haasteita riittää muun muassa siinä, mitä ohjelmistot kannattaa panna tekemään ja miten niiden tulosteita tulkitaan. Toivon kirjastani olevan apua tällaisten kysymysten kanssa painiskelussa.

Kuvaile vielä kirjasi syntytapaa – mielellään lyhyesti.

Minulta on monta kertaa kysytty, onko tästä aihepiiristä hyvää kirjaa. Tyhjentävän vastauksen antaminen on ollut vaikeaa, joten aloin harkita sellaisen kirjoittamista. Ajattelin, että olisi hyödyllistä tiivistää vuosien varrella karttuneita näkemyksiäni kirjalliseen muotoon. Aioin ensin päivittää aiemmin laatimaani monimuuttujamenetelmien monistetta, mutta palautteen ja kommenttien perusteella päätin pian ottaa tavoitteeksi kunnon painotuotteen synnyttämisen. Saatuani työn alulle rohkenin väittää, että kirjani tulisi olemaan hyödyllinen. Toivon, että moni lukijoista olisi kanssani ainakin *osin samaa mieltä*.

Sovelluspainotteisessa lähestymistavassa hyvät esimerkit ovat tärkeitä. Halusin välttää kirjavuutta ja keksin, että kirja voisi rakentua vain yhden tutkimusasetelman varaan, kunhan se olisi riittävän edustava ja monipuolinen.

Olin onnekas, sillä sain käyttööni Maarit Valtarin tekeillä olevan sosiaalipsykologian väitöstutkimuksen kyselyaineiston. Tutkimus käsittelee *suomalaisten naisten suhtautumista omaan ulkonäköönsä*. Aihe on kiinnostava, ja sitä voi tässä lähestyä ilman erityisiä sosiaalipsykologian tietoja.

Kirjan esimerkit ja kuvat pohjautuvat ulkonäkötutkimuksen asetelmaan, sen kyselylomakkeeseen ja aineistoon. Esittämiini sisällöllisiin tulkintoihin on syytä suhtautua varauksellisesti, koska en ole sosiaalipsykologi. Ulkonäkötutkimuksen varsinaiset tulokset on parasta katsastaa aikanaan Maarit Valtarin väitöskirjasta.

Kiitokset

Vaikka kirjan voi kirjoittaa yksin, on mukana joukko ihmisiä, joita ilman työtä ei saisi päätökseen, tuskin edes alulle.

Professori, VTT Lauri Tarkkonen sai kurssillaan minut innostumaan mittauskehikon ulottuvuuksista (Hirvelä & Vehkalahti, 1993).

Omien kurssieni ja koulutuspäivieni osallistujat yllyttivät oppikirjan kirjoittamiseen niin, että se todella tuntui hyvältä idealta.

Tammen kustannuspäällikkö Leena Paunonen tarttui ideaan ja auttoi monin tavoin sen muuntamisessa seoksesta teokseksi.

VTM Maarit Valtari antoi käyttööni väitöskirja-aineistonsa ja kävi viimeistelyvaiheessa käsikirjoitukseni perusteellisesti läpi.

VTM Kati Tiirikainen kommentoi kahtena kesänä keskeneräisiä kehitelmiäni – kappaleita, kuvia, kaavioita. Kiitän kukkasin!

FT, VTM Sirpa Lappalainen seurasi työn edistymistä likeltä ja toi pohdittavakseni useita yhteiskuntatieteilijän kriittisiä näkemyksiä.

Professori, FT Seppo Mustonen esitti arvokkaita kommentteja ja täytti monia Survo-toiveitani. Kirjaa on vaikea kuvitella ilman Survoa.

Dosentti, FT Simo Puntanen perehdytti minua siihen, miten julkaisuja laaditaan, työstetään ja viimeistellään.

Professori, FT Sari Lindblom-Ylänne kannusti kivasti kirjoittamaan lainaamalla kaksi klassikkoa (Jyrinki, 1977; Valkonen, 1981).

VTM Maria Valaste antoi viime vaiheissa hyödyllistä palautetta sekä rakentavia ehdotuksia ja huomautuksia.

LuK Heidi Rand ja VTK Emmi Tikkanen kommentoivat tekstiä työn loppumetreillä ja kysyivät hyviä kysymyksiä.

FT Pekka J. Nieminen ja FM Jarmo Niemelä ratkoivat ystävällisesti LaTeX-ongelmiani asiantuntevilla neuvoillaan.

Perhepiirini merkitys on ollut suuri. Monet kohdat saivat alkunsa kirjoituslomilla Benalmádenassa ja Sastamalassa 2007 ja 2008.

Äitiäni Miiraa kiitän saamastani tuesta ja eräistä lempeistä, mutta tavattoman tarkkanäköisistä huomautuksista.

Isääni Mattia kiitän saamastani tuesta ja klassikosta (Sariola, 1956), joka merkinnöistä päätellen on luettu huolellisesti.

Veljeäni Herkkoa kiitän kadonneeksi luulemani *"Karkkia vain karkkipäivänä"* -elokuvamme esiin kaivamisesta.

Puolisoani Sirpaa kiitän rakkaudesta ja rohkaisusta kirjasavottaan ryhtymisessä ja sen loppuun saattamisessa sekä tukemisesta ja lukemisesta kaikissa työn myötä- ja vastamäissä.

Vuosaaressa 20. syyskuuta 2008

Kimmo Vehkalahti

Kiitokset lukijoille!

Heti ilmestyttyään kirja sai ilahduttavan myönteisen vastaanoton. Sitä käytetään opinnäytteiden lähteenä ja kurssien materiaalina.

Nyt kirja ilmestyy uudelleen Finn Lecturan kustantamana ja tulee taas paremmin saataville. Kansi on uusi, mutta sisältö (sivunumeroita myöten) sama. Uskon kirjasta olevan hyötyä vielä pitkään, sillä tilastollisen tutkimuksen perusasiat eivät kovin nopeasti muutu.

On ollut ilo saada lukijoilta innostuneita yhteydenottoja. Toivotan lisää oivaltavia hetkiä kyselytutkimuksen parissa!

Vuosaaressa 10. maaliskuuta 2014

Kimmo Vehkalahti

Kolmas kerta avoimin ovin!

Kirjaa on hyödynnetty todella monien opinnäytteiden lähteenä! Nyt se tulee avoimesti verkkoon Helsingin yliopiston julkaisemana.

Liitteessä kuvatun SURVO MM:n rinnalle on vakiintunut avoimen lähdekoodin versio Survo R, joka toimii osana R-ohjelmistoa.

Vuosaaressa 13. toukokuuta 2019

Kimmo Vehkalahti

PS. Vinkki jatko-opintoihin: github.com/KimmoVehkalahti/MABS

1 Johdanto

Useimmat meistä ovat vastanneet johonkin kyselyyn. Laajasti käsitettynä kysely kattaa monenlaista toimintaa yksinkertaisista mielipidetiedusteluista laajoihin kyselytutkimuksiin. Kyselyjä tekevät sekä yliopistot, yritykset ja yhteisöt että tiedotusvälineet ja tutkimuslaitokset. Kyselylomakkeeseen voi törmätä yhtä hyvin työssä, kotona tai kadulla kuin kaupassa, ravintolassa tai verkossa.

Kaikkia kyselyjä ei voi pitää tutkimuksena, mutta tässä ei lähdetä tarkemmin rajaamaan, mikä on tutkimusta ja mikä ei. Kirjassa esitettyjä asioita voi hyödyntää kyselyn ”tutkimuksellisuuden” asteesta riippumatta, esimerkiksi palautelomaketta suunniteltaessa.

1.1 Kyselytutkimus

Kyselytutkimus on tärkeä tapa kerätä ja tarkastella tietoa muun muassa erilaisista yhteiskunnan ilmiöistä, ihmisten toiminnasta, mielipiteistä, asenteista ja arvoista. Tämäntyyppiset kiinnostuksen kohteet ovat sekä moniulotteisia että monimutkaisia.

Kyselytutkimuksessa tutkija esittää vastaajalle kysymyksiä kyselylomakkeen välityksellä. Kyselylomake on mittausväline, jonka sovellusalue ulottuu yhteiskunta- ja käyttäytymistieteellisestä tutkimuksesta mielipidetiedusteluihin, katukyselyihin, soveltuvuustesteihin ja palautemittauksiin.

Haastattelututkimuksessa tutkija tai haastattelija esittää kysymyksiä suoraan vastaajalle, esimerkiksi puhelimitse tai kasvotusten. Haastattelulomake muistuttaa kyselylomaketta. Erona on se, että kyselylomakkeen on toimittava omillaan, ilman haastattelijan apua.

Englanninkielinen termi *survey* kattaa sekä kysely- että haastattelututkimuksen, mutta valitettavasti sanalle ei ole vakiintunutta suomennosta. Toisinaan käytetty sana lomaketutkimus kuulostaa liian viralliselta ja hieman kuivalta, ikään kuin tutkittaisiin lomakkeita. Kysely ja haastattelu viittaavat paremmin tutkimustyössä tarvittavaan uteliaisuuteen.

Jatkossa puhutaan kyselytutkimuksesta, sillä tämän kirjan näkökulmasta kysely ja haastattelu eivät eroa käytännössä lainkaan. Kirjan näkökulmaan viittaavat sen nimessä esiintyvät mittarit ja menetelmät, jotka yhdistävät kaikkia kyselytutkimuksia. Juuri mittarien laatimisessa ja menetelmissä tilastotiede kietoutuu kiinnostavalla tavalla sisällöllisiin kysymyksenasetteluihin ja tulkintoihin.

Mittarit

Mielipiteiden, asenteiden ja arvojen tutkiminen ei ole helppoa. Haasteita aiheuttavat lukuisat epävarmuudet: edustivatko kyselyyn osallistuneet tutkimuksen perusjoukkoa, saatiinko tarpeeksi vastauksia, oliko kysymyksiin vastattu riittävän kattavasti, mittasivatko kysymykset tutkittavia asioita, toimivatko mittarit luotettavasti, oliko kyselyn ajankohta hyvä ja niin edelleen. Osa haasteista liittyy tiedonkeruuseen, osa mittaamiseen ja osa tutkimuksen sisällöllisiin tavoitteisiin. Eniten huomiota tässä kirjassa saavat mittaamista tai mittareita koskevat tilastolliset näkökohdat.

Kyselytutkimuksessa mittarilla tarkoitetaan kysymysten ja väitteiden kokoelmaa, jolla pyritään mittaamaan erilaisia moniulotteisia ilmiöitä kuten asenteita tai arvoja. Mittareita voidaan rakentaa itse tai soveltaa aiemmin käytettyjä, ”valmiita” mittareita. Valmiisiin mittareihin on syytä suhtautua jossain määrin varauksellisesti, sillä niiden toimivuus toisessa yhteydessä ei ole itsestäänselvyys. Mitattavat ilmiötkään eivät yleensä ole kovin vakaita; ne voivat muuttua ajan kuluessa tai ilmetä eri ympäristöissä eri tavalla.

Mittarien laatiminen tapahtuu parhaimmillaan sisällön tuntevan tutkijan ja soveltavan tilastotieteilijän yhteistyönä. Tässä kirjassa korostuvat tilastotieteilijän näkemykset, mutta käytännössä tutkija on avainasemassa mittareiden määrittelyssä. Tilastotieteilijähän ei voi tietää, mitä pitäisi mitata, mutta voi auttaa siinä, miten mittaaminen kannattaisi suorittaa.

Menetelmät

Kyselytutkimus on enimmäkseen määrällistä tutkimusta, jossa sovelletaan tilastollisia menetelmiä. Kyselyaineistot koostuvat pääosin mitatuista luvuista ja numeroista, sillä vaikka kysymykset esitetään sanallisesti, niin vastaukset ilmaistaan numeerisesti. Sanallisesti annetaan täydentäviä tietoja tai vastauksia kysymyksiin, joiden esittäminen numeroina olisi epäkäytännöllistä.

Usein sanotaan, että määrällisellä tutkimusotteella tavoitellaan yleiskäsityksiä ja laadullisilla menetelmillä pureudutaan yksityiskohtiin, mutta ei tutkimusote kaikkea ratkaise. Myös tilastollisilla menetelmillä päästään käsiksi yksityiskohtiin. Samassa tutkimuksessa saatetaan hyödyntää molempia lähestymistapoja. Sanallisia vastauksia voi olla antoisampaa analysoida laadullisilla menetelmillä, mutta saatuja tuloksia voi tiivistäen esittää määrällisillä menetelmillä. Olennaisinta on, että osaa valita tarkoituksenmukaiset lähestymistavat sen ilmiön tutkimiseen, josta on kiinnostunut.

Aineiston analysointi ei ole mekaanista käsittelyä, sillä jos se sitä olisi, ei tällaisia kirjoja tarvittaisi. Vaikka eri työvaiheita voidaan ja on syytäkin automatisoida, on menetelmien soveltaminen ja tulkinta paljolti käsityötä, joka edellyttää myös ohjelmistojen ja järkevien työskentelytapojen omaksumista.

Mittauskehikko

Mittareiden ja menetelmien tarkastelun yhdistää käytännön kannalta hyödyllisellä tavalla *mittauskehikko*, jonka varaan kirja merkittävästi rakentuu. Kehikkoon viitataan alustavasti jo luvussa 2, mutta kokonaan se nähdään vasta kuvassa 5.1 (s. 122). Kehikon eri osiin syvennytään vaiheittain luvuissa 4 ja 5.

Mittauskehikon on alun perin esittänyt Lauri Tarkkonen väitöskirjassaan *On Reliability of Composite Scales* (Tarkkonen, 1987). Ideaa ovat soveltaneet käytäntöön useiden eri alojen tutkijat. Mittauskehikon ja sen sovellusten teoriaperusteista kiinnostuneiden kannattaa tutustua aihepiirin viimeaikaisiin julkaisuihin kuten Vehkalahti, Puntanen & Tarkkonen (2008, 2007, 2006); Valaste, Vehkalahti & Tarkkonen (2008); Tarkkonen & Vehkalahti (2005) tai Vehkalahti (2000). Käytännön soveltamisen kannalta se ei kuitenkaan ole välttämätöntä. Tässä kirjassa kehikon matemaattisiin yksityiskohtiin ei perehdytä.

1.2 Kirjan rakenne ja sisältö

Rakenteellisesti kirja seuraa kyselytutkimuksille yhteisiä vaiheita. Sisällöllisesti se yhdistää kaksi kokonaisuutta:

- mittauksen tilastolliset näkökohdat kyselytutkimuksessa
- kyselyaineiston tilastollinen mallintaminen ja kuvailu.

Olennaista käsiteltävien asioiden ymmärtämisessä ovat erilaiset tulosteet ja kuvat, joita kirjassa on runsaasti. Laskukaavoja ei ole lainkaan, sillä niistä ei yleensä ole menetelmien soveltajille hyötyä. Tietokoneet huolehtivat laskemisesta; tutkijalle kuuluvat menetelmien ja ohjelmistojen hallinta sekä lukujen ja numeroiden sanallinen tulkinta.

Luvut 2 ja 3 johdattelevat kyselytutkimuksen aihepiiriin ja luovat pohjan myöhemmissä luvuissa käsiteltäville malleille ja menetelmille. Luvussa 2 perehdytään kyselylomakkeen suunnitteluun mittauksen ja tiedonkeruun kannalta, pohjustetaan alustavasti mittauksen mallintamista sekä tarkastellaan lyhyesti, miten kyselyaineisto syntyy. Luvussa 3 käydään läpi välttämättömät aineiston esikäsittelyvaiheet, joiden kuluessa piirretään kuvia, tuotetaan taulukoita, tutkitaan tunnuslukuja sekä tehdään muunnoksia ja muokkauksia, joilla aineistoa valmistellaan varsinaisiin analyyseihin.

Luvut 4 ja 5 käsittelevät kaikissa kyselytutkimuksissa tarvittavaa aineiston tiivistämistä ja havaintojen vertailua. Aiheita yhdistää mittauskehikko, johon perehdytään vaiheittain. Aluksi keskitytään luvussa 2 pohjustettuun mittausmalliin, jonka avulla pureudutaan tutkittavan ilmiön ulottuvuuksiin. Mallin perusteella aineistoa tiivistetään mitta-asteikoiksi ja tutkitaan niiden ominaisuuksia. Kokonaisuudessaan mittauskehikko esitetään luvussa 5, jossa huomio kääntyy tiivistetyn aineiston havaintoihin. Luvut 4 ja 5 ovat sisällöltään tärkeimmät ja samalla vaativimmat. Mittauskehikon ohella esille tulevat myös keskeisimmät menetelmät, faktorianalyysi ja regressioanalyysi.

Luvuissa 6 ja 7 ryhmitellään aineistoa ja visualisoidaan ryhmiä erilaisilla monimuuttujamenetelmillä. Menetelmiä ei käsitellä yhtä yksityiskohtaisesti kuin kahdessa aiemmassa luvussa. Tarkastelut ovat esimerkkejä siitä, miten aineiston analysointia voidaan jatkaa syvemmälle.

Liitteessä A kuvataan lyhyesti kirjassa käytettyjä Survo- ja SPSS-ohjelmistoja sekä tarkastellaan kuvien ja tulosteiden työkaavioita esimerkkeinä dokumentoivasta työskentelytavasta. Työkaavioihin viitataan luvuista 3–7 marginaaliin sijoitetulla sivunumerolla. Sekä tekstissä että tulosteissa on yhdenmukaisuuden vuoksi käytetty desimaalierottimena pilkun sijasta pistettä.

s. 208

Ulkonäkötutkimus

Sekä rakenteellisesti että sisällöllisesti tärkeä kokonaisuus muodostuu Maarit Valtarin sosiaalipsykologian väitöstutkimukseen perustuvista esimerkeistä. Suomalaisten naisten suhtautumista omaan ulkonäköönsä luotaava, tekeillä oleva tutkimus on edustava esimerkki kyselytutkimuksesta. Sen kyselylomakkeessa on hyödynnetty useita erilaisia mittareita, ja tiedonkeruu on toteutettu rekisteripohjaisena otantana kahtena eri ajankohtana, vuosina 1997 ja 2005.

Ulkonäkötutkimuksen pohjalta laadittujen esimerkkien tarkoitus on kuvata mittareiden ja menetelmien soveltamisen tilastollisia näkökohtia. Sisällöllisiin näkökohtiin on tässä kirjassa syytä suhtautua tilastotieteilijän mielikuvituksen tuotteina.

Muita menetelmäkirjoja

Menetelmäkirjoja on julkaistu Suomessa yli 50 vuoden ajan, pääosin yhteiskunta- ja käyttäytymistieteiden piirissä. Sellaiset teokset kuten *Sosiaalitutkimuksen menetelmät* (Sariola, 1956), *Psykometriikan metodeja II* (Vahervuo, 1956), *Johdatus faktorianalyysiin* (Vahervuo & Ahmavaara, 1958), *Sosiologian tutkimusmenetelmät 2* (Eskola, 1968), *Haastattelu- ja kyselyaineiston analyysi sosiaalitutkimuksessa* (Valkonen, 1981) sekä *Kysely ja haastattelu tutkimuksessa* (Jyrinki, 1977), joiden ensipainokset ilmestyivät vuosina 1956–1974, ovat toimineet opinnäytetöiden lähteinä vielä 2000-luvulla.

Vanhoista teoksista osa on painunut unholaan, osa saavuttanut jo aikoinaan ”klassikon” aseman. Kirjoissa esitetyt tutkimuksen teon periaatteet pätevät edelleen monelta osin, mutta kun miltei kaikki tekniset ratkaisut ovat kauan olleet historiaa, on teosten käytännön hyöty vähentynyt.

Uudemmissakin oppikirjoissa käsitellään usein koko tutkimusprosessia suunnittelusta raportointiin. Tietokoneen käyttö nähdään olennaisena osana työskentelyä, mutta käsin laskeminen ja laskukaavat kulkevat yhä mukana yllättävän monissa teoksissa.

Sosiaalitutkimuksen kvantitatiiviset menetelmät (Alkula, Pöntinen & Ylöstalo, 1994) kuvaa tutkimuksen suunnittelua, aineiston hankintaa ja mittausta; analysoinnin osalta erityisesti ristiintaulukointia ja eräitä yleisimpiä tilastollisia menetelmiä. Kirjassa on SPSS-ohjelmistoon perustuvia esimerkkejä, mutta sen tärkein anti on menetelmien perusteiden selvittäminen ja tulosten tulkinta sosiaalitieteiden näkökulmasta. *Tutki ja mittaa* (Vilkka, 2007) tiivistää sanallisesti määrällisen tutkimuksen perusteita suunnittelusta raportointiin unohtamatta tutkimusetiikkaa.

Tilastolliset monimuuttujamenetelmät (Mustonen, 1995) on perusteellinen esitys monimuuttujamenetelmistä. Vaikka kirja on monilta osin teoreettisesti vaativa, se sisältää myös runsaasti tulosten tulkinnan pohdiskelua. Survo-ohjelmisto (Mustonen, 2001, 1992) on keskeisellä sijalla sekä käytännön esimerkeissä että teoreettisemmissa tarkasteluissa. Myös *Tutkimusaineiston analyysi* (Nummenmaa, Konttinen, Kuusinen & Leskinen, 1997) on matemaattisesti melko vaativa teos, joka perusmenetelmien lisäksi sisältää pidemmälle meneviä mallintamismenettelyjä ja mittaamisen teoriatarkasteluja. Kirjaan sisältyy muun muassa Survo- ja SPSS-esimerkkejä.

Tilastollinen tutkimus (Heikkilä, 2004) kattaa kyselytutkimuksen tiedonkeruun ja mittaamisen, aineiston kuvaamisen, eräiden perusmenetelmien ja testien ohella myös tilastotieteen ja todennäköisyyslaskennan perusteita. Kirjassa työstetään asioita laskukaavojen lisäksi SPSS:llä ja Excelillä. Aihepiiriltään samantapainen teos on *Käyttäytymistieteiden tilastolliset menetelmät* (Nummenmaa, 2004), jossa aineiston käsittelyyn, kuvien piirtoon ja menetelmiin tutustutaan käymällä läpi SPSS:n valikkoja ja tulostuksia. Kirjassa esiintyy melko runsaasti laskukaavoja.

Menetelmien perusteista löytyy myös paljon materiaalia verkosta. Esimerkiksi *Menetelmäopetuksen tietovaranto* (Yhteiskuntatieteellinen tietoarkisto, 2008) ja *Tilastokeskuksen verkkokoulu* (Tilastokeskus, 2006) ovat tutustumisen arvoisia sivustoja.

2 Mittaus ja tiedonkeruu

Määrällisen tutkimuksen peruskivi on mittaus, sillä asioiden tutkiminen tilastollisesti edellyttää, että tietoja voidaan mitata erilaisilla mittareilla. Kyselytutkimuksessa mittarit koostuvat kysymyksistä ja väitteistä, joiden laatimiseen liittyy sekä sisällöllisiä että tilastollisia haasteita. Mittaus tapahtuu kyselylomakkeella, joka on kokoelma mittareita ja yksittäisiä kysymyksiä.

Kyselytutkimuksen kohteet, kuten mielipiteet, asenteet ja arvot, ovat moniulotteisia ja usein myös monimutkaisia, eikä niiden mittaus ei ole aivan yksinkertaista. Mittausvaiheeseen kannattaa panostaa, sillä siinä tehtyjä virheitä ei voi korjata millään analyysimenetelmillä. Tehdyt ratkaisut vaikuttavat myös menetelmien valintamahdollisuuksiin sekä tutkimuksesta tehtävien johtopäätösten luotettavuuteen.

2.1 Johdatteleva esimerkki

Ajatellaan, että haluttaisiin tutkia satunnaisen lenkkeilijän sydämen sykettä, juoksunopeutta ja elämänasennetta. Helppoa, eikö niin? Ei tarvita kuin sykemittari, nopeusmittari ja – asennemittari.

Syke, nopeus ja asenne

Sopivilla laitteilla sykkeen ja nopeuden saa reaaliajassa rannekellosta. Asennemittaus ei sen sijaan onnistu edes päähän asennettavilla antureilla. Siihen tarvitaan asennemittari, joka koostuu joukosta elämänasennetta mittaavia kysymyksiä tai väitteitä. Lisäksi lenkkeilijä olisi vielä saatava suostuteltua vastaamaan niihin.

Sykettä ja nopeutta voidaan mitata suoraan mittalaitteilla, sillä ne ovat ymmärrettäviä, hyvin määriteltävissä olevia käsitteitä. Syke on helppo ilmaista sydämenlyöntien määränä minuutissa ja nopeus esimerkiksi kilometreinä tunnissa tai metreinä sekunnissa.

Asenteita ei voida mitata eikä edes ilmaista yhtä helposti, koska ne ovat käsitteellisesti vaikeatajuisempia ja hankalampia määritellä. ”Lenkkeilijän syke oli 165 lyöntiä minuutissa ja nopeus 11 kilometriä tunnissa” kuulostaa ihan vauhdikkaalta, mutta ”hänen elämänasenteensa oli 7” ei kerrokaan mitään, koska asenteille ei ole yleisesti käytettyä yksikköä tai asteikkoa.

Käsitteiden operationalisointi

On tärkeää erottaa eritasoiset käsitteet toisistaan. Kyselytutkimuksessa kiinnostuksen kohteet ovat yleensä abstrakteja, kuten asenteita tai arvoja, mutta niiden mittaus edellyttää konkreettisia kysymyksiä tai väitteitä. Käsitteet on operationalisoitava, toisin sanoen työstettävä ymmärrettävään ja mitattavaan muotoon.

Äskeistä esimerkkiä voidaan kehitellä myös toiseen suuntaan siirtymällä konkreettisista mittauksista kohti abstraktimpia käsitteitä. Sinällään syke ja nopeus kiinnostaisivat luultavasti vain lenkkeilijää itseään. Tutkijaa saattaisi ennemminkin kiinnostaa lenkkeilijän fyysinen suorituskyky, jonka eräitä osoittimia tai indikaattoreita syke ja juoksunopeus voisivat olla.

Fyysinen suorituskyky on käsitteenä jo huomattavasti abstraktimpi, eikä yhtä helposti määriteltävissä, mutta sen takia se onkin tutkimuskohteena kiinnostavampi kuin pelkkä yksittäinen indikaattori. Jotta päästäisiin takaisin mittaustasolle, pitää miettiä, mistä kaikesta fyysinen suorituskyky voisi koostua.

Ilmiöiden moniulotteisuus

Äkkiä alkaa hahmottua reaalimaailman ilmiöille yhteinen piirre: moniulotteisuus. Fyysinen suorituskykykään ei koostu vain yhdestä ulottuvuudesta kuten nopeudesta, vaan muita aivan ilmeisiä osatekijöitä ovat ainakin voima ja kestävyys. Ilmiö on siis vähintäänkin kolmiulotteinen. Lisäksi mieleen tulee koko joukko muita tekijöitä, joilla lienee oma vaikutuksensa fyysiseen suorituskykyyn: ikä, sukupuoli, pituus, paino, elämäntavat ja kenties jonkinlainen suoritusmotivaatio.

Näin tulee abstrakti käsite purettua useampaan palaseen, joista osa on helposti mitattavissa. Esimerkiksi ikä ja sukupuoli ovat tällaisia, ja ne kuuluvatkin luonnostaan myös useimpien kyselytutkimusten taustatekijöihin. Sen sijaan suorituskyvyn edellä mainitut ulottuvuudet nopeus, voima ja kestävyys ovat sellaisenaan liian abstraktilla tasolla. Ne pitää siis operationalisoida tarkemmin.

Lisäksi on tullut hahmotettua suorituskykyä abstraktimpiakin tekijöitä kuten elämäntavat ja suoritusmotivaatio. Myös nämä ovat kumpikin moniulotteisia yläkäsitteitä, joita on purettava, jotta niitä pystyy mittaamaan. Elämäntapojen mittariin voisi sisältyä kysymyksiä alkoholin käytöstä, tupakoinnista, ravitsemustottumuksista ja ajankäytöstä. Käytännössä mittarien sisällöt määräytyvät tutkijan asiantuntemuksen ja aiempien tutkimusten perusteella; tässä esitetyt ovat yleistiedon pohjalta tehtyjä arvauksia.

Ilmiöiden mitattavuus

Pohditaan nyt vaikkapa nopeutta hieman tarkemmin. Kuten aiemmin todettiin, nopeutta voidaan helposti mitata, mutta on määriteltävä tarkemmin, mitä oikeastaan mitataan. Valitaanko aamulenkin maksiminopeus vai keskinopeus? Kumpikaan ei taida riittää, sillä seuraavaksi pitäisi määritellä, mitä aamulenkki tarkoittaa.

Jotta voidaan operoida käsitteellä nopeus, pitää määritellä tarkemmin, miten sitä mitataan. Yksi mahdollisuus olisi laittaa lenkkeilijä juoksemaan 100 metriä ja mitata siihen kulunut aika. Entä kun olosuhteet, kuten lämpötila tai tuulen nopeus, vaihtelevat? Ei ihme, että tällaisia mittauksia pyritäänkin tekemään laboratorio-olosuhteissa, joissa erilaiset tekijät pystytään vakioimaan. Kenties lenkkeilijä pitäisikin saada testattavaksi jollekin urheiluopistolle.

Voima ja kestävyys ovat vastaavanlaisia, eli on mietittävä, miten niitä voisi mitata. Urheilulajeista tulisi mieleen painonnosto, kiekonheitto ja kuulantyöntö sekä pidemmät juoksumatkat, pyöräily ja uinti. Kymmenottelu on tässä suhteessa erityisen mielenkiintoinen, sillä sen voi ajatella mittaavan sekä kaikkia näitä ulottuvuuksia että muitakin, vaikkapa heittolajeissa tarvittavia, erityistaitoja.

Yhteys kyselytutkimukseen?

Kuten havaitaan, ei fyysisen suorituskyvyn mittaus ole ihan yksinkertaista. Tyypilliset kyselytutkimuksen kohteet kuten asenteet ja arvot ovat ehkä vielä haastavampia. Niiden mittaus edellyttää samantapaista työstämistä kuin äskeisessä esimerkissä.

2.2 Kyselylomake mittausvälineenä

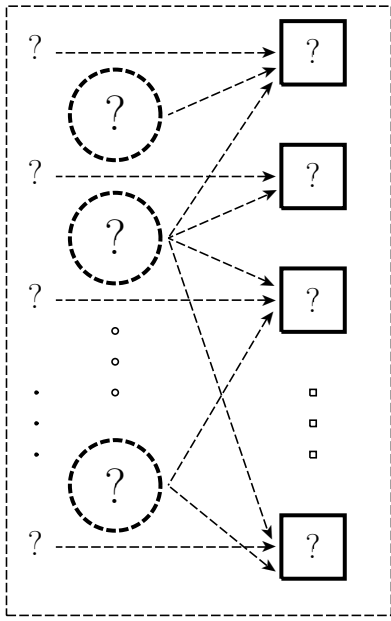
Kyselytutkimuksessa mittaus tapahtuu kyselylomakkeella. Kun vastaaja täyttää lomakkeen, on siihen enää myöhäistä tehdä muutoksia, joten lomake on todella syytä suunnitella huolellisesti. Koko tutkimuksen onnistuminen riippuu mitä suurimmassa määrin lomakkeesta. Ratkaisevaa on se, kysytäänkö sisällöllisesti oikeita kysymyksiä tilastollisesti mielekkäällä tavalla. Kumpikaan ei yksin riitä. Hyvä kyselylomake on kokonaisuus, jossa toteutuvat sekä sisällölliset että tilastolliset näkökohdat.

Tässä luvussa kyselylomaketta tarkastellaan mittausvälineenä korostaen erityisesti tilastollisia näkökohtia. Tiedonkeruu tuo kokonaisuuteen muitakin asioita, joita käsitellään kohdassa 2.6 (s. 47).

2.2.1 Ulottuvuudet

Kuten johdattelevassa esimerkissä todettiin, useimmat ilmiöt ovat moniulotteisia. Aivan aluksi onkin tärkeää hahmottaa kiinnostuksen kohteena olevan ilmiön keskeiset ulottuvuudet. Mitä enemmän on käytettävissä tutkimusalan tunnettua teoriaa, sitä selvemmin ulottuvuudet saadaan johdettua suoraan teoriassa määritellyistä käsitteistä. Tuntemattomammilla alueilla luovittaessa on oltava valmiina havaitsemaan ja tunnistamaan myös täysin uusia ulottuvuuksia.

Ulottuvuuksien ja niiden mittauksen pohtimista auttaa kuvan 2.1 tyyppisten hahmotelmien piirtäminen. Kysymyksessä on mittausmalli, jota Lauri Tarkkonen on kursseillaan kutsunut ”tutkijan kotitehtäväksi”. Kaavioita piirrettyään on moni tutkija kertonut itsekin oivaltaneensa oman tutkimusasetelmansa selkeämmin. Katsotaan nyt, mitä tuo kotitehtävä oikein pitää sisällään.



Kuva 2.1. Mittausmalli alkutekijöissään.

Mittausmallin alkupalat

Myöhemmin tullaan näkemään, kuinka mittausmalli muodostaa osan yleisempää mittareiden ja menetelmien kehikkoa, mutta toistaiseksi kuva 2.1 on pelkkä hahmotelma täynnä arvoituksellisia kysymysmerkkejä. Kuvio selkiytyy sitä mukaa kuin asioille saadaan nimiä.

Isoimpia mysteereitä ovat ilmiön keskeiset ulottuvuudet, joita kuvassa symboloivat ympyrät. Näistä tutkijan pitää aloittaa ja olla ainakin jollain tavoin selvillä ulottuvuuksien lukumäärästä ja nimistä. Tässä kohtaa ei tilastotieteilijä voi paljoa auttaa, sillä kysymysmerkit ympyröiden sisällä vaativat vastauksia perustavaa laatua oleviin kysymyksiin kuten ”Mitä tutkitaan?” ja ”Mistä tutkittava ilmiö koostuu?”. Fyysisen suorituskyvyn esimerkissä ulottuvuuksia tulisi heti kärkeen kolme: nopeus, voima ja kestävyys. Tärkeintä on päästä alkuun; kaaviota voi sen jälkeen helposti täydentää.

Seuraavaksi tullaan suoraan käytännön mittauksen tasolle, kysymyksiin ja väitteisiin, joita kutsutaan *osioiksi*. Niitä symboloivat kuvan 2.1 oikeassa reunassa olevat neliöt, joita olisi oltava ainakin muutama jokaista mitattavaa ulottuvuutta kohti. Kyselylomakkeen osioiden laatimista käsitellään tuonnempana, mutta on hyvä muistaa, että mittausmalli toimii muissakin kuin kyselytutkimuksissa. Fyysisen suorituskyvyn mittauksessa neliöihin voisi sijoittaa esimerkiksi sadan metrin juoksun, kuulantyönnön ja muita kymmenottelun lajeja.

Mittausmallissa ulottuvuuksien ajatellaan vaikuttavan mittauksiin. Esimerkiksi suorituskyvyn ulottuvuuksista nopeus vaikuttaa siihen, miten pärjää sadan metrin juoksussa, voima puolestaan kuulantyönnössä onnistumiseen. Aivan vastaavasti asenteiden tai arvojen ajatellaan vaikuttavan siihen, miten vastaaja lomakkeen äärellä reagoi hänelle esitettyihin kysymyksiin tai väitteisiin. Siksi mittausmallin nuolet osoittavat ulottuvuuksista osioihin. Kuvan 2.1 vinonuolet ovat vain ”suuntaa-antavia”, eihän niitä oikeastaan voi piirtää, ennen kuin kysymysmerkkien tilalle on laitettu oikeita nimiä.

Ajatusrakennelma ja perimmäiset kysymykset

Kuvassa 2.1 neliöinä esitetyt osiot ovat mallin ainoat vastaajalle näkyvät osat. Kaikki muu edustaa ajatusrakennelmaa, jonka tutkija pystyttää tutkimuksen suunnitteluvaiheessa. Mitä enemmän on käytettävissä ilmiötä koskevaa teoriaa tai muuta aiempiin tutkimuksiin perustuvaa tietoa, sitä vankempi tämä rakennelma tulee olemaan. Tärkeää on hahmottaa, että ensisijaisena kiinnostuksen kohteena eivät ole osiot sinänsä vaan ulottuvuudet, joita niillä pyritään mittaamaan. Osiot ovat vain mittausvälineitä, joita voi analyysien perusteella kehittää paremmaksi, tosin vasta seuraavalla mittauskerralla.

Ulottuvuudet ja osiot vastaavat perimmäisiin kysymyksiin siitä, mitä mitataan ja miten. Jäljellä on enää jokaisen osion taustalla häilyvä joukko epävarmuuksia, joista osa liittyy mittaukseen. Kuvan vaakasuorat nuolet viittaavat erityisesti *mittausvirheisiin*, joita ei voi välttää, mutta joiden vaikutuksia voi vähentää. Näissä asioissa tilastotieteilijästä voi olla apua, ja asioihin palataankin vielä monessa kohtaa kirjaa. Kaavion piirtämisen helpottamiseksi mittausvirheet voi jättää pois, kunhan muistaa, että ne ovat mukana riippumatta siitä, onko niitä piirretty vai ei.

Ulkonäkötutkimuksen ulottuvuudet

Maarit Valtarin väitöstutkimus mahdollistaa vastaamisen monenlaisiin tutkimuskysymyksiin. Tässä kirjassa rajoitutaan melko yksinkertaisiin kysymyksenasetteluihin, joissa suomalaisten naisten ulkonäkökäsityksiä tarkastellaan kolmen ulottuvuuden kannalta. Ulottuvuudet ovat 1) *itsetunto ulkonäköasioissa*, 2) *panostaminen ulkonäköön* ja 3) *sosiaaliset ulkonäköpaineet*.

2.2.2 Osiot ja mittarit

Osiolla tarkoitetaan siis yksittäistä kysymystä tai väitettä, joka lähtökohtaisesti mittaa vain yhtä asiaa. *Mittari* on osioista koostuva kokonaisuus, joka mittaa useita, jollain tavoin toisiinsa liittyviä asioita.

Kokonaisuuden kannalta tärkeintä on osioiden sisältö ja se, mitä ulottuvuuksia niillä pyritään mittaamaan. Näihin asioihin ei tässä kirjassa ole mahdollista syvemmin puuttua, sillä ne riippuvat alasta, aiheesta ja tutkimuskysymyksistä. Mittareiden sisällöllinen puoli tai tieto siitä, mitä pitäisi ylipäätään tutkia, ei yleensä ole tilastotieteilijän vastuulla. Hyviä neuvoja antavat johdannossa mainitut Alkula ym. (1994, 130–139) ja Heikkilä (2004, 47–66). Myös Jyrinki (1977, 41–102) on tässä suhteessa yhä kelpo luettavaa, jos kirjan vain jostain saa käsiinsä. Tutustumisen arvoinen ja helppolukuinen teos on *Improving Survey Questions: Design and Evaluation* (Fowler, Jr., 1995).

Tässä kirjassa painottuvat osioiden mittaukseen ja tilastolliseen analysointiin liittyvät näkökohdat, jotka ovat hyödyllisiä alasta riippumatta. Joitakin yleisiä periaatteita on joka tapauksessa hyvä noudattaa. Osioiden on oltava selkeitä, ytimekkäitä ja ymmärrettäviä. On vältettävä monimutkaisia sanamuotoja ja käsitteitä, samoin kuin sanoja *ja*, *sekä*, *sekä–että*, *tai* ja *eli*, sillä ne aiheuttavat monikäsitteisyyksiä. Kysymykseen tai väitteeseen on mahdoton ottaa yksikäsitteisesti kantaa, jos se sisältää samanaikaisesti useita asioita.

Millainen on huono väite?

Esimerkki äärimmäisen huonosta väitteestä voisi olla ”*Radiossa ja tv:ssä on osaavia toimittajia ja kuvaajia*”. Tässä keksityssä väitteessä on useampiakin ongelmia. Vastauksista ei pystyisi päättelemään, onko vastaaja tarkoittanut radiotoimittajia, tv-kuvaajia vai joitain lukuisista

muista mahdollisista yhdistelmistä. Väite on myös liian yleisluonteinen, vaikka sen kohderyhmää tarkennettaisiin. Sanamuotojen on oltava yksityiskohtaisempia ja selkeämpiä. Epäselvät kysymykset ja väitteet saavat vastaajan turhautumaan ja pahimmassa tapauksessa jättämään koko kyselyn kesken.

Äskeisen kaltaiset väitteet saattavat paljastaa, millaisista asioista ollaan kiinnostuneita, mutta mittareiden rakennusaineiksi ne ovat raakileita. Väitteiden ja mittareiden rakentaminen pitäisi malttaa aloittaa miettimällä ilmiön ulottuvuuksia ja purkamalla niitä sopiviin, mitattavissa oleviin osiin. Tässä mittausmallikaavio (ks. kuva 2.1, s. 21) on avuksi. On turha yrittää puristaa kiinnostavaa kysymystä suoraan yhdeksi väitteeksi ja toivoa, että vastaajat ymmärtäisivät, mitä tutkijalla on ollut mielessä.

Ulkonäkötutkimuksen osiot ja mittarit

Edellä mainittuja ulkonäkökäsitysten ulottuvuuksia mitataan kolmella mittarilla, joista ensimmäiseen sisältyy 22, toiseen 20 ja kolmanteen 11 osiota. Mittausmallin oletetaan siis koostuvan kolmesta ulottuvuudesta, joita mitataan 53 osiolla.

2.2.3 Avoimet ja suljetut osiot

Osioita voidaan kutsua avoimiksi tai suljetuiksi. Avoimeen osioon vastataan vapaamuotoisesti, kun taas suljetun osion vastausvaihtoehdot on annettu valmiiksi lomakkeessa. Valmiiden vaihtoehtojen on oltava toisensa poissulkevia, toisin sanoen ne eivät saa mennä päällekkäin.

Esimerkissä 2.1 ensimmäinen kysymys on avoin ja toinen suljettu. Jälkimmäisen vaihtoehdot ovat toisensa poissulkevia, joten ei pitäisi syntyä epäselvyyksiä siitä, mikä vaihtoehto tulisi valita. Alle 18-vuotiaat ja yli 74-vuotiaat eivät tässä tapauksessa kuulu tutkimuksen kohderyhmään, eivätkä he siksi sisälly myöskään vaihtoehtoihin.

Periaatteessa avoimella kysymyksellä saataisiin tieto vuoden tarkkuudella. Ikä kuuluu kuitenkin niihin asioihin, joita ei useinkaan pidä kysyä suoraan, ei naisilta eikä miehiltä. Kysymys voi tuntua vastaajasta epäkohteliaalta, ja vastaukset voivat olla yllättävän epäluotettavia.

Esimerkki 2.1. Kaksi tapaa kysyä vastaajan ikää.

```
Minkä ikäinen olette? ______ vuotta

Minkä ikäinen olette? 18-29 / 30-39 / 40-49 / 50-59 / 60-74 vuotta
```

Suljettu kysymys, jossa on valmiit ikäluokat, on vähän parempi tapa. Näin mittaustarkkuus on karkea, tässä vain kymmenen vuoden luokkaa, mutta se ei haittaa, jos se vain tutkimuksen kannalta riittää. Käytettävä luokittelu on mietittivä tarkoin etukäteen, koska jälkikäteen sitä ei voi muuttaa – korkeintaan tiivistää vielä karkeammaksi. Monesti kannattaa soveltaa aiemmissa tutkimuksissa käytettyjä luokituksia, jolloin voi helpommin vertailla tutkimuksia toisiinsa.

Luotettavimmin ikä mitataan epäsuorasti, kysymällä vastaajan syntymävuotta. Syntymävuoden avulla on aineiston esikäsittelyvaiheessa helppo laskea vastaajan ikä vastausajankohtana. Tarkemmin ikä saadaan selville, jos voidaan käyttää tukena rekisteritietoja. Niihin vertaamalla voidaan myös tutkia vastaajan antamien tietojen luotettavuutta. Myös kysymysten sijoittelu lomakkeessa vaikuttaa vastausten luotettavuuteen. Ikää, kuten muitakin tärkeitä taustatekijöitä, on usein parempi kysyä vasta lomakkeen lopuksi, sillä niistä aloittaminen voi tuntua vastaajasta tungettelevalta.

Kyselytutkimuksessa käytetään enimmäkseen suljettuja osioita, mutta avoimiakin tarvitaan. Molemmilla on hyvät ja huonot puolensa. Valmiit vastausvaihtoehdot selkeyttävät mittausta sekä helpottavat tietojen käsittelyä olennaisesti. Tilastollisten analyysien kannalta keskeisiä ovat eri tavoin valmiiksi koodatut numeeriset vastaukset.

Sanalliset vastaukset ovat työläämpiä käsitellä, mutta joissain tilanteissa avoimet osiot toimivat suljettuja valintavaihtoehtoja paremmin. Avovastauksista saatetaan saada tutkimuksen kannalta tärkeää tietoa, joka voisi jäädä muuten kokonaan havaitsematta. Välttämättömiä avoimet osiot ovat tilanteissa, joissa vaihtoehtoja ei haluta tai ei voida luetella. Vaihtoehtoja voi olla liikaa, tai niitä ei vain ole mahdollista etukäteen rajata riittävästi.

Osioista jatkuviin ja diskreetteihin muuttujiin

Esimerkissä 2.1 käsitelty ikä on ominaisuutena tai käsitteenä *jatkuva*, koska se liittyy ajan jatkuvaan kulumiseen. Jatkuvien ilmiöiden mittaus ja esittäminen tietokoneella on kuitenkin mahdollista vain rajallisella tarkkuudella. Ikää voidaan mitata paljon tarkemmin kuin vuosina, mutta kyselytutkimuksessa se on harvoin tarpeellista.

Luvussa 3, jossa tutustutaan kyselyaineistoon, tullaan osioiden ohella puhumaan *muuttujista*. Muuttujalla tarkoitetaan aineistoon talletettua, usein numeeriseen muotoon koodattua tietoa mitatun osion sisällöstä. Mittaustarkkuuden rajallisuudesta huolimatta muuttujaa voidaan sanoa jatkuvaksi, jos se saa monia eri arvoja. Vastaavasti muuttujaa, joka saa vain harvoja eri arvoja, sanotaan *diskreetiksi*.

Esimerkki 2.2. Kolme tapaa kysyä vastaajan vointia.

```
Kuinka voitte?  _____________________________________________

Kuinka voitte?  1) hyvin  2) kohtalaisesti  3) huonosti

Kuinka voitte?  hyvin  10  9  8  7  6  5  4  3  2  1  0  huonosti
```

Keksityssä esimerkissä 2.2 esiintyy kolme vaihtoehtoista tapaa kysyä vastaajan vointia: yksi avoin ja kaksi suljettua kysymystä, joista toinen on tyypiltään karkea luokitus ja toinen numeerinen arviointi yksinkertaisella asteikolla. Avoimeen kysymykseen saadaan luultavasti paljon ”mielenkiintoisia” vastauksia, joiden koodaaminen voi olla melko työlästä. Suljetuista kysymyksistä ensimmäinen johtaa aineistossa diskreettiin muuttujaan, kun taas jälkimmäisen voi joissain tapauksissa tulkita jatkuvaksi. Kaikki kysymykset mittaavat kuitenkin vastaajan vointia. Selvimmin ne erottaa toisistaan *mittauksen taso*, johon seuraavaksi perehdytään.

2.3 Mittauksen taso

Osiot ovat siis mittausväline. Seuraavassa rajoitutaan suljettuihin osioihin, joiden mittaus voi edelleen olla monentasoista. Mittauksen taso vaikuttaa siihen, miten osiota voidaan jatkossa käyttää ja millaisissa tilastollisissa analyyseissa sitä voidaan hyödyntää. Mittaustaso vaikuttaa myös mittauksen laatuun, jota käsitellään kohdassa 2.4.

Mitä korkeampi mittaustaso on, sitä enemmän vaihtoehtoisia analysointitapoja on tarjolla. Jälkikäteen mittaustasoa ei voi nostaa, joten on syytä pyrkiä mittaamaan mahdollisimman korkeatasoisesti. Jos on mahdollista mitata määrää, ei kannata tyytyä luokittelemaan, toisin sanoen mittaamaan vain laadullista eroa. Toisaalta kaikkia asioita ei voi mitata määrällisesti. Silloin on vastaavasti luokiteltava mahdollisimman hyvin.

Mittaustason määrittelee lopulta se, miten osioon voidaan vastata. Mahdolliset vastaustavat ja siten mittaustasot voi tiivistää kolmeen päätyyppiin: 1) luokittelu, 2) järjestäminen ja 3) mittaaminen. Seuraavassa tarkastellaan näitä kolmea mittaustasoa lähinnä ulkonäkötutkimuksesta poimittujen esimerkkien avulla.

2.3.1 Luokittelu

Luokittelu edustaa puhtaasti laadullista mittaustasoa. Siinä ei määrällisillä asioilla ole sijaa, vaikka eri vaihtoehdot usein koodataankin numeroilla. Numerot ovat tässä yhteydessä enemmänkin koodeja kuin lukuja. Urheilijoiden peliasuissakin on yleensä numerot, jotta pelaajat voidaan tunnistaa, mutta pelinumeroilla ei silti ole mielekästä tehdä laskutoimituksia, ei edes laittaa pelaajia mihinkään varsinaiseen järjestykseen.

Analyysivaiheessa luokittelutason muuttujista voidaan laskea lukumääriä, niitä voidaan ristiintaulukoida (ks. luku 3) ja tehdä taulukoille jatkotarkasteluja muun muassa korrespondenssianalyysilla (ks. luku 7). Luokittelutason mittaukset ovat kyselytutkimuksessa tärkeitä, mutta jos vain on mahdollista mitata tarkemmin, ei pidä tyytyä pelkkään luokitteluun.

Esimerkki 2.3 on selvä luokittelutilanne. Vaikka vaihtoehdot onkin numeroitu, ei niillä ole sisällöllisesti mitään järjestystä. Ensimmäiseksi on sijoitettu odotettavasti yleisin ja viimeiseksi avoin kohta niitä tilanteita varten, joissa annetut vaihtoehdot eivät riitä. Vaihtoehtojen

numeroista on hyötyä lähinnä lomakkeen tallentajalle. Nettilomakkeessa ei vaihtoehtoja tarvitse edes numeroida; riittää kun annetaan lista, josta voi valita vain yhden vaihtoehdon. Koodauksesta huolehtii tällöin tiedonkeruuohjelma. Vaihtoehtoon 10 sisältyy myös tyypillinen avovastausmahdollisuus, sillä kaikkia mahdollisia vaihtoehtoja voi olla vaikea luetella. Tilanteesta riippuu, käytetäänkö näitä avovastauksia tarkemman luokittelun muodostamiseen vai tyydytäänkö raportoimaan ”muu vaihtoehto” sellaisenaan.

Esimerkki 2.3. Työllisyystilanne, kymmenen vaihtoehtoa.

```
Mikä on nykyinen työllisyystilanteenne? Ympyröikää yksi vaihtoehto.

  1. Kokopäiväinen palkansaaja
  2. Osapäivätoiminen palkansaaja
  3. Maatalousyrittäjä tai työssä perheen maatilalla
  4. Muu yrittäjä
  5. Työtön tai lomautettu
  6. Eläkeläinen
  7. Opiskelija
  8. Kotia hoitamassa/kotiäiti
  9. Äitiysloma
 10. Muu vaihtoehto, mikä? _____________________________________________
```

Työllisyystilanteet, joita esimerkki 2.3 kartoittaa, voivat nykyään olla kovin kirjavia, ja niinpä joillain vastaajilla voi olla vaikeuksia rajoittua vain yhteen vaihtoehtoon. Usein on yksinkertaistamisenkin uhalla syytä vaatia, että luokittelut ovat yksikäsitteisiä eli että eri vaihtoehdot ovat toisensa poissulkevia. Toisinaan vastaajalle saatetaan antaa lupa valita useampia annetuista vaihtoehdoista.

Useita valintoja samalla kertaa

Esimerkissä 2.4 on tyypillinen valintatilanne, jossa vastaaja saa valita niin monta kohtaa kuin haluaa. Valittavien vaihtoehtojen määrää voidaan myös rajata. Tällaiset valintatehtävät saattavat tuntua houkuttelevilta mittaustavoilta, mutta niiden huono puoli on karkea mittaustaso. Jokaisesta luonnehdinnasta välittyy vain tieto, onko vastaaja valinnut sen vai ei. Valitut voidaan koodata esimerkiksi ykkösellä ja loput nollilla. Näin tullaan käyttäneeksi dikotomista eli kaksiarvoista

asteikkoa, tässä esimerkissä yhteensä 74 kertaa. Aineistossa se tulee tarkoittamaan 74:ää muuttujaa, joten tiiviiksi ajatellusta mittarista tulee helposti laaja joukko tasoltaan karkeita mittauksia.

Esimerkki 2.4. Ulkonäön kuvailu valmiilla vaihtoehdoilla.

```
Ympyröikää tai alleviivatkaa seuraavista ilmaisuista ne, jotka mieles-
tänne kuvaavat ulkonäköänne. Voitte valita niin monta kuin haluatte.


alipainoinen / edustava / epämiellyttävä / epäsiisti / epäsuhtainen /
erikoinen / erittäin kaunis / erittäin ruma / hauskannäköinen /
hento / herttainen / hienostunut / hoikka / hyvin säilynyt /
hyvävartaloinen / iso / kaunis / kivannäköinen / klassinen / komea /
kookas / kurvikas / kurviton / leidimäinen / laiha / lattarintainen /
lauta, luuviulu / leveä / lihava / liian lyhyt / luonnollinen /
lyhyt / läski / miellyttävä / muodikas / muodoton / naisellinen /
nuhruinen / nuorekas / näyttävä / paksu / pehmeä / persoonallinen /
pieni / pitkä / poikamainen / poikkeava / pulska, pullea / pyöreä /
rehevä / ruma / rupsahtanut / ryhdikäs / ryhditön / seksikäs /
sensuelli / sievä / siisti / siro / sopusuhtainen / suuri / söpö /
tanttamainen / tasapaksu / tavallisennäköinen / tukeva /
tyttömäinen / tyylikäs / upea / urheilullinen / vastenmielinen /
viehättävä / voimakasrakenteinen / ylipainoinen
```

Esimerkin 2.4 asioita voisi mitata tarkemminkin laittamalla jokaisen luonnehdinnan kohdalle oman, esimerkiksi viisiportaisen asteikon, mutta tällöin lomake pidentyisi huomattavasti ja kävisi liian raskaaksi täyttää.

Jos asioita kysytään pelkästään esimerkin 2.4 tyyppisillä valintatehtävillä, mittaus jää kovin ohueksi. Osana laajempaa lomaketta, jossa suurin osa mittareista on tarkempia, valintatehtävä puolustaa kuitenkin paikkaansa. Tällä tavoin monesta vaihtoehdosta saadaan kokonaisuuden kannalta kiintoisia tietoja: voidaan katsoa, mitkä ovat yleisimpiä ilmaisuja, mitkä harvinaisimpia, miten ne ylipäätään jakautuvat tai miten monia tai harvoja ilmaisuja kukin vastaaja on valinnut. Luvussa 6 ilmaisuja tarkastellaan tiivistetysti kuvallisessa muodossa.

Luokitteluun viitataan usein sanalla *”luokitteluasteikko”*, mutta on selvempää puhua pelkästään luokittelusta, sillä kyseessä ei ole varsinainen asteikko, johon vaihtoehdot voitaisiin sijoittaa. Sana asteikko sisältää ajatuksen asioiden välisestä järjestyksestä.

2.3.2 Järjestäminen

Mahdollisuus asettaa luokat johonkin sisällön kannalta mielekkääseen järjestykseen nostaa mittauksen tasoa jonkin verran. Yleensä luokat järjestelee kysymyksen suunnittelija, mutta joskus järjestäminen annetaan vastaajan tehtäväksi. Järjestämiseen perustuvasta mittaustasosta käytetään myös nimeä *järjestysasteikko*.

Järjestämisen pitää siis perustua vaihtoehtojen sisältöön. Esimerkin 2.4 vaihtoehtojen aakkosjärjestys on vain ulkoinen järjestys. Se ei muuta luokittelua järjestysasteikoksi, kuten ei myöskään se, että sattumalta ensimmäisenä esiintyy luokka ”alipainoinen” ja viimeisenä ”ylipainoinen”. Koska välissä olevat luokat eivät muodosta mitään jatkumoa näiden välille, kyseessä on pelkkä luokittelu, jossa vaihtoehdot eivät ole edes toisensa poissulkevia. Sinänsä jokainen vaihtoehdoista muodostaa yksinkertaisimman mahdollisen järjestysasteikon sen perusteella, onko sana valittu vai ei, esimerkiksi ”iso” tai ”ei iso”.

Toisinaan vastaajaa pyydetään asettamaan joitakin annetuista vaihtoehdoista esimerkiksi paremmuusjärjestykseen. Mittaustavan huono puoli on se, että mittaus on epätarkkaa ja vastausten analysointi pinnallista. Asteikko, jolla vertailu tapahtuu, jää ikään kuin vastaajan määriteltäväksi, eivätkä vastaukset ole välttämättä vertailukelpoisia. On parempi, että tutkija määrittelee käytettävän asteikon, jolloin mittaustaso ja vertailumahdollisuudet paranevat selvästi.

Taustatiedot

Monet tyypilliset kyselytutkimuksen taustatiedot, kuten koulutustaso, edustavat järjestystasoista mittausta. Esimerkissä 2.5 koulutustasoa kuvaavilla luokilla 1–4 on ulkoisten numeroiden lisäksi selvä sisäinen järjestys alemmasta peruskoulutuksesta korkeampaan. Periaatteessa vaihtoehdot voitaisiin koodata vaikka kirjaimin A–D, mutta ohjelmistot käsittelevät yleensä sujuvammin lukuja ja numeroita kuin sanoja ja kirjaimia.

Numeroista huolimatta laskeminen ei tässäkään ole kovin mielekästä. Ei siis kannata raportoida, että ”peruskoulutus oli keskimäärin 3.6”. Mittaustaso vaikuttaa myös numeeristen tietojen esittämiseen, jota sivutaan useasti myöhemmissä luvuissa.

Esimerkki 2.5. Peruskoulutus, neljä vaihtoehtoa.

```
Mikä on peruskoulutuksenne?
1. Osa kansa- tai peruskoulua tai vähemmän 3. Keskikoulu tai peruskoulu
2. Kansakoulu tai kansalaiskoulu           4. Ylioppilastutkinto
```

Aktiivisuuden aste

Esimerkissä 2.6 mitataan liikuntaharrastamisen aktiivisuutta. Vastausvaihtoehdot muodostavat jatkumon, jonka ääripäät ovat ”joka päivä” ja ”en koskaan”. Muuttamalla sanalliset vaihtoehdot liikunnan harrastamiskertojen määriksi viikossa nähtäisiin, etteivät ne asetu asteikolle tasaisin välein.

Järjestysasteikossa vaihtoehtojen ei tarvitse olla tasavälisiä, kunhan järjestys pätee. Lisäksi välit näyttävät asteikon alkupäässä selvemmiltä kuin loppupäässä; vaihtoehdot 4 ja 5 ovat sanamuodoltaankin epämääräisempiä. Ne menevät myös osittain päällekkäin, vaikka järjestys säilyykin. Analyysivaiheessa luokkia saattaisi olla hyvä yhdistää, mutta mittausvaiheessa on parempi tarjota enemmän vaihtoehtoja.

Esimerkki 2.6. Liikunnan harrastaminen, aktiivisuusasteikko.

```
Kuinka usein harrastatte liikuntaa?
1. Joka päivä             4. Kerran tai kaksi kertaa viikossa
2. Viisi kertaa viikossa  5. Muutaman kerran kuukaudessa tai harvemmin
3. Kolme kertaa viikossa  6. En koskaan
```

Esimerkin 2.6 viimeinen vaihtoehto on selvä: ”en koskaan” tarkoittaa, ettei vastaaja harrasta lainkaan liikuntaa. Samalla saadaan siis vastaus kysymykseen ”harrastatteko liikuntaa?”. Jos sitä kysyttäisiin erikseen, vaihtoehtoina ”kyllä” ja ”en”, saataisiin vain karkea tieto. Nyt liikunnan harrastajilta saadaan monipuolisempi arvio harrastamisen aktiivisuudesta. Mittaustaso on tarkempi huolimatta edellä mainituista asteikon epätarkkuuksista.

Laatua ja määrää samassa

Liikunnan harrastamiskysymys (esimerkki 2.6) edustaa tyypillistä tilannetta, jossa asteikko sisältää sekä laadullisen että määrällisen osan. Laadullista on tässä tapauksessa se, harrastaako lainkaan liikuntaa, ja määrällistä se, kuinka usein harrastaa. Olennainen ero on, harrastaako lainkaan liikuntaa. Analyyseissa tällaisten asteikkojen laadullinen vaihtoehto on usein syytä ottaa erikseen huomioon eikä ajatella sitä vain osana määrällistä jatkumoa.

Kaikkiaan esimerkin 2.6 kysymys on varsin yleisluonteinen, koska siinä ei mitenkään täsmennetä liikunnan lajia. Tutkimuksen tavoitteista riippuen tästä voisi jatkaa ja kysyä tarkemmin liikunnan harrastamisesta niiltä, jotka vastasivat 1–5. Vastaavasti muilta voisi tiedustella syitä liikuntaharrastuksen puutteeseen.

Käsityksiä, luonnehdintoja ja huolia

Esimerkissä 2.7 vastaajaa pyydetään luonnehtimaan omaa painoaan. Nyt aiemmin mainitut luonnehdinnat ”alipainoinen” ja ”ylipainoinen” muodostavat asteikon ääripäät, vieläpä lisättynä täsmennyksellä ”reilusti”. Tässä aletaan lähestyä numeerista mittausta, sillä vaihtoehdot muodostavat selvän jatkumon, jonka keskellä on neutraaliksi vaihtoehdoksi tarkoitettu ”normaalipainoinen”.

Esimerkki 2.7. Painon luonnehdinta viisiportaisena.

```
Oletteko omasta mielestänne?     3. Normaalipainoinen
1. Reilusti alipainoinen         4. Vähän ylipainoinen
2. Vähän alipainoinen            5. Reilusti ylipainoinen
```

On muistettava, että mittauksen kohteena tässä on käsitys tai luonnehdinta omasta painosta, ei varsinaisesti paino, jota voisi mitata tarkemminkin. Vastauksia esimerkin 2.7 kysymykseen kannattaisi verrata esimerkissä 2.4 valittuihin luonnehdintoihin. Koska painoa pidetään yleensä ulkonäön kannalta keskeisenä, sitä kannattaa mitata useammasta näkökulmasta.

Esimerkki 2.8 puolestaan luotaa painosta huolissaan olon astetta jatkumolla, jonka ääripäinä ovat ”aina” ja ”en koskaan”. Viimeinen vaihtoehto on sama kuin liikuntaharrastusesimerkissä 2.6, mutta muut vaihtoehdot ovat epämääräisempiä. Mitattava ilmiö määrää vaihtoehtojen sanavalinnat. Liikunnan harrastaminen on fyysistä toimintaa, kun taas huolissaan olo on mielentila.

Esimerkki 2.8. Huolissaan olo painosta, ääripäät ja välit.

```
Kuinka usein olette huolissanne painostanne?
1. Aina          3. Usein      5. Harvoin
2. Lähes aina    4. Joskus     6. En koskaan
```

Liikunnan harrastamisen osalta mittausta voitaisiin haluttaessa tarkentaa lisäämällä vaihtoehtoja. Huolissaan olon kohdalla se tuskin olisi mahdollista, ainakaan jos haluttaisiin ilmaista vaihtoehdot sanallisesti. Esimerkiksi sanavalinnoilla ”miltei aina”, ”lähes aina” tai ”erittäin usein”, ”todella usein”, ”hyvin usein”, ”kovin usein” tulisi kyseenalaiseksi, olisivatko vaihtoehdot enää edes yksikäsitteisessä järjestyksessä. Mittaus ei tarkentuisi, vaan se sumentuisi.

On siis pyrittävä mahdollisimman tarkkaan mittaukseen, muttei toisaalta pidä yrittää liikaa. Ehdottomasti esimerkissä 2.8 on järkevämpää käyttää mainittua kuutta vaihtoehtoa kuin tyytyä kysymään vain ”Oletteko huolissanne painostanne?”. Tällaisesta kysymyksestä seuraisi vain epäselvyyksiä: tarkoitetaanko juuri nyt vai yleensä? Kysymättäkin on selvää, että monet ovat ainakin joskus huolissaan painostaan. Kysymys pitää asettaa niin, että vastaajalle on selvää, mitä kysytään. Tällöin kysymyksiin saadaan myös tarkempia vastauksia.

Sanallisiin vaihtoehtoihin palataan vielä seuraavassa, sillä etenkin kyselytutkimuksessa järjestämisellä ja mittaamisella on pieni mutta merkittävä ero.

2.3.3 Mittaaminen

Mittaamiseksi tässä kirjassa kutsutaan varsinaista numeerista mittausta, joka kattaa sen, mihin luokittelu ja järjestäminen eivät yllä. Kirjallisuudessa puhutaan myös *väliasteikosta* ja *suhdeasteikosta*. Aineiston analyysin kannalta niillä ei ole suurta eroa.

Väliasteikossa asteikon pykälien välit ovat yhtäsuuria, mutta nollakohta ei ole yksikäsitteisesti määritelty ja näin ollen suhteelliset tarkastelut eivät ole mahdollisia. Yleisin esimerkki on lämpötila, vaikkakaan sitä ei ole tapana mitata kyselytutkimuksilla. Toisinaan näkee kyllä kiinnostusta vaaleihin mitattavan ”vaalilämpömittareilla”.

Suhdeasteikolle on ominaista hyvin määritelty mittayksikkö ja määrän mittaaminen, jolloin asteikossa on yksikäsitteinen nollakohta. Suhdeasteikolla mitatut arvot eivät siis voi olla negatiivisia. Esimerkissä 2.9 on näytteitä suhdeasteikollisesta mittaamisesta senttimetreinä, kilogrammoina ja euroina.

Esimerkki 2.9. Pituus, paino ja rahan käyttö vaatteisiin.

```
Kuinka pitkä olette ja paljonko painatte?

Pituus _____ cm  Paino _____ kg


Paljonko suurin piirtein käytätte rahaa vuosittain vaatetukseenne?

Noin ___________ euroa vuodessa
```

Jos suhdeasteikosta poistetaan mittayksiköt ja desimaalit, jäljelle jäävät lukumääriä eli frekvenssejä mittaavat kokonaisluvut. Lukumääriin päädytään myös minkä tasoisista mittauksista tahansa tarkastelemalla niiden frekvenssijakaumia (ks. luku 3).

Onko asteikolla väliä?

Käytännön kannalta olennaisin ero koskee järjestämistä ja mittaamista, tarkemmin sanottuna järjestysasteikkoa ja väliasteikkoa. Vaikka joissain tilanteissa ero onkin selvä, on myös paljon tilanteita, joissa ero voi vaikuttaa hämärämmältä.

Monet kyselytutkimuksen keskeiset mittaustavat, kuten asennemittaukset, käsitetään kirjallisuudessa järjestysasteikoiksi. Yleisin näistä tunnetaan *Likertin asteikkona*. Esimerkissä 2.10 on tyypillinen näyte tästä asteikosta, jota useimmiten sovelletaan viisiportaisena.

Esimerkki 2.10. Käsityksiä ulkonäön merkityksestä.

```
Ympyröikää omaa käsitystänne parhaiten vastaava vaihtoehto.

Vaihtoehdot ovat:   1: Täysin samaa mieltä, 2: Osin samaa mieltä,
  3: Ei samaa eikä eri, 4: Osin eri mieltä, 5: Täysin eri mieltä.

Ulkonäkö on liian arvostetussa asemassa.   1    2    3    4    5
Miellyttävästä ulkonäöstä on hyötyä.       1    2    3    4    5
Hyvännäköiset ihmiset pärjäävät paremmin.  1    2    3    4    5
```

Likertin asteikko täyttää hyvin järjestysasteikon tunnusmerkit, mutta jos siihen tyydytään, ei päästä pitkälle, sillä järjestysasteikolle soveltuvia tilastollisia menetelmiä on vähän. Valtaosa tässäkin kirjassa esitetyistä menetelmistä nojaa keskiarvoihin, hajontoihin ja korrelaatioihin, joiden laskeminen edellyttää väliasteikollista mittausta.

Käytännössä Likertin asteikoilla tehdään tilastollista analyysia ikään kuin kyseessä olisi väliasteikko. Ilman mitään perusteluja toiminta voi tuntua aika ristiriitaiselta, joten on syytä pohtia tarkemmin sen edellytyksiä ja epävarmuuksia.

Ensinnäkin on tärkeää, että käytettävä asteikko muodostaa selvän, yksiulotteisen jatkumon jostain ääripäästä toiseen. Tyypilliset ääripäät ”täysin samaa mieltä” ja ”täysin eri mieltä” eivät tässä suhteessa ole ongelma. Ongelma ilmeneekin yleensä asteikon keskellä, johon saatetaan sijoittaa kaikenlaisia vaihtoehtoja. Likertin asteikon rakenteeseen kuuluu, että keskimmäinen vaihtoehto on *neutraali*, esimerkiksi ”ei samaa eikä eri mieltä”.

Keskimmäinen vaihtoehto

Voidaan kysyä, tarvitaanko keskimmäistä vaihtoehtoa lainkaan. Joissain tapauksissa, kuten yhdistyksen jäsenkyselyissä, voi olettaa vastaajien ottavan kantaa, mutta yleensä neutraali vaihtoehto on syytä olla. Jos kynä on paperilomakkeessa alkanut piirtää pystyviivaa kolmosen kohdalle, on syynä liian pitkä lomake tai liian vaikeat kysymykset. Mikäli neutraalia vaihtoehtoa ei ole, vastaaja voi helposti jättää kokonaan vastaamatta. Neutraali vastaus on parempi kuin puuttuva tieto.

Usein käytetty vaihtoehto ”en osaa sanoa” (eos) voi sen sijaan olla kaukana neutraalista, useastakin eri syystä. Vastaaja ei ehkä ole ymmärtänyt kysymyksen sisältöä riittävästi ottaakseen siihen kantaa, tai on ymmärtänyt, muttei ole halunnut ilmaista kantaansa. Syitä eos-vastauksiin on mahdoton erottaa toisistaan lomakkeiden perusteella. Eos-vaihtoehdon sijoittaminen keskimmäiseksi onkin huono ajatus, koska se mittaa eri asiaa kuin kysymys muuten. Koska eos-vaihtoehto rikkoo sekä mittauksen jatkumon että yksiulotteisuuden, vastauksista tehdyt analyysit ja johtopäätökset jäävät epämääräisiksi.

Kysymyksestä riippuen eos-vaihtoehdon voi mieluummin tarjota omana kohtanaan asteikon ulkopuolella, sillä sekin on parempi kuin puuttuva tieto. Analyysivaiheessa on vain muistettava, että eos-vaihtoehto ei kuulu varsinaiseen asteikkoon. Esimerkiksi numeroksi 9 koodatut eos-vastaukset nostaisivat laskelmiin päätyessään ”yllättävästi” asteikon keskiarvoa. Aineiston muokkausten yhteydessä (ks. kohta 3.5) on hyvä tutkia eos-vastauksia ja miettiä, mitä niille voi tehdä ennen analyyseja.

Vaihtoehtojen välit

Väliasteikossa vaaditaan myös, että vaihtoehtojen välit ovat yhtä suuria. Numeroina ajateltuna on itsestään selvää, että asteikossa 1–5 vaihtoehtojen 1 ja 2 väli on ykkösen mittainen, samoin vaihtoehtojen 3 ja 4. Peräkkäiset vaihtoehdot ovat siis yhtä kaukana toisistaan.

Kun numerot korvataan sanallisilla ilmaisuilla, ei olekaan enää selvää, minkä mittaisia eri vaihtoehtojen välit ovat. Millä perusteella esimerkiksi ”täysin samaa mieltä” ja ”osin samaa mieltä” olisivat yhtä kaukana toisistaan kuin vaikkapa ”ei samaa eikä eri mieltä” ja ”osin eri mieltä”?

Usein kuultu vastaus tähän kysymykseen on, ettei mitään perustetta ole. Vastaus tarkoittaa, että Likertin asteikkoa on syytä pitää vain järjestysasteikkona. Silloin välien suuruudet voivat vaihdella, kunhan järjestys säilyy. Tästä kuitenkin seuraa jo edellä mainittu ongelma: järjestysasteikolle soveltuvia tilastollisia menetelmiä on vähän. Analyyseissa ei päästä eteenpäin, jos ei voida laskea tarvittavia tunnuslukuja kuten korrelaatioita.

Tyypillinen Likertin asteikkojen käytännön soveltaminen on perusteltavissa yksinkertaisesti mittauksen käsittein. Likertin asteikko voidaan mieltää väliasteikoksi, jossa poikkeamat yhtä suurista väleistä johtuvat kohdassa 2.2.1 mainituista mittauksen häiriötekijöistä, mittausvirheistä. Jos niitä ei olisi, Likertin asteikko olisi puhdas väliasteikko. Se on kuitenkin epärealistista, sillä mittausvirhettä sisältyy kaikkiin mittauksiin.

Likertin asteikoista voi siis vallan hyvin laskea keskiarvoja, hajontoja ja korrelaatioita, kunhan ei tyydy pelkästään niihin vaan soveltaa menetelmiä, joilla mittausvirheiden vaikutuksia saadaan hälvennettyä. Näihin kysymyksiin perehdytään luvussa 4.

Vaihtoehtojen lukumäärät

Tyypillisesti tällaisissa asteikoissa on viisi vaihtoehtoa, jota pidetään yleisesti sopivana määränä käsitettäväksi yhtaikaa. Joissain tilanteissa voi olla hyödyllistä laajentaa valikoimaa seitsemään sisällyttämällä mukaan vielä painavammat ääripäät, esimerkiksi ”ehdottomasti samaa mieltä” ja ”ehdottomasti eri mieltä”. Vastaavasti voidaan tietysti vaihtoehtojen lukumäärää supistaakin, mutta jos vain mahdollista ja luontevaa, ei kannata suotta karkeistaa mittausta.

Seitsenportaisen asteikon numeroarvot voivat olla 1, 2, . . . , 6, 7, mutta aivan yhtä hyvin 7, 6, . . . , 2, 1 tai vaikka –3, –2, . . . , 2, 3. Tilastollisen analyysin kannalta näillä ei ole eroa. Mahdolliset erot tulevat muista seikoista, esimerkiksi siitä, missä järjestyksessä vaihtoehdot esiintyvät lomakkeella. Periaatteessa luontevinta on koodata asteikot siten, että suurin numero vastaa eniten samanmielistä vaihtoehtoa, jolloin asteikkojen suunnat ovat loogisia. Suunnat voi kuitenkin kääntää myös aineiston muokkausvaiheessa.

Vaihtoehtojen esitysjärjestys riippuu myös tutkittavasta asiasta ja voi ilmentää tutkijan tekemiä sisällöllisiä olettamuksia. Selkeyden vuoksi on hyvä pitää kiinni valitsemastaan järjestyksestä eikä vaihdella sitä lomakkeen mittarista toiseen. Vastaaja turhautuu, jos joutuu joka mittarin kohdalla selvittämään erikseen, miten sen väitteisiin vastataan. Paperilomakkeissa numerot on syytä esittää, jottei vastauksia tallennettaessa tulisi epäselvyyksiä. Verkkolomakkeessa, jonka tallennus tapahtuu automaattisesti, voidaan numerot häivyttää kokonaan vastaajan näkyvistä.

Kouluarvosana-asteikko

Varsinkin palautelomakkeissa näkee käytettävän niin sanottua kouluarvosana-asteikkoa. Se tarkoittaa seitsemänportaista asteikkoa, jonka numeroarvot ovat 4, 5, . . . , 9, 10. Tilastollisesti on samantekevää, mitä kohtaa lukusuorasta käytetään, mutta sisällöllisesti kouluarvosana-asteikko on sen verran ongelmallinen, ettei sitä kannata käyttää.

Kaikille ei ole edes selvää, mitä kouluarvosana-asteikolla tarkoitetaan, koska se on saatettu useissa kouluissa korvata toisenlaisilla asteikoilla. Vaikka asteikko olisi omilta kouluajoilta tuttu, ei ole sama asia arvioida jotain asteikolla, jolla on tullut itse aikanaan arvioiduksi.

Käsityksiin kouluarvosana-asteikosta heijastuu myös vastaajan oma koulumenestys. Harva osaa käyttää tätä asteikkoa koko laajuudelta, esimerkiksi arvosanoihin 9 ja 10 tottuneelle asteikon alkupää on tuntematon ja toisinpäin. Tämänkaltaiset arvolataukset eivät kuulu hyvään mittaukseen.

Kouluarvosana-asteikkoa ei siis kannata käyttää, koska se tuo mukanaan liikaa turhia epävarmuuksia. Asioita voi mitata luotettavammin muuntyyppisillä asteikoilla.

Laatusanat ja mittaus

Eräs vertailuun perustuva mittaustapa tunnetaan nimellä *semanttinen differentiaali* tai *Osgoodin asteikko*. Esimerkissä 2.4 lueteltuja asioita voisi mitata tällä tekniikalla hakemalla sopivia laatusanapareja, kuten ulkonäkötutkimuksen pohjalta keksityssä esimerkissä 2.11. Yleensä on tapana käyttää seitsemänportaista asteikkoa, joka voi olla näkyvissä numeroina, tyhjinä kohtina, joihin laitetaan rasti tai verkkolomakkeessa liukurina, joka asetetaan haluttuun kohtaan.

Esimerkki 2.11. Ulkonäön luonnehdintoja laatusanapareilla.

```
Arvioikaa ulkonäköänne seuraavien sanaparien avulla:

    alipainoinen   1  2  3  4  5  6  7   ylipainoinen
     miellyttävä   1  2  3  4  5  6  7   epämiellyttävä
       epäsiisti   1  2  3  4  5  6  7   siisti
            ruma   1  2  3  4  5  6  7   kaunis
           laiha   1  2  3  4  5  6  7   lihava
     tyttömäinen   1  2  3  4  5  6  7   poikamainen
```

Sanaparit kannattaa laittaa vaihtelevaan järjestykseen eikä kasata kaikkia ”hyviä” tai ”huonoja” ominaisuuksia samalle puolelle. Osgoodin asteikko on kätevä mittaustapa tilanteisiin, joissa sanoille löytyy selviä vastinpareja.

Myös Likertin asteikkoa käytetään paljon siten, että vain ääripäät kuvataan sanallisesti, siis ikään kuin Osgoodin asteikon tapaan. Selvempää on kuitenkin kuvata kaikki vaihtoehdot sanallisesti, jotta vastaajat ymmärtävät asteikon mahdollisimman yhtenäisellä tavalla. Tällöin tulee myös selvemmin esiin, mitä keskimmäisellä vaihtoehdolla tarkoitetaan.

Lopuksi vain kaksi vaihtoehtoa

Edellä on monessa kohtaa noussut esiin asteikko, jossa on vain kaksi vaihtoehtoa. Tällainen *dikotominen* asteikko on siitä erikoinen, että se voi edustaa mitä tahansa mittaustasoa. Kun vaihtoehtoja on kaksi, ne voidaan asettaa järjestykseen. ”Oikea” suunta riippuu tulkinnasta samaan tapaan kuin useampiportaisissakin järjestysasteikoissa. Vaihtoehtojen väleistäkään ei tule mitään ongelmaa, sillä niitä on vain yksi. Näin dikotominen asteikko on myös väliasteikko, joten sen arvoilla voidaan tehdä kaikkia tarvittavia laskelmia.

Dikotomisella asteikolla voi olla käyttöä aineiston analyysivaiheessa, kun halutaan yksinkertaistaa tarkasteluja ”joko – tai” -tyyppisiksi. Mittaukset on ehkä tehty tarkemmin, mutta lopulta kiinnostaa vain, ylittyykö jokin sisällöllisesti merkittävä raja vai ei. Usein luokituksia saatetaan tiivistää dikotomisiksi, jotta saadaan käsiteltyä mahdollisimman monia asioita yhtaikaa.

Mittausvaiheessa dikotomioita on parasta välttää, jos on mahdollista mitata tarkemmin. Ei siis pidä ainakaan pelkästään kysyä ”Kyllä vai ei?” tilanteissa, joissa voi kysyä ”Miten usein?” tai ”Miten paljon?”. Vastaavasti on myös kysymyksiä, joihin on luontevinta vastata vain joko ”kyllä” tai ”ei”, esimerkiksi ”Onko teillä vakituinen parisuhde?”.

Aikanaan dikotomiset kysymykset olivat yleisiä, koska tuloksia voitiin laskea jopa käsin suhteellisen helposti. Tällaisilla perusteluilla ei ole nykyään mitään virkaa, koska ohjelmistojen ansiosta laskemiseen ei tarvitse enää kiinnittää niin paljon huomiota. Kannattaa mitata niin tarkasti kuin mahdollista, jolloin analyysivaiheessa on enemmän mahdollisuuksia.

2.4 Mittauksen luotettavuus

Kuten edellä on havaittu, mittaus ei kyselytutkimuksessa ole niin suoraviivaista kuin ehkä voisi kuvitella. Mittauksen luotettavuuteen ja laatuun vaikuttavat sisällölliset, tilastolliset, kulttuuriset, kielelliset ja teknisetkin seikat, joten on selvää, että laadukas mittaus edellyttää usean asiantuntijan yhteistyötä.

Monesti tilastotieteilijältä kysytään neuvoa vasta, kun aineisto on kerätty ja mietitään, millä menetelmillä päästäisiin parhaiten kiinni tutkimuskysymyksiin. Valitettavasti siinä vaiheessa on monelta osin liian myöhäistä.

Mittaus on ainutkertaista, eikä huonosti mitattuja osioita voi jälkikäteen parantaa millään menetelmillä. Mittauksen laatuun voi kuitenkin vaikuttaa etukäteen. Neuvoja kannattaa kysyä asiantuntijoilta jo lomakkeen suunnitteluvaiheessa. Kohdassa 2.2.1 (s. 20) mainittua ”kotitehtävää” on myös hyvä pohtia; kokemusten mukaan sillä on pelkästään positiivisia vaikutuksia, ei ainoastaan mittauksen vaan koko tutkimuksen luotettavuuteen.

Mittauksen luotettavuudesta puhuttaessa erotetaan kaksi perustetta: *validiteetti* ja *reliabiliteetti*. Edellistä näkee toisinaan kutsutun pätevyydeksi ja jälkimmäistä joko luotettavuudeksi tai toistettavuudeksi. Luotettavuus on kuitenkin laajempi käsite kuin pelkkä reliabiliteetti. Toistettavuus on puolestaan liian suppea määritelmä reliabiliteetille.

2.4.1 Validiteetti

Tiiviisti ilmaistuna validiteetti kertoo, mitataanko sitä, mitä piti, ja reliabiliteetti kertoo, miten tarkasti mitataan. Toimivia suomennoksia voisivat olla pätevyys ja tarkkuus, mutta tässä kirjassa käytetään sanoja validiteetti ja reliabiliteetti, koska ne ovat jokseenkin vakiintuneita ilmaisuja.

Osio tai mittari saattaa todella mitata jotain muuta kuin sen luultiin mittaavan. Esimerkkinä voidaan ajatella tilannetta, jossa tutkimus toistetaan eri maassa kuin missä se on alun perin tehty. Tyypillisesti kyselylomake joudutaan tällöin kääntämään eri kielelle. Jos keskitytään vain osioiden huolelliseen kääntämiseen, voi seurata kohtalokkaita yllätyksiä, sillä ne saattavat mitata eri maissa ja kulttuureissa tyystin eri asioita.

Sama koskee yleensäkin tutkimuksen toistamista, sillä mikään ei takaa mittareiden tai osioiden ajallista pysyvyyttä. Kuten kohdassa 2.2.1 korostettiin, on ensisijaisesti ajateltava mitattavia ulottuvuuksia – siis ensin asiat, sitten osiot. Luonnollisesti ilmiön ulottuvuudetkin muuttuvat ajassa. Mittareihin pitää olla valmiina tekemään muutoksia, jotta ne toimisivat luotettavasti.

Validiteetti on mittauksen luotettavuuden kannalta ensisijainen peruste, sillä ellei mitata oikeaa asiaa, ei reliabiliteetilla ole mitään merkitystä. Validiteetti voidaan luokitella useaan eri tyyppiin, joista kertoo esimerkiksi Alkula ym. (1994, 88–93). Koska validiteetti on ennen kaikkea tutkittavan ilmiön sisällöllinen kysymys, sitä voi vain osittain lähestyä tilastollisesti. Luvuissa 4 ja 5 käsitellään kahta tilastollisesti arvioitavaa validiteetin muotoa, rakennevaliditeettia ja ennustevaliditeettia.

2.4.2 Reliabiliteetti

Validiteetin lisäksi on tavoittelemisen arvoista saada mittaus reliabiliteetiltaan mahdollisimman hyvälle tasolle. Mittauksen reliabiliteetti on sitä parempi, mitä vähemmän siihen sisältyy mittausvirhettä. Näihin asioihin syvennytään luvuissa 4 ja 5.

Kirjallisuudesta mainittakoon tässä yhteydessä *Margins of Error: A Study of Reliability in Survey Measurement* (Alwin, 2007), joka korostaa reliabiliteetin osuutta tutkimuksen luotettavuudessa etenkin yhteiskuntatieteissä. Kirjan esimerkit perustuvat yhdysvaltalaisiin kyselytutkimuksiin, kun taas teoksen *Design, Evaluation, and Analysis of Questionnaires for Survey Research* (Saris & Gallhofer, 2007) esimerkit pohjautuvat vahvasti *European Social Survey* -tutkimuksiin. Jälkimmäinen kirja keskittyy kyselylomakkeiden, kysymysten ja vastausvaihtoehtojen laatimisen ja kehittämisen haasteisiin.

Mittauksen ohella myös tiedonkeruu on merkittävä epävarmuuden aiheuttaja tilastollisessa tutkimuksessa. Tutkimuksen kokonaisluotettavuus edellyttää luotettavuutta sekä mittaukselta että tiedonkeruulta.

2.5 Tiedonkeruu

Tilastotiede on perinteisesti keskittynyt tiedonkeruun haasteisiin. Otantateorialla on näissä kysymyksissä keskeinen asema. Tiedonkeruusta johtuvia epävarmuuksia pystytään nykyisin hallitsemaan yhä paremmin muun muassa kehittyneiden otantamenetelmien avulla.

Tilastotieteen oppikirjoista voi välillä saada sellaisen vaikutelman, että aineistot vain yhtäkkiä ”ilmestyvät” jostain tutkijan työpöydälle. Eri alojen tutkijat tietävät hyvin, ettei se niin yksinkertaista ole. Tilastotieteilijä on ehkä vain tottunut tulemaan ”valmiiseen pöytään”. Aineiston kerääminen vaatii kovaa työtä aivan kuten mittareidenkin laatiminen. Tutkimuksen luotettavuutta ajatellen sekä mittauksesta että tiedonkeruusta on huolehdittava mahdollisimman hyvin, sillä molemmat ovat ainutkertaisia vaiheita.

Kyselytutkimuksen aineiston voi kerätä monella tavalla, esimerkiksi kirjekyselynä tai verkkolomakkeella. Keskeisimmät vastaajien valintatavat perustuvat *otantaan*, jonka perusmenetelmistä kertoo esimerkiksi Alkula ym. (1994, 106–123). Yksityiskohtaisemmin aihepiiriä käsittelee *Otanta-asetelmat ja tilastollinen analyysi* (Pahkinen & Lehtonen, 1989).

Pidemmälle meneviin otanta-asetelmiin ja otanta-aineistojen analysointiin perehdyttää samojen tekijöiden englanninkielinen teos *Practical Methods for Design and Analysis of Complex Surveys* (Lehtonen & Pahkinen, 2004). Laaja kyselytutkimuksen tiedonkeruun ja mit-

tauksen perusteos on *Survey Methodology* (Groves ym., 2004), joka lähestyy aihepiiriä suureksi osaksi ilman matemaattisia kaavoja.

Seuraavassa pohditaan tavallisimpien tiedonkeruutapojen vaikutuksia tutkimuksen luotettavuuteen ja tulosten yleistämiseen. Lisäksi kuvataan ulkonäkötutkimuksen otanta-asetelmaa ja palataan vielä kyselylomakkeeseen, nyt tiedonkeruun näkökulmasta.

2.5.1 Perusjoukko ja otos

Perusjoukko ja *otos* ovat otannan tärkeimmät käsitteet. Perusjoukon muodostavat ne, joista tutkimuksessa ollaan kiinnostuneita, esimerkiksi ”työikäiset suomalaiset”. Otoksen muodostavat tutkimukseen valituiksi tulleet vastaajat. Otannan idea on, että kooltaan perusjoukkoa huomattavasti pienemmän otoksen perusteella saadut tulokset voidaan yleistää koskemaan perusjoukkoa. Johtopäätösten tekemistä otoksen perusteella kutsutaan *tilastolliseksi päättelyksi*, johon perehdytään luvussa 4.

Otannan toimivuuden takaavat huolellisesti laadittu *otanta-asetelma*, tarkoituksenmukainen otantamenetelmä ja ennen kaikkea vastaajien valintaan sisältyvä *satunnaisuus*. Jokaisella perusjoukkoon kuuluvalla tulee olla sama todennäköisyys tulla valituksi otokseen. Muussa tapauksessa otos ei edusta perusjoukkoa.

Otoskoko on myös tärkeä, vaikkakin jossain määrin toissijainen kysymys, ja vain yksi tutkimuksen luotettavuuteen vaikuttavista tekijöistä. Otoskoko vaikuttaa siihen, miten tarkasti otos kuvaa perusjoukkoa. Tämä tarkkuus ei valitettavasti kasva suorassa suhteessa otoskokoon vaan ainoastaan otoskoon neliöjuureen. Esimerkiksi tarkkuuden kaksinkertaistaminen vaatisi nelinkertaista otoskokoa. Otoskoon kasvattaminen voi siis käydä kalliiksi. Hyvä otanta-asetelma mahdollistaa luotettavien johtopäätösten tekemisen pienemmälläkin otoskoolla. Tutkimuksen kokonaisluotettavuutta ajatellen on syytä panostaa myös mittaukseen (vrt. kohta 2.4).

Aineistoista käytetään usein sanaa ”otos”, vaikka tiedonkeruussa ei olisikaan kyse otannasta. Tiedonkeruuseen liittyvä epävarmuus on tällöin vaikeammin arvioitavissa. Mittausepävarmuudet ovat mukana riippumatta siitä, onko kyse otannasta vai ei. Tästä syystä mittaukseen on kiinnitettävä huomiota kaikessa tilastollisessa tutkimuksessa. On muistettava myös, että huonosti tehty mittaus ei tarkennu kasvattamalla otoskokoa.

Virhemarginaali

Tiedotusvälineissä viitataan yleensä tutkimusten ”virhemarginaaliin”. Tyypillinen virhemarginaali esimerkiksi puolueiden kannatusosuuksia mittaavissa tutkimuksissa on parin prosenttiyksikön luokkaa suuntaansa. Marginaali sisältää otannasta johtuvan epävarmuuden sekä korkeintaan karkean arvion muista epävarmuuksista. Näiden tutkimusten otoskoko on yleensä noin tuhat. Virhemarginaalin kaventaminen vaatisi helposti useita tuhansia vastaajia, mikä lisäisi tiedonkeruun kustannuksia, mutta vähentäisi silti vain otannasta johtuvaa epävarmuutta. Muut epävarmuudet johtuvat muun muassa mittauksesta, ja niiden vaikutusta virhemarginaaleihin on vaikeampi arvioida.

Vastausprosentti ja kato

Vastausprosentti on eräs tutkimuksen luotettavuuden ilmaisin. Se kertoo, kuinka moni otokseen valituista vastasi, siis täytti ja palautti kyselylomakkeen. Otoshan poimitaan niin, että se edustaa perusjoukkoaan, mutta mikäli vastausprosentti jää kovin alhaiseksi, edustavuus voi jäädä kyseenalaiseksi. Kato on puolestaan sitä suurempi, mitä useampi jättää vastaamatta, joko kokonaan tai osittain. Vaillinnaisia vastauksia voidaan tietyin ehdoin paikata, mutta osa vastauksista joudutaan yleensä hylkäämään.

Tyypilliset kyselytutkimuksen vastausprosentit lienevät nykyisin alle 50 %:n suuruisia. Mikäli kato käy vielä pahemmin, voi vastausprosentti pudota vaikkapa kymmeneen. Jos 90 % päätti jättää vastaamatta, voi kysyä, keitä ovat ne erikoiset henkilöt, jotka vastasivat kyselyyn! Tällaisilla prosenteilla otos muuttuisi varsin ”epäedustavaksi”. Tutkimuksen luotettavuuden arvioinnin kannalta on syytä raportoida vastausprosentti, jotta nähdään, kuinka moni ylipäätään vastasi kyselyyn. Jos kyseessä on otos, voidaan vastaamatta jättäneistä saada tarkempi käsitys niin sanotulla kadon analyysilla, jossa verrataan otoksen taustatietoja perusjoukon vastaaviin tietoihin. Lisäksi on hyvä tehdä selkoa saatujen vastausten laadusta, esimerkiksi siitä, miten paljon niissä esiintyy puutteellisia tietoja.

Ulkonäkötutkimuksen otanta-asetelma

Ulkonäkötutkimuksen aineisto on kerätty postikyselynä kahtena eri ajankohtana, vuosina 1997 ja 2005. Molemmilla kerroilla kyselylomake lähetettiin 500 suomalaiselle naiselle. Naisten osoitetiedot hankittiin Väestörekisterikeskuksesta, joka suoritti satunnaisotannan Suomessa asuvien Suomen kansalaisten, äidinkielenään suomea puhuvien 18–74-vuotiaiden naisten perusjoukosta. Ahvenanmaa rajattiin perusjoukon ulkopuolelle. Kyselyn suorittaminen kahtena eri ajankohtana mahdollistaa naisten ulkonäkösuhtautumisen ajallisen vertaamisen, vaikka eri vuosina kyselyyn ovatkin vastanneet eri henkilöt.

Otanta-asetelman puolesta tutkimuksen tulokset ovat yleistettävissä suomenkielisiin suomalaisiin aikuisiin naisiin, joiden kyselyiden aikainen kotipaikka ei ollut Ahvenanmaa. Tutkimukseen hyväksyttäviä vastauksia palautui 273 (55 %) vuonna 1997 ja 223 (45 %) vuonna 2005. Aineiston edustavuutta tarkasteltiin vertaamalla sitä perusjoukkoon iän, koulutuksen, siviilisäädyn ja sosioekonomisen aseman suhteen sekä vastaajan itse ilmoittamasta painosta ja pituudesta lasketulla painoindeksillä. Vuoden 1997 otoksesta tehty kadon analyysi osoitti otoksen edustavan perusjoukkoa kohtalaisen hyvin (Valtari, 2001).

2.5.2 Kokonaistutkimus ja rekisterit

Kokonaisvaltaisesti puolueiden kannatuksia mitataan vaaliuurnilla. Perusjoukon muodostavat äänestysikäiset kansalaiset, joista jokaisella on yhtäläinen mahdollisuus osallistua tähän ”kokonaistutkimukseen” äänestämällä. Kaikki eivät äänioikeuttaan käytä, joten lopullisiin vaalituloksiin sisältyy runsaasti spekuloinnin mahdollisuuksia.

Varsinaista kokonaistutkimusta Suomen väestöstä tekevät lähinnä väestötieteilijät, mutta rajatumpien perusjoukkojen osalta on muidenkin mahdollista tehdä kokonaistutkimusta. Se voi olla kustannuksiltaan kalliimpaa, mutta toisinaan perusteltua, kun halutaan minimoida otannasta johtuvat epävarmuudet. Kokonaistutkimuksen aineistohan ei ole otos, vaan se kuvaa suoraan perusjoukkoa. Kyselyn tapauksessa käy kuitenkin samoin kuin vaaleissa: kaikki eivät vastaa, jolloin tiedonkeruuseen jää aukkoja ja sitä myöten epävarmuuksia.

Suomessa on kansainvälisestikin poikkeuksellisen kattavat ja monipuoliset rekisterit, joita ylläpitävät viralliset tahot kuten Väestörekisterikeskus ja Tilastokeskus. Rekistereistä voidaan sekä poimia otoksia että tehdä kokonaistutkimuksia monissa sellaisissa tilanteissa, joissa vastaajilta ei tarvitse kysyä mitään. Tällöin on periaatteessa yhtä helppo tehdä saman tien kokonaistutkimus.

Rekisteritietojen avulla voidaan myös täydentää otannalla kerättyjä aineistoja ja parantaa tältä osin tietojen kattavuutta ja luotettavuutta. Tietosuojasäädökset rajoittavat kuitenkin huomattavasti rekisterien käyttöä ja varsinkin niiden yhdistelyä. Haasteita asettaa myös se, että rekisteriaineistot on yleensä suunniteltu vain hallinnolliseen käyttöön eivätkä ne siten välttämättä sovellu tutkimustarkoituksiin.

Mittauksesta johtuva epävarmuus on läsnä joka tapauksessa. Jos on varaa kokonaistutkimukseen, ei kannata pilata sitä huonolla mittauksella. Tutkimuksen suunnittelun ja toteutuksen yhteydessä eri virhelähteet ovat erotettavissa, mutta lopulta ne kaikki yhdessä muodostavat edellä mainitun virhemarginaalin ja koko perustan tulosten luotettavuudelle.

2.5.3 Näyteaineistot

Jos perusjoukkoa on vaikea tai mahdoton määritellä, ei pidä puhua otoksestakaan. Mihin tahansa kyselyyn voidaan saada ”satunnaisia” vastauksia, verkossa toteutettuna runsaastikin, mutta satunnaisuus ei tässä tarkoita samaa kuin edellä. Vastaukset ovat pikemminkin sattumanvaraisia.

Harkinnanvarainen näyte

Aineistoja, jotka eivät täytä otoksen kriteerejä, kutsutaan *näytteiksi*. Jos etukäteen päätetään, keille tutkimuksen tarpeisiin soveltuville vastaajille kysely suunnataan, kyseessä on *harkinnanvarainen näyte*. Esimerkiksi voidaan lähettää osalle tietyn tuotteen rekisteröityneistä käyttäjistä paperinen tai sähköinen kirjekysely, jolloin voidaan saada vastauksia tuotteen tuntevilta. Tehtävät johtopäätökset rajoittuvat lähinnä kyselyyn vastanneisiin, joskin houkutus päätellä jotain yleisempää tuotteen käyttäjistä on varmasti suuri.

Sattumanvarainen näyte

Jonkin verkkosivun lukijoista tiettynä ajankohtana voidaan saada näyte tarjoamalla osalle vierailijoista kyselylomake esimerkiksi erillisessä selainikkunassa. Vaikka tässä on myös jotain harkintaa, kyseessä on *sattumanvarainen näyte*, riippuen paljon siitä, mitä kysytään ja mihin sivulla vierailevat kiinnostuvat vastaamaan. Tämäntyyppistä aineistoa kutsutaan myös *itse valikoituvaksi näytteeksi*.

Verkossa tehtävillä kyselyillä saatetaan helposti saada kokoon tuhansia, jopa kymmeniätuhansia vastauksia. Aineiston koko ei kuitenkaan sinänsä oikeuta sen ihmeellisempiin johtopäätöksiin. Näyte kertoo varmasti jotain keruuaikana sivulla vierailleista kävijöistä, mutta yliampuvia tulkintoja tunnutaan tekevän liian helposti, esimerkiksi kuvittelemalla, että aineiston koko jotenkin oikeuttaisi tekemään johtopäätöksiä ”kaikista suomalaisista”.

Ilman aitoa otanta-asetelmaa johtopäätösten yleistäminen on vain tutkijan asiantuntemuksen varassa. Yhden tutkimuksen tai aineiston perusteella on luultavasti mahdoton sanoa mitään kovin varmaa. Lisää tukea johtopäätöksille voidaan saada esimerkiksi toistamalla tutkimus ja kenties kohdentamalla se tarkemmin. Tulosten yleistäminen ei ole ilmoitusasia, varsinkaan näyteaineistolla.

2.6 Kyselylomake tiedonkeruuvälineenä

Kyselylomaketta mittausvälineenä käsiteltiin kohdassa 2.2, mutta siihen liittyy myös tiedonkeruun näkökohtia, joita seuraavassa tarkastellaan lyhyesti. Näkökohdat eivät ole lainkaan tilastollisia, mutta niillä on sitäkin suurempi merkitys kokonaisuuden kannalta.

Saatekirjeen merkitys

Saatekirje on kyselytutkimuksen julkisivu. Se kertoo vastaajalle tutkimuksen perustiedot, siis mistä tutkimuksessa on kysymys, kuka tutkimusta tekee, miten vastaajat on valittu ja mihin tutkimustuloksia tullaan käyttämään.

Saatekirjeen merkitystä ei voi aliarvioida, koska sen perusteella vastaaja voi joko motivoitua vastaamaan kyselyyn tai hylätä koko lomakkeen. Ei siis välttämättä auta, vaikka lomake olisi kuinka hyvä. Vastaaja ei ehkä edes vilkaise sitä, jos saatekirje on epämääräinen tai ylimalkainen. Ehkä paras vastaamismotivaatio tulee siitä, että aihe kiinnostaa jo valmiiksi, mutta hyvin laaditulla saatekirjeellä voi herättää vastaajan kiinnostuksen ja vaikuttaa vastausten luotettavuuteen.

Verkkolomake vai paperilomake?

Verkkolomakkeet ovat voimakkaasti yleistyneet, ja niillä on kieltämättä paljon hyviä puolia. Kun vastaukset tallettuvat suoraan sähköiseen muotoon, niitä ei tarvitse erikseen tallentaa. Paperilomakkeiden tallentaminen on aikaa vievä ja virhealtis vaihe. Toisaalta verkkolomakkeisiin voi liittyä tavoitettavuusongelmia. Edustavaksi tarkoitettu otos voi valikoitua sen mukaan, onko käytettävissä tietokonetta ja verkkoyhteyttä tai kokeeko verkossa vastaamisen luontevaksi vastaustavaksi. Joskus voi olla paikallaan toimittaa vastaava lomake myös paperiversiona.

Vastausväsymys ja lomakkeen testaus

Kun kyselyiden määrä on jatkuvasti kasvanut, on alkanut ilmetä vastausväsymystä. Tutkimusten vastausprosentit ovat huonontuneet huolestuttavasti. Monesti olisi syytä miettiä vakavasti, olisiko mahdollista tiivistää lomaketta ja jättää osa kysymyksistä pois. Vastaaminen pitäisi tehdä mahdollisimman helpoksi, sillä harva viitsii käyttää aikaansa pitkien lomakkeiden kanssa painimiseen. Myös kielen selkeyteen ja lomakkeen ulkoasuun kannattaa kiinnittää paljon huomiota.

Luultavasti kyselyn laatijakin väsähtää jossain vaiheessa eikä kykene havaitsemaan kaikkia mahdollisia ongelmatilanteita, joita vastaajille voi tulla. Kyselylomaketta onkin ehdottomasti testattava etukäteen. Hyviä testaajia ovat tutkimuksen kohderyhmään kuuluvat, sillä tällöin saadaan todenmukainen käsitys siitä, onko kysymykset ja ohjeet ymmärretty oikein, onko lomakkeessa turhia kysymyksiä ja onko jotain olennaista kenties jäänyt kysymättä. Testaajia ei tarvitse olla kovin paljon – muutamakin riittää, jotta ainakin pahimmat ongelmat saadaan korjattua ennen varsinaista tiedonkeruuta.

Lomakkeen testauksen vaiheita käsittelee muun muassa Fowler, Jr. (1995, 104–137). Laajalti tietoa näistä asioista tarjoaa kyselylomakkeiden suunnitteluun ja testaamiseen erikoistunut *SurveyLaboratorio* (Tilastokeskus, 2005). Perusteellisesti lomakkeiden testausmenetelmiä kartoittaa myös *Methods for Testing and Evaluating Survey Questionnaires* (Presser ym., 2004), jonka kirjoittajina on tutkijoita Yhdysvaltojen, Kanadan ja Euroopan tilastokeskuksista, yliopistoista ja tutkimuslaitoksista.

Ulkonäkötutkimuksen kyselylomake

Ulkonäkötutkimuksen kyselylomake on paperilomake, pituudeltaan 24 sivua. Se sisältää tyypillisten taustakysymysten lisäksi useita mittareita sekä yksittäisiä osioita, jotka mittaavat naisten ulkonäköön liittyviä käsityksiä, kokemuksia, käyttäytymistä ja asenteita. Valtaosa on kansainvälisiä ulkonäkötyytyväisyyden mittareita, joita on kehitetty viimeisen 50 vuoden aikana. Kaikkiaan mittarit koostuvat yli 200 osiosta, joista suurin osa on viisiportaisia Likertin asteikkoja. Lisäksi lomakkeeseen sisältyy kansainvälisten esikuvien pohjalta itse kehitettyjä mittareita. Lomakkeeseen voi tutustua tarkemmin kirjan kotisivulla.

Tässä kirjassa viitataan vain osaan ulkonäkötutkimuksessa käytetyistä mittareista. Tarkemmin niistä kertoo Maarit Valtarin valmisteilla oleva väitöskirja sekä pro gradu -työ *Suomalaisten naisten ulkonäkötyytyväisyys* (Valtari, 2001).

Lomakkeesta aineistoksi

Tiedonkeruun myötä siirrytään tutkimusaineiston tarkasteluun. Täysin automaattisesti se ei tapahdu, vaan aineiston muodostaminen on oma työvaiheensa. Voidaan puhua aineiston perustamisesta, koska tavallaan aineisto rakennetaan perustuksista lähtien. Perustamisvaihe on tehtävä huolellisesti, koska aineisto on jatkossa kaiken työskentelyn keskipiste.

Kyselytutkimuksessa aineisto rakennetaan kyselylomakkeen pohjalta. Paperilomakkeen tapauksessa luodaan vastaavan rakenteinen havaintotiedosto, annetaan osioita vastaaville muuttujille sopivat nimet ja tallennetaan tiedot. Verkkolomakkeella toiminta on suoraviivaisempaa, koska aikaavievä ja virhealtis tallennusvaihe jää pois.

Käytettävästä järjestelmästä riippuen muuttujien nimetkin voidaan päättää jo valmiiksi lomaketta laatiessa.

Ohjelmistoilla on omat tiedostomuotonsa, joita muut ohjelmistot eivät pääsääntöisesti ymmärrä. Ainakaan ei kannata liikaa luottaa ohjelmistojen kykyyn lukea muiden ohjelmistojen tiedostoja. Versioiden vaihtuessa yhteydet voivat varoituksetta lakata toimimasta – toisinaan jopa saman ohjelmiston eri versioiden välillä.

Yleispätevä siirtomuoto eri ohjelmistojen ja niiden versioiden sekä käyttöjärjestelmien ja aikakausien välillä on tekstitiedosto, jota yleensä kaikki tilastolliset ohjelmat, verkkolomakejärjestelmät, jopa tekstinkäsittelyohjelmat ja taulukkolaskimet osaavat käsitellä. Tekstitiedosto ei kuitenkaan sovellu aineiston analysointiin, vaan tiedot on siirrettävä käytettävän ohjelmiston omaan muotoon. Aineiston perustamista käsitellään yksityiskohtaisemmin liitteessä A.

Kun aineisto vihdoin on olemassa, alkaa mielenkiintoinen vaihe, jossa valmistaudutaan varsinaisiin tilastollisiin analyyseihin tutustumalla aineistoon perinpohjaisesti.

3 Aineiston esikäsittely

Kun aineisto on koossa, siihen päästään tutustumaan piirtämällä kuvia, tekemällä taulukoita ja tutkimalla tunnuslukuja. Tämä on vaihe, jossa ei pidä hätäillä. Vaikka tekisi jo mieli rynnätä regressioanalyyseihin, on viisainta malttaa mielensä. Perusteellinen aineiston *esikäsittely* luo pohjan varsinaisille analyyseille. Samalla se auttaa löytämään virheitä; niitä ei voi kokonaan välttää. Virheet kannattaa yrittää korjata saman tien, sillä myöhemmin ne aiheuttavat enemmän vaikeuksia.

3.1 Aineistoon tutustuminen

Aineistoon tutustuminen on hyvä aloittaa yksinkertaisesti selailemalla ja katselemalla, miltä aineiston sisältämät tiedot näyttävät. Perusteellisin tapa on tallentaa koko aineisto itse, jolloin on varmasti nähnyt sen jokaisen luvun ja numeron. Tallentaminen on kuitenkin käynyt harvinaisemmaksi verkkolomakkeiden myötä. Selailu on hyvä tapa varmistaa, että aineistossa on kaikki ainakin päällisin puolin kunnossa.

Havainnot ja muuttujat

Tyypillisin kyselytutkimusaineiston muoto on *havaintomatriisi*. Sen vaakarivejä kutsutaan *havainnoiksi*. Ne koostuvat kyselyyn osallistuneiden henkilöiden vastauksista. Yleensä yhtä vastaajaa kohti on yksi havainto. Havaintomatriisin pystyrivejä kutsutaan *muuttujiksi*. Jokaista kyselylomakkeen osiota vastaa yksi tai useampi muuttuja. Hyvä tapa on sijoittaa ensimmäiseksi muuttujaksi havainnon yksikäsitteinen tunniste, paperilomakkeista esimerkiksi juokseva numero.

Aineistoa on hyvä selailla sekä havaintojen että muuttujien suunnassa. Selailu paljastaa äkkiä ainakin laajat tietojen puuttumiset, järjettömän oloiset arvot ja monet muut outoudet, joita saattaa löytyä useimmista aineistoista, tietojen tallennustavasta riippumatta.

Ulkonäkötutkimuksen havainnot ja muuttujat

Ulkonäkötutkimuksen aineistossa on 496 havaintoa ja 454 muuttujaa. Osa kysymyksistä on koodattu useina muuttujina. Erilaisten muunnosten myötä muuttujien määrä nousee yli viidensadan, mutta jo aluksi yksittäisiä tietoja, joko numeerisia tai sanallisia, on *yli kaksisataatuhatta*. Pelkkä selailu ei riitä, vaan tarvitaan tehokkaita tapoja tarkastella tietoja tiivistetymmin.

Aineiston perustarkastelut ja esikäsittely

Tässä luvussa perehdytään ensin yksittäisten muuttujien perustarkasteluihin: jakaumiin, tunnuslukuihin ja kuviin. Eräiden muuttujamuunnosten jälkeen jatketaan tarkasteluja kahdella muuttujalla. Lopuksi tutustutaan aineiston muokkauksiin. Mikään aineisto ei ole sellaisenaan ”valmis”. Perustarkastelut ja esikäsittely ovat välttämättömiä vaiheita, jotka pitää viedä läpi ennen varsinaisia analyyseja.

3.2 Yhden muuttujan tarkastelu

Yksittäisistä havaintoarvoista voi päätellä monenlaista, mutta vasta tiivistämällä tietoa saadaan käsityksiä suuristakin tietomääristä. Seuraavassa aineiston yksittäisiä muuttujia tarkastellaan tiivistettyinä erilaisten jakaumien, tunnuslukujen ja kuvien avulla.

3.2.1 Jakaumat

Aineistoon tutustuminen kannattaa aloittaa muuttujien *jakaumista*, joista näkee nopeasti, mitä arvoja mikäkin muuttuja sisältää. Jakaumista voidaan edetä tiivistämällä arvoja tunnusluvuiksi, mutta siitä ei pidä mennä aloittamaan; se olisi kuin työskentelisi silmät ummessa.

Tuloste 3.1 tiivistää tiedot vastaajien syntymävuodesta luokittaisia lukumääriä kuvaavaksi *frekvenssijakaumaksi* ja vastaavaksi *prosenttijakaumaksi*. Luokkien ylärajoista (`up.limit`) nähdään, että syntymävuosi on luokiteltu tasavälisesti 14 luokkaan viiden vuoden välein. Jakaumien vieressä olevasta yksinkertaisesta vaakapylväskuvasta nähdään yhdellä silmäyksellä, että jakauma vaikuttaa muodoltaan melko symmetriseltä. Suurimman luokan muodostavat vuosina 1956–1960 syntyneet, joita aineistossa on 62 (12.5 %).

Tuloste 3.1. Syntymävuoden luokiteltu jakauma ja tunnusluvut.

```
Basic statistics: UN2007 N=496
Variable: sv       syntymävuosi
Interval scale
min=1923     in obs.#6
max=1987     in obs.#383
mean=1957.893 stddev=14.93334 skewness=-0.101139 kurtosis=-0.834285
lower_Q=1946.95  median=1957.782 upper_Q=1969.573
up.limit      f     %      *=2 obs.  class width=5
    1925      3   0.6 *
    1930     15   3.0 *******
    1935     19   3.8 *********
    1940     32   6.5 ****************
    1945     46   9.3 ***********************
    1950     50  10.1 *************************
    1955     49   9.9 ************************
    1960     62  12.5 ******************************
    1965     59  11.9 *****************************
    1970     41   8.3 ********************
    1975     49   9.9 ************************
    1980     39   7.9 *******************
    1985     27   5.4 *************
    1990      5   1.0 **
```

s. 203

Frekvenssi- ja prosenttijakaumien etu on se, että niitä voidaan käyttää mittaustasosta riippumatta. Jos muuttuja on jatkuva (ks. kohta 2.2.3, s. 26), on tulkinnan kannalta usein mukavampaa siirtyä luokiteltuun jakaumaan, ikään kuin karkeistaa alkuperäiset mittaukset järjestys- tai jopa luokittelutasolle. Toisinpäin tämä ei tietenkään toimi, sillä mittauksia ei voi jälkikäteen tarkentaa. Luokittelu tiivistää tarkasteluja ja tekee ne ymmärrettävämmiksi. Syntymävuoden luokittelua voisi edelleen tihentää muodostamalla vaikka seitsemän luokkaa 10 vuoden välein, riippuen siitä mihin huomio halutaan kohdistaa.

3.2.2 Tunnusluvut

Jakaumien lisäksi muuttujia voidaan tarkastella tiivistämällä niitä tilastollisiksi tunnusluvuiksi. Tunnuslukuja on paljon, mutta muutama keskeisin riittää käytännön tarpeisiin. Sopivien tunnuslukujen valintaan vaikuttaa muun muassa muuttujan mittaustaso.

Edellä esitettyyn tulosteeseen 3.1 sisältyy myös joukko syntymävuoden tunnuslukuja. Keskeisimmät tunnusluvut ovat *minimi* (`min`) eli pienin arvo ja *maksimi* (`max`) eli suurin arvo sekä *keskiarvo* (`mean`), *keskihajonta* (`stddev`) ja havaintojen lukumäärä (`N`).

Minimi ja maksimi ovat *järjestystunnuslukuja*, joihin kuuluvat myös *mediaani* (`median`) eli keskimmäinen arvo sekä *ala-* ja *yläkvartiilit* eli järjestetyn aineiston neljännekset (`lower_Q`, `upper_Q`). Näihin kaikkiin perehdytään tämän luvun kuluessa tarkemmin. Sen sijaan *vinous* (`skewness`) ja *huipukkuus* (`kurtosis`) eivät ole niin keskeisiä tunnuslukuja. Jakaumien muotoa on kätevämpi arvioida kuvista kuin tunnusluvuista.

Keskiarvo ja keskihajonta

Kaikkein yleisin tunnusluku on tavallinen *keskiarvo*, joka on sinänsä helppo käsittää: keskiarvo kuvaa muuttujan keskimääräistä arvoa. Koska se muodostetaan laskemalla muuttujan arvot yhteen ja jakamalla summa havaintojen lukumäärällä, edellytetään numeerista mittaustasoa (ks. kohta 2.3.3, s. 34). Tulos voi silti olla yllättävän hankala tulkita. Jos muuttujan jakauma on kovin vino tai muuten erikoinen, keskiarvo ei anna järkevää kuvaa muuttujasta. Jakauman muoto selviää parhaiten kuvista, joihin perehdytään kohdassa 3.2.3 (s. 61).

Erittäin moniin tilanteisiin keskiarvo kuitenkin sopii, joten se on ainakin ”keskeinen” tunnusluku. Pelkkä keskiarvo ei kuitenkaan tule kysymykseen; on myös nähtävä, mitä sen ympärillä tapahtuu, siis paljonko ja millaista *vaihtelua* muuttujan arvoissa esiintyy.

Vaihtelun tutkiminen on keskeisellä sijalla tilastotieteessä. Teoreettisemmissa tarkasteluissa puhutaan *varianssista*, mutta käytännössä paras kumppani keskiarvolle on *keskihajonta*, koska ne molemmat ilmaistaan samoina yksiköinä kuin muuttujan arvot on mitattu. Kumpaakaan ei ole mieltä tuijotella pelkästään vaan pikemminkin tunnuslukuparina. Niiden käyttöedellytykset ovat samat ja tulkinta vastaavanlainen: keskihajonta kuvaa keskimääräistä hajontaa.

Vaihtelun laatu ja määrä

Keskiarvon ja keskihajonnan avulla muodostuu jo jonkinlainen kuva muuttujasta. Mitä pienempi keskihajonta on, sitä tiiviimmin arvot ovat sijoittuneet keskiarvon ympärille. Tilanteesta riippuu, onko pieni hajonta hyvä vai huono asia. Periaatteessa mitä suurempi on hajonta, sitä enemmän muuttujassa on määrällistä informaatiota. Kaikki vaihtelu ei kuitenkaan ole samanarvoista, sillä osa voi johtua esimerkiksi mittausvirheestä. Vaihtelun tarkastelussa on kiinnitettävä huomiota sekä määrään että laatuun. Näihin aiheisiin palataan tässä kirjassa useasti.

Ääritapauksessa keskihajonta on nolla, jolloin muuttuja on pelkkä *vakio*, toisin sanoen kaikki sen arvot ovat samoja. Tällöin pelkkä keskiarvo olisi todella onneton tunnusluku; eihän siitä kävisi millään tavoin ilmi, että jakauma on täysin surkastunut. Vakio välittää vain laadullisen tiedon, esimerkiksi ”kaikki vastasivat samalla tavalla”, mikä voi tietenkin olla sisällöllisesti hyvinkin kiinnostavaa. Tilastolliseen käyttöön tällainen muuttuja ei kelpaa, koska siinä ei ole lainkaan määrällistä informaatiota.

Tuloste 3.2. Keskeisimpiä tunnuslukuja taulukkona.

s. 203

	N	Minimum	Maximum	Mean	Std. Deviation
syntymävuosi	496	1923	1987	1957.89	14.933
Valid N (listwise)	496				

Kattavamman käsityksen muuttujan vaihtelusta saa tarkastelemalla sen *vaihteluväliä*, jonka rajaavat minimi ja maksimi. Tulosteesta 3.2 nähdään syntymävuoden keskeisimmät tunnusluvut hieman eri muodossa kuin edellä. Tunnusluvuista voidaan päätellä vaihteluvälin olevan 1923–1987. Keskiarvo on 1958 ja keskihajonta 15, joten keskimääräinen hajonta molempiin suuntiin keskiarvosta kattaa vaihteluvälistä vuodet 1943–1973. Koska jakauma on aiemman tulosteen 3.1 perusteella melko symmetrinen, niin tällainen väli kattaa yleensä noin kaksi kolmasosaa koko vaihteluvälistä.

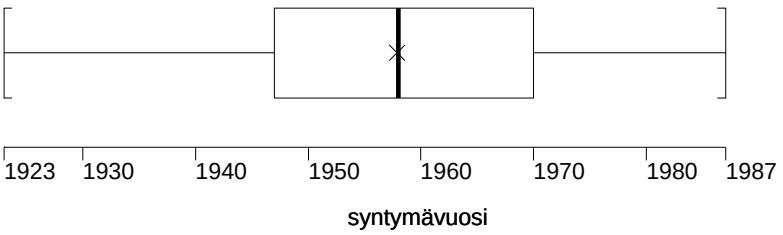
Jos tarkasteltavaa väliä levennetään keskiarvosta kumpaankin suuntaan *kahden hajonnan* päähän, saadaan alue, joka yleensä kattaa noin 95 % vaihteluvälistä. Tämä alue ja sen ulkopuolelle jäävät viisi prosenttia ovat huomion keskipisteenä erittäin monissa tilastollisissa analyyseissa. Syntymävuoden osalta kahdella hajonnalla suuntaansa päästään väliin 1928–1988, joka tutkimuksen otanta-asetelmasta johtuen (ks. kohta 2.5.1, s. 43) kattaa jo lähes koko vaihteluvälin.

Laatikollinen lisää lukuja

Keskiarvo saattaa olla huono tunnusluku, vaikka muuttujan mittaustaso olisi riittävä. Jos jakauma on vino tai sisältää huomattavan pieniä tai suuria arvoja, keskiarvo vääristyy herkästi. Keskiarvo ja keskihajonta käyvät käsi kädessä, joten jos keskiarvo ei käy, ei käy keskihajontakaan.

Keskikohdan ja vaihtelun hahmottaminen ilman keskiarvoa ja keskihajontaa tapahtuu siirtymällä *järjestystunnuslukuihin*, joissa muuttujan arvoja tutkitaan suuruusjärjestyksessä. Tärkeimmät järjestystunnusluvut ovat jo edellä mainitut minimi, ensimmäistä neljännestä (25 %) osoittava alakvartiili, keskimmäistä havaintoa merkkaava mediaani, kolmatta neljännestä (75 %) osoittava yläkvartiili ja maksimi.

Vaihteluväli voidaan jakaa muillakin tavoin, mutta tavallisin on jako neljään, sillä se on yhteydessä *laatikkokuvaksi* (*box-plot*) kutsuttuun graafiseen esitykseen (ks. kuva 3.1). Laatikkokuva on niin kiinteä osa järjestystunnuslukujen tarkastelua, että se esitetään jo tässä vaiheessa. Muihin kuviin tutustutaan kohdassa 3.2.3 (s. 61).



Kuva 3.1. Laatikkokuva syntymävuoden jakaumasta.

Laatikkokuva voidaan tehdä vaaka- tai pystysuuntaisena. Pystykuva on tyypillisempi, mutta tilan säästämiseksi kuva 3.1 on piirretty vaakatasoon. Kuva visualisoi muuttujan jakauman kätevästi pelkkien järjestystunnuslukujen avulla. Keskellä olevan laatikon vasen reuna vastaa alakvartiilia ja oikea yläkvartiilia. Puolet havainnoista jää tällöin laatikon sisään muodostaen *kvartiilivälin*. Paksu viiva symboloi mediaania. Rastilla kuvaan on merkitty myös keskiarvo.

Kuvasta 3.1 nähdään, että syntymävuoden mediaani, 1958, on sama kuin sen keskiarvo (vrt. tuloste 3.1, s. 53). Siitäkin voidaan päätellä, että jakauma on symmetrinen, toisin sanoen keskikohdan molemmilla puolilla on suunnilleen yhtä paljon havaintoja. Kvartiiliväli 1947–1970 on hieman kapeampi kuin aiemmin tarkasteltu suuntaansa yhden hajonnan väli.

Järjestystunnusluvut soveltuvat tilanteisiin, joissa keskiarvo ja keskihajonta vääristyvät, sillä järjestystunnusluvut eivät ole niin herkkiä jakaumien vinoudelle ja poikkeaville arvoille. Niitä voidaan käyttää myös järjestystasoisten muuttujien kuvailuun, johon keskiarvo ja keskihajonta eivät sovellu.

Kaikki edellä tarkastellut syntymävuoden tunnusluvut ovat lopulta aika hankalia tulkita. Syynä ei ole muuttujan mittaustaso vaan sen sisältö. Syntymävuotta kiinnostavampaa olisi tarkastella vastaajan ikää – etenkin kun tiedot on kerätty eri vuosina. Tähän palataan kohdassa 3.3 (s. 64).

Luokittelutason tunnusluvut

Luokittelutason muuttujien tunnuslukujen tulkinnassa on syytä olla tarkkana. Esimerkiksi ulkonäkötutkimuksessa vuonna 2005 kysytty maakunta on koodattu lukuina 1–19. Kyseessä on luvuista huolimatta luokittelutasoinen muuttuja, joten keskiarvojen laskemiseen tai ”tulkintaan” ei ole perusteita. Mitä muka kertoisi, että *”maakunta on keskimäärin 13”*, kun arvoa 13 vastaa Pohjois-Karjala? Maantieteellisiä tulkintoja ei kannata yrittää, koska luvut vastaavat nimiä aakkosjärjestyksessä; esimerkiksi 1 on Etelä-Karjala ja 19 Varsinais-Suomi. Mediaani ei ole tässä tapauksessa yhtään parempi kuin keskiarvo, koska aakkosjärjestys ei ole sisällöllinen järjestys.

Sen sijaan minimi ja maksimi ovat käyttökelpoisia tunnuslukuja jo luokittelutasolla, ainakin jos muuttujat on koodattu numeroina.

Niistä nähdään helposti, jos joukkoon on eksynyt minimin alittavia tai maksimin ylittäviä tietoja.

Yleiskäyttöisiä tunnuslukuja ovat myös havaintojen lukumäärät ja prosenttiosuudet. Joskus mainitaan myös tyypillisimmän luokan frekvenssi, jota kutsutaan *moodiksi* tai *tyyppiarvoksi*. Sellaisenaan se on melko vähäinen tieto. Parempia esitystapoja tarjoaa tilastollinen grafiikka, esimerkiksi pylväskuvat (ks. kohta 3.2.3, s. 63). Usean muuttujan yhtaikaisen tarkastelun myötä alkaa löytyä muitakin aineiston tiivistämiskeinoja.

Määrää ja laatua samassa

Ennen kuin siirrytään kuvallisempiin tarkasteluihin, tutkitaan tilannetta, jossa määrällinen ja laadullinen tieto kietoutuvat ovelasti samaan muuttujaan. Tunnuslukujen kanssa on tällöin syytä olla entistäkin tarkkaavaisempi.

Tuloste 3.3. Yhdessäolon jakauma ja tunnusluvut.

```
Variable: yhdessa  Yhdessäoloaika (vuosina)
N(missing)=13
Ratio scale
min=0        in obs.#4
max=53       in obs.#257
mean=14.07764 stddev=13.53234 skewness=0.761888 kurtosis=-0.434746
lower_Q=1.390625 median=10.33333 upper_Q=23
up.limit      f     %      *=4 obs.  class width=10
       0     99  20.5 ***********************
      10    144  29.8 ***********************************
      20     98  20.3 ***********************
      30     66  13.7 ***************
      40     54  11.2 ************
      50     20   4.1 *****
      60      2   0.4 :
```

Asiaa havainnollistaa tuloste 3.3, jossa on tietoja yhdessäoloajasta nykyisen kumppanin kanssa. Tieto on mitattu avoimella kysymyksellä ja vastaukset koodattu vuosina yhden desimaalin tarkkuudella. Muuttujan mittaustaso sallii kaikkien edellä käsiteltyjen tunnuslukujen soveltamisen. Keskiarvo on 14 vuotta ja mediaani 10, joten jakauma on vino – sen näkee heti karkeasta kuvastakin. Kun keskiarvo vääristyy, on mediaani parempi keskiluvun kuvaaja.

Haasteelliseksi tällaisissa muuttujissa saattaa muodostua nolla. On paikallaan miettiä, mitä nolla kulloinkin sisällöllisesti tarkoittaa. Tässä tapauksessa ”nollan vuoden yhdessäoloaika” viittaa ilmeisesti siihen, että vastaaja ei ole ollut vakituisessa parisuhteessa. Näitä vastaajia aineistossa on 99, mikä on pääteltävissä tulosteesta 3.3, sillä nolla on samanaikaisesti muuttujan minimi ja frekvenssijakauman ensimmäisen luokan yläraja. Kyseiseen luokkaan sisältyy siis ainoastaan nollahavaintoja. Vastaajista 13 näyttää jättäneen vastaamatta kysymykseen.

Tuloste 3.4 näyttäisi kertovan saman täsmällisemmin, sillä parisuhteen statusta on tiedusteltu myös erikseen. Samalla havaitaan, että parisuhdemuuttujan keskiarvo on 1.8, mikä on kohtalaisen kömpelö tapa ilmaista, että suurin osa vastaajista on ilmaissut elävänsä vakituisessa parisuhteessa. Dikotominen muuttujahan edustaa periaatteessa mitä tahansa mittaustasoa, kuten luvussa 2 (s. 39) todettiin. Parisuhteen statusta olisi silti selvempää tyytyä ajattelemaan vain luokittelutasoisena muuttujana ja jättää tunnusluvut suosiolla väliin.

Tuloste 3.4. Vakituisen parisuhteen perustiedot.

```
Variable: pari     Vakituinen parisuhde (1=ei, 2=kyllä)
N(missing)=1
Dichotomous variable
mean=1.8      stddev=0.400405 skewness=-1.498484 kurtosis=0.243434
pari          f     %      *=8 obs.
       1     99  20.0 ************
       2    396  80.0 **************************************************
```

Taulukoihin tutustutaan kohdassa 3.4.1 (s. 68), mutta jo tässä tekee mieli katsoa, ovatko tulosteessa 3.4 esiintyvät 99 vastaajaa *samat* 99, jotka edellä havaittiin (vrt. tuloste 3.3). Yksittäisistä muuttujista tällaista ei voi aukottomasti päätellä. Sen sijaan tulosteissa erikseen esiintyneet muuttujat `yhdessa` ja `pari` voidaan taulukoida vastakkain, jolloin nähdään, miten niiden jakaumat menevät ristiin toistensa kanssa. Taulukointia kutsutaankin myös ristiintaulukoinniksi.

Tulosteessa 3.5 edellä mainitut muuttujat on taulukoitu ristiin. Taulukko todistaa selvästi, että kyseessä todellakin ovat samat 99 vastaajaa. Samalla se tarkentaa myös puuttuvista tiedoista edellä tehtyä huomiota: 13 vastaajasta yksi on jättänyt vastaamatta molempiin, loput 12 vain yhdessäoloaikaa koskevaan kysymykseen. Puuttuvista tiedoista aiheutuu merkittäviä hankaluuksia, joita käsitellään tarkemmin kohdassa 3.5 (s. 81).

Tuloste 3.5. Parisuhdemuuttujat ristiintaulukoituna.

```
     yhdessa      ei      on       ?
pari *******
ei                99       0       0
on                 0     384      12
?                  0       0       1
```

s. 204

Kaikkiaan on siis huomattava, että yhdessäoloaikaa kuvaava muuttuja ei ole pelkkä määrällinen jatkumo, vaan se sisältää myös laadullisen tiedon vakituisesta parisuhteesta. Leikillisesti tätä voisi kutsua todelliseksi ”suhdeasteikoksi”! Asiaan on kuitenkin syytä suhtautua vakavasti ja ottaa muuttujan kaksoisrooli huomioon kaikissa siitä tehtävissä laskelmissa. Esimerkiksi edellä mainitut keskiarvo ja mediaani antavat harhaanjohtavan kuvan asiasta, koska nollat on laskettu mukaan. Ilman nollia – siis rajaten tarkastelut vain vakituisessa parisuhteessa oleviin – yhdessäoloajan keskiarvo olisi 18 vuotta ja mediaani 15. Ero on huomattava aiemmin mainittuihin lukuihin 14 ja 10 verrattuna.

Samantyyppisiä muuttujia ovat erilaiset lukumäärät tai vaikkapa liikunnan harrastamisen aktiivisuus (ks. esimerkki 2.6, s. 31). Usein nolla on laadullisesti eri asia, ja määrällinen jatkumo alkaa vasta sen jälkeen.

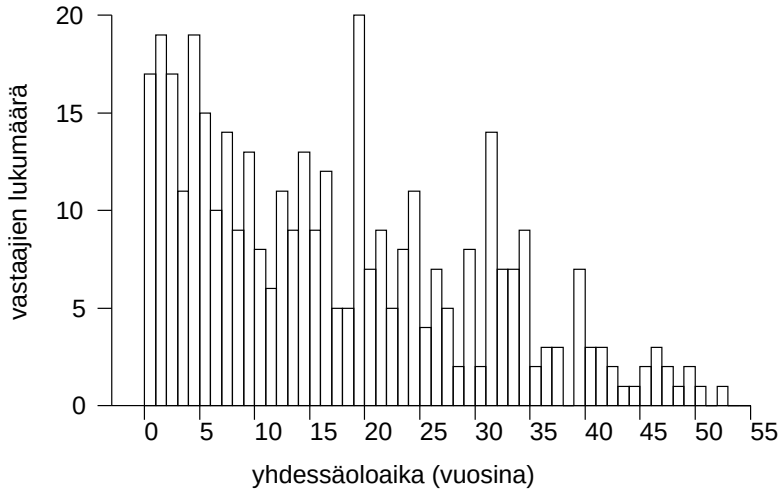
On hyvä muistaa, että tunnuslukuihin tiivistäminen myös kadottaa informaatiota, joten siinä ei pidä mennä liian pitkälle. Yksi luku ei välttämättä kuvaa kovin hyvin mitään. Sen sijaan yksi kuva voi kertoa enemmän kuin kuuluisat tuhat sanaa – tai tunnuslukua.

3.2.3 Kuvat

Jakaumien ja tunnuslukujen yhteydessä esiintyi tulosteita, joihin sisältyi merkkipohjaisia pylväskuvia. Ne ovat käteviä ennen kaikkea aineistoon tutustuessa, ei niinkään tulosten raportoinnissa. Järjestystunnuslukujen yhteydessä esitettyyn laatikkokuvaan palataan vielä kahden muuttujan kuvien yhteydessä kohdassa 3.4.2 (s. 71). Seuraavassa perehdytään erikseen jatkuville ja diskreeteille muuttujille soveltuviin yhden muuttujan pylväskuviin.

Histogrammi

Jatkuvan muuttujan pylväskuvaa kutsutaan *histogrammiksi*. Jatkuvuutta ilmennetään histogrammissa piirtämällä pylväät kiinni toisiinsa. Pystyakseli kuvaa yleensä havaintojen lukumäärää. Kuten kohdassa 2.2.3 (s. 26) todettiin, jatkuvuus on tulkinnanvaraista; joissain tapauksissa diskreetin muuttujan jakaumaa voi olla mielekästä kuvata histogrammilla, jos muuttuja sisältää ”paljon” eri arvoja.

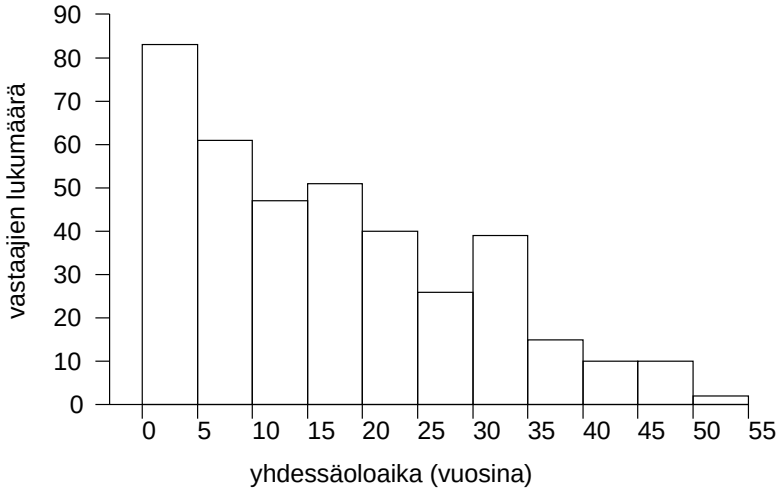


s. 204

Kuva 3.2. Yhdessäolon histogrammi vuoden välein.

Seuraavassa rajoitutaan yhdessäoloaikaa kuvaavan muuttujan jatkuvaan osuuteen, toisin sanoen niihin vastaajiin, joilla on vakituinen parisuhde. Tältä osin muuttuja saa 71 erilaista numeroarvoa, mikä on ilman muuta paljon verrattuna vaikka viisiportaiseen asennemuuttujaan. Osa arvoista esiintyy vain kertaalleen, loput 2–20 kertaa.

Kuvassa 3.2 histogrammi on piirretty vuoden tarkkuudella. Luokitus siis pyöristää desimaaliluvut lähimpään kokonaislukuun. Ensimmäiseen luokkaan kuuluvat korkeintaan vuoden yhdessä olleet, joita aineistossa on 17. Kuvan mukaan yleisin yhdessäoloaika olisi noin 20 vuotta. Kuva on tässä muodossa hyödyksi lähinnä aineistoon tutustumisessa; tietojen raportoimiseen se on turhan yksityiskohtainen.



Kuva 3.3. Yhdessäolon histogrammi viiden vuoden välein.

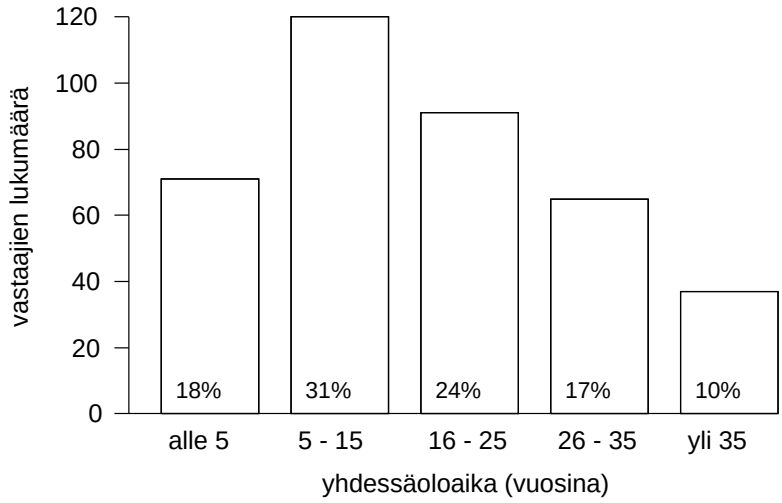
Kuvassa 3.3 esitystä on tiivistetty siirtymällä viiden vuoden tarkkuuteen. Pystyakselin asteikko on luokituksen myötä muuttunut, mutta vaaka-akselin asteikko on sama kuin edellä. Kuvan yleisvaikutelma on nyt selkeämpi. Tällä esitystarkkuudella yleisin yhdessäoloaika näyttäisi olevan nollasta viiteen vuotta. Viimeisessä luokassa on vain yksi havainto. Kuvan perusteella tiedetään vain, että vastaava luku on yli 50, mutta enintään 55. Tarkemman kuvan 3.2 tai tulosteen 3.3 (s. 58)

tunnuslukujen avulla voi päätellä, että kyseessä on aineiston maksimi, 53 vuotta.

Luokituksen tiivistämisestä huolimatta kuvassa 3.3 on ehkä yhä liikaa luokkia. Tarkempi luokitus saattaisi kiinnostaa enemmän jakauman alkupäässä, mutta sen kuvaamiseksi on parempi siirtyä histogrammista pylväskuvaan.

Pylväskuva

Vaaka- ja pystysuuntaiset pylväät ovat tilastollisista kuvista yleisimpiä. Ne soveltuvat diskreetin muuttujan frekvenssi- ja prosenttijakaumien kuvaamiseen. Hyvä oppikirja näiden kuvien historiaan, piirtosääntöihin ja tulkintaan on *Tilastografiikan perusteet* (Kuusela, 2000).



s. 205

Kuva 3.4. Yhdessäolon pylväskuva viisiluokkaisena.

Toisinaan visuaalisesti raskaille pylväskuville on paikallaan harkita vaihtoehtoisia esitystapoja. Niistä ja monista muista tilastollisista kuvista kertoo lisää helppolukuinen teos *Creating More Effective Graphs* (Robbins, 2005).

Histogrammista pylväskuvaan siirtyminen tarkoittaa, että muuttujaa pidetään luokittelu- tai järjestystasoisena ja luovutaan sen mahdollisesta jatkuvuustulkinnasta. Merkkinä tästä ja erona histogrammiin pylväät piirretään erilleen toisistaan. Kuvassa 3.4, joka jatkaa edellä aloitettua yhdessäoloajan jakauman kuvailua, havainnot on luokiteltu viiteen luokkaan. Ensimmäinen ja viimeinen ovat erikokoisia, muut kattavat kukin noin 10 vuotta. Pylväisiin on merkitty luokkien prosenttiosuudet.

Nyt yleisimmältä yhdessäoloajalta näyttää 5–15 vuotta, sillä ensimmäinen luokka ei sisällä tasan viiden vuoden yhdessäoloaikaa. Luokitusten valinta vaikuttaa huomattavasti kuvien ja tulosten tulkintaan ja on erinomainen muistutus siitä, että määrällinen tutkimus ei ole sen objektiivisempaa kuin mikään muukaan tutkimus, vaikka näin toisinaan näkee väitettävän.

3.3 Muunnokset

Aineistoon tutustumisen myötä huomataan tilanteita, joissa alkuperäisiä muuttujia on syytä muuntaa tai koodata uudelleen analysoinnin tai tulkinnan helpottamiseksi. Seuraavassa on eräitä tyypillisiä esimerkkejä.

Ikämuuttujan muodostaminen

Tarkastellaan aluksi syntymävuotta, jonka jakaumia ja tunnuslukuja katsastettiin jo edellä useampaan otteeseen. Ikä kannattaa mitata kysymällä syntymävuotta, kuten esimerkin 2.1 (s. 25) yhteydessä todettiin. Helpommin tulkittava ikämuuttuja syntymävuodesta saadaan vähentämällä se aineiston keruuvuodesta.

Ulkonäkötutkimuksen aineisto on kerätty kahtena eri vuonna. Tiedot ovat samassa aineistossa, mutta keruuvuosi on tallennettu omaksi muuttujakseen. Näin vastaajan ikä saadaan muodostettua koko aineistoon yhdellä muuttujamuunnoksella. Tuloste 3.6 esittää vuoden 1997 osa-aineiston ikäjakauman. Otanta-asetelmasta johtuen vuoden 2005 jakauma on varsin samanlainen, joten sitä ei ole tässä esitetty.

Ikä voitaisiin edelleen luokitella uudelleen tulkinnan helpottamiseksi (vrt. kuva 3.4). Eri tarkoituksiin voidaan tehdä erilaisia luokitteluja ja tallettaa ne aineistoon uusiksi muuttujiksi.

Tuloste 3.6. Ikäjakauma vuoden 1997 aineistosta.

```
Basic statistics: UN2007 N=273
Variable: ika      Vastaajan ikä (vuosina)
Ratio scale
min=18       in obs.#42
max=74       in obs.#6
mean=41.94139 stddev=14.02506 skewness=0.238176 kurtosis=-0.826945
lower_Q=30.44355 median=40.57143 upper_Q=51.92708
up.limit      f     % class width=5
      20     13   4.8 *************
      25     27   9.9 ***************************
      30     26   9.5 **************************
      35     31  11.4 *******************************
      40     36  13.2 ************************************
      45     35  12.8 ***********************************
      50     28  10.3 ****************************
      55     24   8.8 ************************
      60     21   7.7 *********************
      65     15   5.5 ***************
      70     12   4.4 ************
      75      5   1.8 *****
```

Muuttujan suunnan kääntäminen

Toisella usein käytetyllä muunnoksella käännetään muuttuja ”toisinpäin”, esimerkiksi viisiportaisen asennemuuttujan ykkönen muunnetaan viitoseksi, kakkonen neloseksi ja niin edespäin. Muunnostarve johtuu osittain perinteestä kysyä asioita sekä positiivisesti että negatiivisesti, mutta myös paperilomakkeiden suunnittelutottumuksista.

Tilastollisesti muuttujien suunnat ovat jokseenkin samantekeviä, ja on makuasia, kääntääkö niitä vai ei. Kääntäminen samaan, yleensä positiiviseen suuntaan helpottaa monesti tulkintoja.

Tuloste 3.7 sisältää kahden samaa asiaa mittaavan muuttujan frekvenssi- ja prosenttijakaumat koko aineistosta. Muuttuja ”Pidän ulkonäöstäni juuri sellaisena kuin se on” on sanalliselta ilmaisultaan positiivinen ja ”En pidä ulkonäöstäni” negatiivinen, mutta vastausvaihtoehdot ovat samat. Jos jälkimmäinen muuttuja käännettäisiin, se vastaisi vielä selvemmin ensimmäistä. Frekvenssijakauman luvut vain vaihtaisivat paikkoja, esimerkiksi ”täysin samaa mieltä” olisikin 187 vastaajaa ja ”osin samaa mieltä” 154. Keskimmäinen vaihtoehto pysyisi muunnoksessa ennallaan. Myös sanamuoto olisi syytä muistaa kääntää muotoon ”Pidän ulkonäöstäni”.

Tuloste 3.7. Kaksi erisuuntaista asennemuuttujaa.

Pidän ulkonäöstäni juuri sellaisena kuin se on.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Täysin eri mieltä	25	5.0	5.1	5.1
	2 Osin eri mieltä	86	17.3	17.5	22.6
	3 Ei samaa eikä eri	90	18.1	18.3	40.9
	4 Osin samaa mieltä	205	41.3	41.7	82.5
	5 Täysin samaa mieltä	86	17.3	17.5	100.0
	Total	492	99.2	100.0	
Missing	System	4	.8		
Total		496	100.0		

En pidä ulkonäöstäni.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Täysin eri mieltä	187	37.7	38.0	38.0
	2 Osin eri mieltä	154	31.0	31.3	69.3
	3 Ei samaa eikä eri	100	20.2	20.3	89.6
	4 Osin samaa mieltä	40	8.1	8.1	97.8
	5 Täysin samaa mieltä	11	2.2	2.2	100.0
	Total	492	99.2	100.0	
Missing	System	4	.8		
Total		496	100.0		

s. 206

Tulosteen 3.7 taulukoissa esiintyy kaksi prosenttijakaumaa, otsikoilla `Percent` ja `Valid Percent`. Niiden luvuissa on pieniä eroja, joita selvitellään taulukoinnin yhteydessä, kohdassa 3.4.1 (s. 68).

Työllisyystilanteen uudelleenluokittelu

Esimerkissä 2.3 (s. 28) tiedusteltiin työllisyystilannetta yhdeksällä valmiilla ja yhdellä avoimella vaihtoehdolla. Analyysivaiheissa luokkia on usein liikaa, joten muuttujia joudutaan tiivistämään. Luokittelumuuttujilla se tapahtuu yhdistämällä luokkia. Parhaat perusteet yhdistelylle ovat sisällöllisiä, mutta toisinaan joudutaan tiivistämään senkin vuoksi, ettei joissain luokissa ole kuin muutamia havaintoja.

Tulosteessa 3.8 kymmenluokkainen `tyo`-muuttuja on muunnettu uudeksi, viisiluokkaiseksi `tyo5`-muuttujaksi, minkä jälkeen muuttujat on taulukoitu vastakkain. Mitään muunnoksia ei parane tehdä tarkistamatta lopputuloksen vastaavuutta siihen, mitä tavoiteltiin.

Tuloste 3.8. Uudelleenluokittelun tarkistaminen.

```
    tyo5   1   2   3   4   5
tyo ****
  1      227                      1  Kokopäiväinen palkansaaja
  2       33                      2  Osapäivätoiminen palkansaaja
  3            5                  3  Maatalousyrittäjä/työssä maatilalla
  4           23                  4  Muu yrittäjä
  5               37              5  Työtön tai lomautettu
  6                   79          6  Eläkeläinen
  7                       44      7  Opiskelija
  8               18              8  Kotia hoitamassa/kotiäiti
  9               13              9  Äitiysloma
 10               17             10  Jokin muu vaihtoehto
     sum 260  28  85  79  44
```

Tulosteen 3.8 taulukosta näkyy, että ainoastaan eläkeläiset ja opiskelijat muodostavat edelleen omat luokat; kaikkia muita on yhdistelty mielivaltaisesti. Joitakin luokkia olisi todellisuudessa hankala tulkita.

Dokumentointi

Aineistoon tutustuminen ja sen esikäsittely varsinaisia analyyseja varten on laaja ja kauaskantoinen vaihe kokeiluja, tarkisteluja, yrityksiä ja erehdyksiä. Osana varsinaista työskentelyä on tärkeää *dokumentoida* työvaiheet, jotta myöhemmin ei ainoastaan nähtäisi, *mitä* kaikkea on tehty, vaan myös, *miten* kaikki on tehty. Liitteessä A pohditaan aihepiiriä tarkemmin ja näytetään esimerkkeinä kirjan tulosteiden ja kuvien tekoon käytettyjä työkaavioita.

3.4 Kahden muuttujan tarkastelu

Edellä havaittiin jo, että yhden muuttujan tarkasteluissa tulee välittömästi tilanteita, joissa on otettava huomioon myös jonkin toisen muuttujan tietoja. Siirryttäessä tutkimaan kahta muuttujaa samanaikaisesti päästään kiinni tilastollisen tutkimuksen kiintoisimpiin asioihin, muuttujien välisiin yhteyksiin. Kahden muuttujan tarkasteluista päästään luontevasti kohti useampiulotteisia analyyseja, mutta monesti nekin perustuvat olennaisesti kaksiulotteisiin tarkasteluihin.

3.4.1 Taulukot

Taulukot tarjoavat hyviä tapoja tutkia kahden muuttujan yhteyksiä, oli näiden mittaustaso mikä hyvänsä. Luokittelutasonkin muuttujia voidaan mainiosti tutkia taulukoimalla. Taulukoiden tekemistä kutsutaan osuvasti *ristiintaulukoinniksi*. Sitä voidaan pitää yhtenä tärkeimmistä yhteiskuntatutkimuksen perusmenetelmistä, vaikka sitä tässä hyödynnetään vain aineistoon tutustumisessa. Ristiintaulukointiin ja taulukoiden tulkintaan perehdyttää Alkula ym. (1994, 175–219).

Hyvin laadittu taulukko on havainnollinen tapa esittää tietoja tiiviisti. Vastaavat tiedot voidaan havainnollistaa myös kuvallisesti, tyypillisimmin pylväskuvina. Seuraavassa esitetyt taulukot ovat ohjelmien tulosteita, jotka kelpaavat lähinnä aineistoon tutustumiseen. Viimeisteltyjen taulukoiden laatiminen kuuluu tutkimuksen raportointiin ja julkaisemiseen, jota ei käsitellä tässä kirjassa. Hyviä neuvoja näistä aiheista sisältää monilla aloilla sovellettu ohjeisto *Publication Manual of the American Psychological Association* (APA, 2001).

Tuloste 3.9. Kahden asennemuuttujan ristiintaulukko.

		Pidän ulkonäöstäni juuri sellaisena kuin se on.					
		1	2	3	4	5	Total
En pidä ulkonäöstäni.	1	2	7	18	97	62	186
	2	3	29	31	78	12	153
	3	5	33	32	25	5	100
	4	8	13	9	5	4	39
	5	6	3	0	0	2	11
Total		24	85	90	205	85	489

s. 206

Tulosteessa 3.9 on ristiintaulukoituina edellä esitellyt kaksi viisiportaista asennemuuttujaa. Muuttujien suuntia ei ole tässä käännetty, joten niiden arvojen yhdistelmät, *solufrekvenssit*, painottuvat taulukon oikeaan yläkulmaan. Muutamaa ristiriitaiselta vaikuttavaa käsitystä lukuun ottamatta suurin osa vastaajista näyttää pitävän ulkonäöstään.

Tulosteen 3.9 ristiintaulukko edustaa näiden kahden asennemuuttujan *yhteisjakaumaa*, jonka *reunajakaumina* ovat niiden omat frekvenssijakaumat (vrt. tuloste 3.7, s. 66). Tulosteita vertailemalla huomataan kuitenkin, että reunajakaumien luvuissa on pieniä eroja. Tällaisten outouksien tarkistaminen on aineistoon tutustumista parhaimmillaan. Erot saattaisivat johtua virheestä, joka vielä tässä vaiheessa voisi olla helppo korjata. Ainakaan ei kannata antaa asian vain olla, vaan pitää selvittää, mistä erot johtuvat.

Tällä kertaa erot johtuvat *puuttuvista tiedoista*, joihin syvennytään vielä erikseen kohdassa 3.5 (s. 81). Tulosteissa havaitut lukuerot ansaitsevat kuitenkin tulla selvitetyiksi saman tien.

Luvut muuttuvat kun tiedot puuttuvat

Tulosteen 3.9 taulukossa on kaikkiaan 489 havaintoa. Koska aineistossa on aiemmin esitetyn mukaisesti 496 havaintoa, voisi päätellä, että seitsemän vastaajan tiedot ovat pudonneet pois. Laskelma ei täsmää, sillä tulosteen 3.7 (s. 66) mukaan molemmista muuttujista puuttuu neljä havaintoa.

Ero selittyy tilastollisten ohjelmistojen oletusarvoisesti soveltamalla toimintaperiaatteella, josta tutkijan on syytä olla tietoinen. *Ohjelmistot jättävät havainnon kokonaan pois*, jos sen kohdalta puuttuu tieto yhdestäkin analyysiin valitusta muuttujasta. Toisin sanottuna ohjelmistoille kelpaavat vain täydelliset havainnot.

Esimerkiksi kahden muuttujan tapauksessa havainto putoaa pois niin kuvista, taulukoista kuin muistakin analyyseista, jos tieto puuttuu jommastakummasta tai molemmista muuttujista. Tulosteen 3.9 tapauksessa käy juuri näin: yksi vastaajista on jättänyt vastaamatta molempiin väittämiin ja loput jompaankumpaan. Tämä selittää edellä havaitut erot havaintojen määrissä.

Moiset erot saattavat vaikuttaa joutavilta, mutta ne ovat kaikkea muuta. Kahden muuttujan taulukko antaa vasta pientä esimakua siitä, mitä on odotettavissa, kun siirrytään usean muuttujan tarkasteluihin. Tulkinnat voivat jo kahden muuttujan osalta käydä hankaliksi, jos taulukoiden havaintomäärät ovat mitä milloinkin. Puuttuvat tiedot ovat jokaisen kyselytutkimuksen rasite. Niille on syytä yrittää tehdä jotain aineiston muokkauksen yhteydessä (ks. kohta 3.5, s. 81).

Prosenttitaulukot

Tulosteessa 3.10 on vastaajien viisiluokkainen ikäjakauma taulukoituna aineiston keruuvuoden mukaan. Tulosteeseen sisältyy kolme taulukkoa, joista ensimmäinen näyttää lukumäärät ja kaksi muuta kertoo lukumääriä vastaavat prosentuaaliset osuudet, ensin sarakkeittain ja sitten riveittäin. Viimeksi mainitut tunnistaa ilman tarkempia otsikoitakin lukujen summista.

Taulukoista nähdään muun muassa, että ikäryhmien osuudet ovat eri vuosien otoksissa melko samat. Selvimmät erot otosten välillä ilmenevät ryhmissä 30–39 ja 60–74 vuotta. Koska kyseessä ovat aidot satunnaisotokset 18–74-vuotiaista naisista, voidaan vertailua tehdä myös perusjoukon vastaaviin tietoihin. Tällöin saadaan käsitys siitä, miten hyvin otoksen vastaajat edustavat eri ikäryhmiä ja mikä on vastauskadon vaikutus.

Tuloste 3.10. Kolme taulukkoa iästä ja aineiston keruuvuodesta.

```
       ika   18-29   30-39   40-49   50-59   60-74     sum
Vuosi  ***
 1997           63      65      62      46      37     273
 2005           54      36      52      37      42     221
       sum     117     101     114      83      79     494


       ika   18-29   30-39   40-49   50-59   60-74     sum
Vuosi  ***
 1997         53.8    64.4    54.4    55.4    46.8    55.0
 2005         46.2    35.6    45.6    44.6    53.2    45.0
       sum   100.0   100.0   100.0   100.0   100.0   100.0


       ika   18-29   30-39   40-49   50-59   60-74     sum
Vuosi  ***
 1997         23.1    23.8    22.7    16.8    13.6   100.0
 2005         24.4    16.3    23.5    16.7    19.0   100.0
       sum    23.7    20.4    23.1    16.8    16.0   100.0
```

Prosenttitaulukoiden yhteydessä on syytä jollain tavoin tuoda esiin havaintojen lukumäärä, ainakin kokonaismäärä. Yllättäen se tässäkin poikkeaa hieman 496:sta. Tarkemmin ero tulee esiin seuraavaksi piirrettävästä laatikkokuvasta. Kaikkiin aineistoihin kätkeytyy virheitä ja epäjohdonmukaisuuksia. Tutustumisen ja esikäsittelyn yhtenä tavoitteena on löytää, selvittää ja korjata ne. Varsinaiset analyysit pirstaloituvat pilalle, jos roskia nousee myöhemmin jatkuvasti pintaan.

Taulukoita voidaan laatia myös useamman kuin kahden muuttujan suhteen, mutta tämä tie on melko rajallinen, koska luokkien yhdistelmien lukumäärä kasvaa hyvin äkkiä liian suureksi. Parhaimmillaan taulukot ovat, kun muuttujia on kaksi tai korkeintaan kolme.

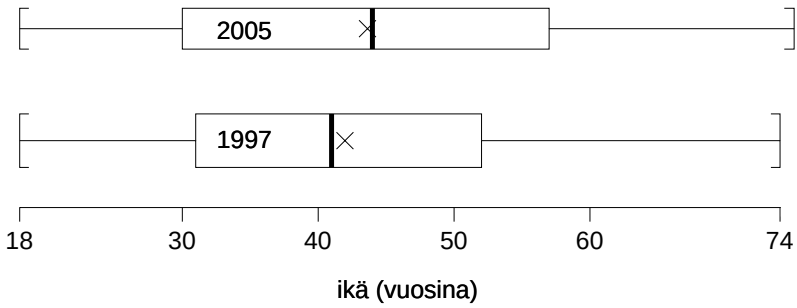
Taulukot ovat hyödyllisiä, koska ne tuovat aineiston tietoja esiin varsin yksityiskohtaisesti. Ne eivät kuitenkaan anna kunnollista yleiskäsitystä muuttujien välisistä suhteista. Etenkin jatkuvien muuttujien kuvailuun taulukko on kömpelö, koska se vaatii tietojen luokittelua. Erilaiset luokitukset johtavat erilaisiin tulkintoihin.

3.4.2 Kuvat

Parempia keinoja kahden muuttujan yhtaikaiseen tarkasteluun avautuu tilastollisista kuvista, joista seuraavassa tutustutaan laatikkokuvaan ja hajontakuvaan.

Laatikkokuva

Järjestystunnuslukujen yhteydessä esitetty laatikkokuva yleistyy välittömästi kahden muuttujan kuvaamiseen, joista toinen on jatkuva ja toinen diskreetti. Samaan kuvaan voidaan helposti piirtää jatkuvan muuttujan jakaumaa symboloivat laatikot ja viivat kullekin diskreetin muuttujan luokalle. Juuri näin on tehty vastaajien ikäjakaumaa aineiston molempina keruuvuosina esittävässä kuvassa 3.5.



Kuva 3.5. Iän laatikkokuva vuosina 1997 ja 2005.

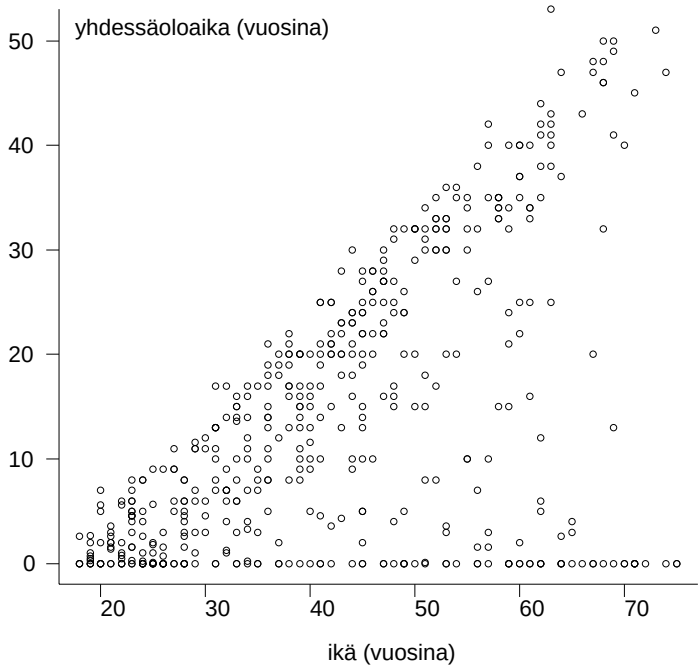
Tarkasti katsottuna kuvan 3.5 oikea reuna paljastaa jo edellä havaitun pienen epäjohdonmukaisuuden: vuoden 2005 aineistossa piileskelee 75-vuotiaita vastaajia, vaikka otanta-asetelma rajaa aineiston 18–74-vuotiaisiin. Kyseessä voi olla tallennusvirhe tai muu epäselvyys, joka olisi hyvä tarkistaa.

Kuvassa 3.5 rasteina näkyvistä keskiarvoista voidaan päätellä, että vuoden 1997 aineiston ikäjakauma on hieman vinompi, koska keskiarvo on mediaania suurempi. Erot ovat pieniä. Keskiarvo- tai mediaanierot keruuajankohtien välillä näkyvät sen sijaan selvemmin. Koska alempi laatikko on lyhyempi, on vuoden 1997 ikäjakauma keskittynyt vähän tiiviimmin mediaaninsa ympärille. Laatikoiden paksuudet ovat puolestaan verrannollisia vastaajamääriin. Vuodelta 2005 on vähemmän vastauksia ja ne ovat hieman enemmän hajallaan.

Hajontakuva

Tilastollisista kuvista tärkeimpiä on *hajontakuva*. Se esittää kahden muuttujan yhteisjakaumaa niin, että kutakin havaintoa vastaa yksi piste muuttujien arvojen rajaamassa koordinaatistossa. Nimi tulee siitä, että kuvaan muodostuva pisteparvi esittää molempien muuttujien hajonnan koko aineiston laajuudelta. Samalla hajontakuva paljastaa muuttujien mahdollisen *riippuvuuden* ja sen luonteen sekä yksityiskohtia, kuten poikkeavia arvoja. Hajontakuva on keskeinen myös sen vuoksi, että se on läheisessä yhteydessä riippuvuutta kuvaaviin korrelaatioon ja regressioanalyysiin, joita käsitellään myöhemmin tässä ja seuraavissa luvuissa. Hajontakuvasta on lukuisia muunnelmia, joita tullaan näkemään eri menetelmien yhteydessä.

Kuvaan 3.6 on piirretty vastakkain kaksi aiemmin tarkasteltua muuttujaa: vastaajan ikä ja yhdessäoloaika nykyisen kumppanin kanssa. Molempien yksikkönä ovat vuodet, mutta yhdessäoloaika on mitattu hieman tarkemmin. Molempia voidaan hajontakuvaa piirrettäessä ajatella jatkuvina, koska ne saavat paljon eri arvoja vaihteluväleillään. Periaatteessa on samantekevää, kummin päin muuttujia tarkastellaan, mutta koska olisi luontevampaa ajatella iän selittävän yhdessäoloaikaa eikä toisinpäin, piirretään ikä vaakasuuntaan. Vakiintuneen käytännön mukaisesti *selittävä muuttuja* piirretään hajontakuvan vaaka-akselille ja *selitettävä muuttuja* pystyakselille. Näihin käsitteisiin palataan tarkemmin regressioanalyysin yhteydessä, luvussa 5.



Kuva 3.6. Iän ja yhdessäolon hajontakuva.

Kuvasta 3.6 on hahmottuvinaan selvä suoraviivainen yhteys: mitä vanhempi, sitä kauemmin on ollut yhdessä. Kulmasta kulmaan voisi nähdä etenevän linjan, jonka perusteella voisi myös päätellä, että parisuhde solmitaan tyypillisesti noin 20-vuotiaana. Hajontaakin on runsaasti; nämä ovat vain yleisvaikutelmia, jotka hajontakuva kertoo nopeasti. Ylipäätään hyvä tilastollinen kuva kertoo jo yhdellä vilkaisulla jotain, ja sen tarkempi tutkiskelu täydentää kokonaiskuvaa yksityiskohdilla.

Kuvan alareunassa näkyvät ne 99, jotka eivät olleet vastaushetkellä vakituisessa parisuhteessa (vrt. tuloste 3.3, s. 58). Kuva kertoo selvästi, kuinka nämä 99 asettuvat iän koko vaihteluvälille kuten aineiston muutkin vastaajat. Vakituisen parisuhteen olemassaolo ei siis ainakaan ole iästä kiinni.

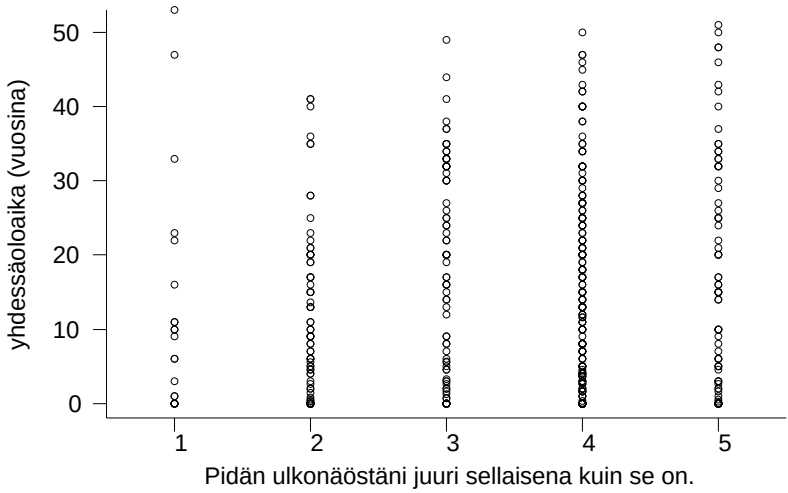
Vaikka tässä on ajateltu yhdessäoloaikaa jatkuvana muuttujana, on muistettava sen aiemmin todettu kaksijakoisuus: muuttuja sisältää myös laadullisen tiedon siitä, onko vastaaja vakituisessa parisuhteessa vai ei. Tähän liittyvät juuri mainitut 99 vastaajaa. Hajontakuva tarjoaa hyvän ikkunan aineistoon muuttujien koko laajuudelta, mutta syvällisempiä analyyseja varten olisi parempi tarkentaa näkymiä. Esimerkiksi jos tehtäisiin regressioanalyysi koko aineistolla, kyseisistä 99:stä tulisi niin sanotusti *vaikutusvaltaisia* havaintoja, sillä ne vetäisivät suoraviivaista yhteyttä oikeasti kuvaavan *regressiosuoran* huomattavasti silmämääräisesti kuviteltua alemmas. Näihin käsitteisiin palataan luvussa 5.

Jos vielä katsastetaan hetki kuvaa 3.6, tulee esiin hajontakuvan ylivoimaisuus aineiston poikkeavuuksien hahmottamisessa. Läheltä kuvan oikeaa yläkulmaa paljastuu epäilyttävältä vaikuttava havainto: vähän yli 60-vuotias vastaaja on ilmoittanut yhdessäoloajaksi nykyisen kumppaninsa kanssa 53 vuotta. Se on tottakai melkoinen saavutus, mutta ehkä asia olisi ainakin hyvä tarkistaa kyselylomakkeesta.

Mitä jos toinen muuttuja on diskreetti?

Hajontakuva on perusmuodossaan kahden jatkuvan muuttujan kuvaaja, onhan hajonnastakin parempi puhua vasta jatkuvien muuttujien osalta. Käytännössä rajanveto jatkuvan ja diskreetin välillä on kuitenkin monesti tulkintaa. Kuten seuraavassa ja myöhemminkin tullaan näkemään, hajontakuva joustaa perusmuodostaan varsin pitkälle. Esimerkiksi aikasarjojen kuvaajat ovat hajontakuvan sovelluksia: eri ajankohtia kuvaavia pisteitä yhdistetään toisiinsa viivoilla. Tässä kirjassa ei aikasarjakuvia tarvita, mutta luvussa 7 samaa tekniikkaa sovelletaan diskreettien muuttujien luokkien yhdistämiseen. Seuraavassa esiintyy myös diskreettejä muuttujia, mutta tässä vaiheessa niitä käytetään vielä jatkuvien muuttujien tapaan.

Kuva 3.7 on myös hajontakuva, mutta kun toinen muuttujista on diskreetti, kuvan ilmiasu on aika erilainen kuin edellä. Diskreetin muuttujan luokkien kohdille muodostuu suoria pisteviivoja jatkuvan muuttujan arvoista. Tällaista kuvaa kutsutaan toisinaan *pistekuvaksi* (*dot-plot*). Kuvan vaaka-akselilla on toinen tulosteessa 3.7 (s. 66) esiintyneistä muuttujista.



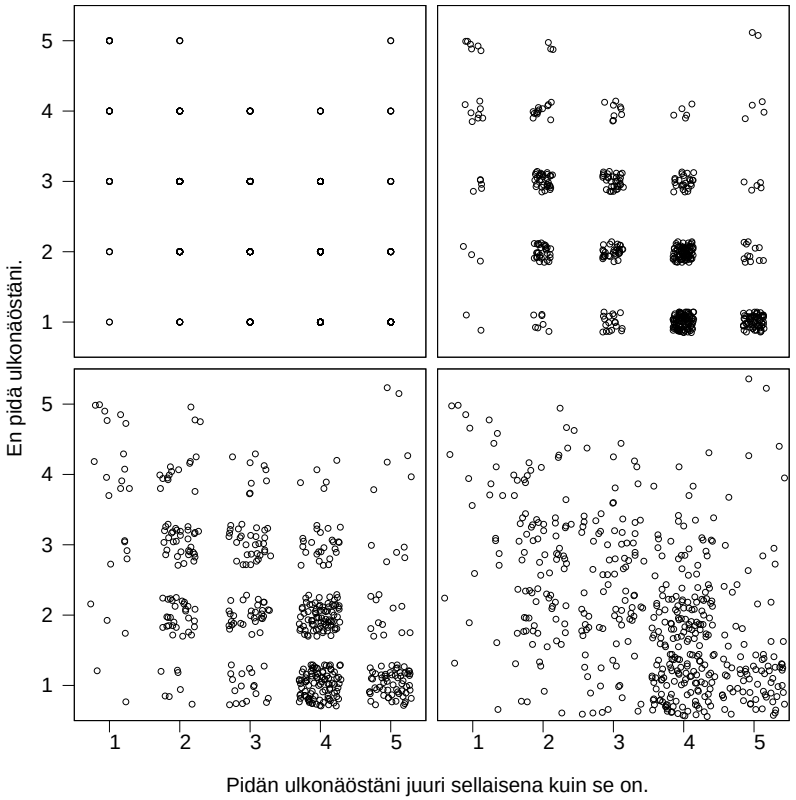
Kuva 3.7. Asennemuuttujan ja yhdessäolon hajontakuva.

Aineistoon tutustuessa kaikki mahdolliset kuvat ovat hyödyksi, mutta tulosten esittämiseen kuva 3.7 saattaisi olla liian yksityiskohtainen. Esitystä voisi tiivistää siirtymällä laatikkokuvaan (vrt. kuva 3.5, s. 71), jolloin jokainen diskreetin muuttujan arvo kuvautuisi vain laatikkona ja viivoina. Laatikkokuvakin on siis hajontakuvan erikoistapaus.

Entä jos molemmat ovat diskreettejä?

Kun molemmat muuttujat ovat diskreettejä, ei laatikkokuvasta ole iloa, muttei hajontakuvastakaan – ainakaan sellaisenaan. Hajontakuvien mahdollisuuksista löytyy kuitenkin keinoja näidenkin tilanteiden kuvaamiseen. Kuva 3.8 sisältää neljä hajontakuvaa, jotka kaikki ilmentävät kahden asennemuuttujan keskinäistä riippuvuutta. Muuttujat ovat samat kuin tulosteen 3.9 (s. 68) ristiintaulukossa.

Alkutilanne on kuvan 3.8 vasemman yläkulman hajontakuva, josta ei ole mitään hyötyä. Siitä näkyy vain, että lähes kaikkia arvojen 1–5 yhdistelmiä löytyy aineistosta; ainoastaan kaksi loistaa poissaolollaan.



Kuva 3.8. Sarja hajontakuvia kahdesta asennemuuttujasta.

Aineiston noin 500 havaintoa sijoittuu siis vain 23 kohtaan, joten päällekkäin piirtyy valtavasti pisteitä; kuva ei vain kerro montako mihinkin kohtaan.

Mielenkiintoisen informaation esiin houkuttelemiseksi muissa sarjakuvan 3.8 ruuduissa käytetään jännittävällä tavalla hyväksi tilastotiedettä. Pisteitä *täristetään* irralleen, jolloin aletaan nähdä kaikkea, mikä alkutilanteessa jäi pimentoon. Täristys tarkoittaa, että jokaiselle havainnolle *arvotaan* satunnaisesti uusi paikka sen oikean paikan läheisyydestä. Tuloksena on kuva, jossa käytännössä yksikään piste ei ole oikealla kohdallaan, mutta juuri sen ansiosta kuva kertoo olennaisesti enemmän kuin se, jossa kaikki ovat oikeilla paikoillaan.

Täristystä kuvan 3.8 ruuduissa lisätään vaiheittain, saman verran molempien muuttujien suunnassa, ja vähitellen muuttujien diskreetti luonne alkaa hämärtyä. Oikean alakulman kuva muistuttaa jo huomattavasti kahden jatkuvan muuttujan hajontakuvaa. Tästä on se hyöty, että muuttujien välinen riippuvuus tulee paremmin näkyviin. Melko hyvin se hahmottuu jo vasemman alakulman kuvasta: pisteet keskittyvät niihin yhdistelmiin, joissa vastaaja on samaa mieltä positiivisesta ja eri mieltä negatiivisesta väitteestä. Myös hieman poikkeukselliset vastaukset ja niiden määrät erottuvat hyvin. Kuvaa kannattaa verrata muuttujien ristiintaulukkoon tulosteessa 3.9 (s. 68).

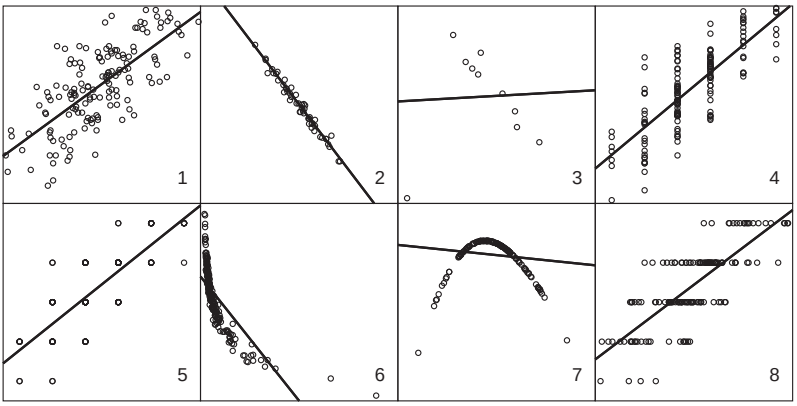
3.4.3 Tunnusluvut

Hajontakuvan yhteydessä viitattiin jo muuttujien yhteyksien ja riippuvuuksien tarkasteluun. Sinänsä muuttujat ovat näissä tarkasteluissa samanarvoisia. Syiden ja seurausten analysointi edellyttää sisällöllisiä tulkintoja. Tilastolliset kahden muuttujan tunnusluvut kertovat korkeintaan muuttujien välisten yhteyksien voimakkuudesta.

Korrelaatio

Tärkein kahden muuttujan yhteyden tai riippuvuuden voimakkuutta kuvaava tunnusluku on niiden *korrelaatiokerroin* tai lyhyemmin korrelaatio. Keskiarvon ja keskihajonnan tavoin se edellyttää numeerista mittaustasoa. Arkikielessä puhutaan toisinaan, kuinka jotkut asiat ”korreloivat”, mutta tilastotieteen kielessä korrelaatio tarkoittaa vain *lineaarista* eli suoraviivaista riippuvuutta. Riippuvuuden luonnetta on arvioitava hajontakuvista; pelkkä korrelaatio ei siitä kerro mitään.

Kuva 3.9 näyttää kahdeksan simuloiduista aineistoista piirrettyä hajontakuvaa. Vain osa riippuvuuksista vaikuttaa lineaarisilta, selvimmin ruudut 1 ja 2. Ruudussa 3 riippuvuus on muuten lineaarista, mutta vasemman alakulman erillinen piste ei tunnu kuuluvan joukkoon. Kuviin on piirretty *regressiosuorat*, jotka korrelaation tavoin ilmentävät ainoastaan lineaarista riippuvuutta. Ruudussa 3 suora asettuu erillisen pisteen takia aivan toisin kuin äkkipäätä ajattelisi. Ruuduissa 4, 5 ja 8 riippuvuus näyttää lineaariselta, mutta etenkin ruudussa 5 muuttujien diskreettiys saattaa piilottaa alle myös jotain muuta. Ruudut 6 ja 7 edustavat *epälineaarisia* riippuvuuksia, joita regressiosuora ja korrelaatio eivät kuvaa lainkaan.



Kuva 3.9. Simuloituja hajontakuvia regressiosuorineen.

Korrelaatio ilmentää lineaarista riippuvuutta positiivisena tai negatiivisena, itseisarvoltaan korkeintaan ykkösen suuruisena lukuna. Nolla tarkoittaa, ettei korrelaatiota ole lainkaan. Miten lähellä nollaa on käytännössä ”ei lainkaan”, on tulkintakysymys. Positiivinen korrelaatio vertautuu nousevaan ja negatiivinen laskevaan regressiosuoraan, kuten kuvan 3.9 ruuduissa 1 ja 2, joissa vastaavat korrelaatiot ovat noin 0.8 ja –0.99. Kun korrelaatio lähestyy itseisarvoltaan ykköstä, muuttujien hajontakuvan pisteet asettuvat yhä lähemmäksi regressiosuoraa.

Korrelaation ja hajontakuvan läheinen suhde on syytä pitää mielessä. Ensin on piirrettävä hajontakuvia. Korrelaatioita kannattaa tutkia vasta sen jälkeen, jos se kuvan perusteella vaikuttaa järkevältä. Kuvan 3.9 ruutu 3 on hyvä muistaa: korrelaatio olisi –0.95, mutta alakulman erillisen pisteen takia se onkin 0.04, siis käytännössä nolla. Jos korrelaatioita tarkastellaan piirtämättä kuvia, tällaiset tilanteet jäävät piiloon, jolloin sekä analyysit että johtopäätökset menevät pieleen.

Regressiosuora on esimerkki yksinkertaisesta *tilastollisesta mallista*. Regressioanalyysin yhteydessä luvussa 5 tutustutaan keinoihin, joilla poikkeavia havaintoja pyritään haravoimaan esiin aineistosta vielä mallin rakentamisen yhteydessä. Aina parempi on, jos niitä havaitaan jo aineistoon tutustuessa.

Korrelaatiomatriisi

Kun edetään aineistoon tutustumisesta kohti aineiston tiivistämistä, korrelaatioiden merkitys kasvaa. Kunhan edellä pohditut edellytykset täyttyvät, niin kahden muuttujan välinen riippuvuus tiivistyy korrelaationa yhdeksi ainoaksi luvuksi. Koska sen muodostamiseen käytetään kaikkia tarkasteltavana olevia havaintoja, niin aineisto tiivistyy heti huomattavasti. Tilastotieteessä puhutaan kuvaavasti *tyhjentävistä tunnusluvuista*, joilla monesti viitataan juuri korrelaatioihin, mutta myös keskiarvoihin ja keskihajontoihin.

Useat myöhemmissä luvuissa käsiteltävät *monimuuttujamenetelmät*, kuten faktorianalyysi ja erotteluanalyysi, tiivistävät aineistoa edelleen. Niiden lähtökohtana on muuttujien välisistä korrelaatioista koostuva *korrelaatiomatriisi*, joka ikään kuin edustaa alkuperäistä havaintomatriisia, mutta on kooltaan huomattavasti pienempi.

Tuloste 3.11. Yhdeksän asennemuuttujan korrelaatiomatriisi.

	k26.1	k26.2	k26.3	k30.1	k30.2	k30.3	k71.1	k71.2	k71.3
k26.1	1.00	**0.59**	**0.76**	0.10	0.04	-0.03	-0.10	-0.11	-0.08
k26.2	**0.59**	1.00	**0.61**	0.11	0.14	0.12	-0.13	-0.07	-0.07
k26.3	**0.76**	**0.61**	1.00	0.10	0.06	-0.03	-0.05	-0.12	-0.14
k30.1	0.10	0.11	0.10	1.00	**0.47**	**0.28**	0.02	0.13	0.06
k30.2	0.04	0.14	0.06	**0.47**	1.00	**0.42**	0.02	0.20	**0.25**
k30.3	-0.03	0.12	-0.03	**0.28**	**0.42**	1.00	-0.08	0.13	**0.24**
k71.1	-0.10	-0.13	-0.05	0.02	0.02	-0.08	1.00	0.18	0.13
k71.2	-0.11	-0.07	-0.12	0.13	0.20	0.13	0.18	1.00	**0.41**
k71.3	-0.08	-0.07	-0.14	0.06	**0.25**	**0.24**	0.13	**0.41**	1.00

Tulosteessa 3.11 on esitetty yhdeksän asennemuuttujan korrelaatiomatriisi. Muuttujat, ensimmäiset osiot kolmesta eri mittarista, tulevat sisällöiltään tutummiksi seuraavassa luvussa. Tässä vaiheessa katsotaan vain pinnallisesti, mitä korrelaatiomatriisista nähdään. Hahmottamisen helpottamiseksi voimakkaimpia korrelaatioita on korostettu niiden *tilastollisen merkitsevyyden* perusteella. Käsitteeseen perehdytään seuraavassa luvussa, mutta jo nyt voidaan nähdä, ettei tilastollinen merkitsevyys yksin riitä johtopäätösten tekemiseen. Korostetuista korrelaatioista muutamat ovat selvästi suurempia kuin toiset. Mitä luultavimmin vain osa niistä on *sisällöllisesti merkittäviä*.

Korrelaatiomatriisi on symmetrinen ykkösistä muodostuvan lävistäjänsä suhteen. Nuo ykköset voi tulkita muuttujan korrelaatioksi itsensä kanssa, mutta oikeastaan ne viittaavat muuttujien omaa vaihtelua kuvaaviin *variansseihin*. Korrelaatioita laskettaessa muuttujista tehdään vertailukelpoisia muuntamalla erisuuruiset varianssit ykkösiksi. Vastaavasti muuttujien yhteisvaihtelua ilmentävät *kovarianssit* muuttuvat korrelaatioiksi. Käytännössä toimitaan varianssien ja kovarianssien sijaan keskihajonnoilla ja korrelaatioilla.

Kaikkiaan tulosteen 3.11 korrelaatiomatriisissa on 81 lukua, joista korkeintaan 36 kiinnostaa, sillä symmetrinen puolikas ja sinänsä turha lävistäjä voidaan unohtaa. Lopulta kiinnostavia lukuja on vain kymmenkunta. Niistä voi päätellä, että mittareissa on joitakin paljolti samoja asioita mittaavia muuttujia ja että osa muuttujista ei juuri korreloi muiden kanssa. Johtopäätökset eivät ole yllättäviä.

Yksittäisten korrelaatioiden perusteella ei ole edes syytä tehdä pidemmälle meneviä tulkintoja tai johtopäätöksiä. Kun muuttujia on enemmän, korrelaatiomatriisin tutkiskelu muuttuu myös varsin epäkäytännölliseksi. Riippuvuuksien tarkempi analysointi ja aineiston todellinen tiivistäminen edellyttävät tilastollista mallintamista, johon perehdytään luvuissa 4 ja 5.

Luokittelu- tai järjestystason riippuvuudet

Luokittelu- tai järjestystason mittausten yhteyksiä on parasta tarkastella ristiintaulukoimalla (ks. kohta 3.4.1, s. 68), sillä korrelaatio ei niihin sovellu. Taulukko on monipuolisempi kuin pelkkä korrelaatiokerroin, koska se ei rajoitu lineaarisiin riippuvuuksiin.

Taulukon hankaluutena korrelaatioon verrattuna on tilankäyttö. Kun kahden numeerisesti mitatun muuttujan lineaarinen riippuvuus tiivistyy parhaimmillaan yhdeksi luvuksi, vaatii kahden luokittelumuuttujan tarkastelu kokonaisen, mahdollisesti kymmenistä luvuista koostuvan frekvenssitaulukon. Ei siis ihme, että taulukoidenkin välittämää informaatiota halutaan tiivistää tunnusluvuiksi.

Luvussa 7 perehdytään tilastolliseen testiin, jolla kahden muuttujan ristiintaulukon kuvaama riippuvuus tiivistetään yhdeksi tunnusluvuksi. Samassa luvussa nähdään myös, että korrelaatiomatriisia vastaa luokittelu- tai järjestystasolla tavallisen ristiintaulukon laajennus, jonka avulla usean muuttujan riippuvuuksia voidaan visualisoida korrespondenssianalyysi-nimisellä menetelmällä.

3.5 Muokkaukset

Aineiston muokkaus sisältää muun muassa edellä käsitellyt muunnokset ja havaittujen virheiden korjaukset. Erityisesti muokkauksilla tarkoitetaan tässä useita muuttujia koskevia muunnoksia, joilla aineistoa valmistellaan tiivistämistä ja analyyseja varten.

Puuttuvat tiedot

Tärkeimpiä muokkauksen kohteita ovat *puuttuvat tiedot*, joita esiintyy kaikissa kyselyaineistoissa. Niiden osalta, jotka täyttävät lomakkeen, mutta jättävät osan tiedoista pois, puhutaan *eräkadosta*. Ne jotka eivät edes täytä lomaketta, jäävät tietenkin kokonaan pois aineistosta. Tätä kuvastaa termi *yksikkökato*. Otanta-asetelmissa molempia kadon tyyppejä voi olla mahdollista paikata jälkikäteen rekisteritietojen avulla tai vastauksia painottamalla, mutta mikään paikkaustapa ei aidosti korvaa puuttuvaksi jäänyttä tietoa. Puuttuneisuuden syiden ja mekanismien tarkempi selvittäminen on haasteellista. Asiaan on hyvä yrittää vaikuttaa jo mittausvaiheessa. Mitä enemmän puuttuvia tietoja aineistossa on, sitä enemmän epävarmuuksia siihen sisältyy.

Tilastollisin keinoin tapahtuva tietojen paikkaus eli *imputointi* on tasapainoilua erilaisten epävarmuuksien kanssa. Usein jonkinasteinen paikkaus on perusteltua, koska muuten tietoja yksinkertaisesti menetetään liikaa. Toisaalta on selvää, ettei tutkimusta voida perustaa pelkkien paikkausten varaan, tällöinhän koko aineisto voitaisiin yhtä hyvin rakentaa simuloimalla! Imputointi vertautuu tavallaan kohdassa 3.4.2 (s. 71) käsiteltyyn hajontakuvan täristämiseen. Molemmissa hyödynnetään tilastotieteen omia keinoja aineiston sisältämän informaation esille saamiseksi.

Tilastollisia aineiston paikkausmenetelmiä käsittelevät tarkemmin muun muassa tiedonkeruuta koskevat uudemmat teokset kuten luvussa 2 mainitut Groves ym. (2004) ja Lehtonen & Pahkinen (2004). Hyvä suomenkielinen katsaus aiheeseen on Sovio & Läärä (2002). Perusteos näistä menetelmistä on *Statistical Analysis with Missing Data* (Little & Rubin, 1987).

Puuttuvuuden seuraukset

Kohdassa 3.4.1 (s. 68) nähtiin, millaisia seurauksia tietojen puuttumisesta voi olla jo kahden muuttujan kanssa. Kun siirrytään analysoimaan kymmeniä muuttujia yhtaikaa, kuten seuraavassa luvussa tehdään, voi puuttuvien tietojen takia käydä todella huonosti. Kun menetelmät kelpuuttavat vain täydelliset havainnot, voi pahimmillaan menettää kaikki havainnot, ennen kuin analyysi pääsee edes alkuun. Niin täydellisen hävikin havaitsisi varmasti. Ongelmia seuraa enemmän siitä, kun kato käy vähän pienemmässä mittakaavassa, jolloin analyysit saattavat mennä ”läpi” ja tulokset voivat vaikuttaa järkeviltä. Havaintojen karsinta tapahtuu korrelaatioita laskiessa, ja koska siihen teknisesti riittää vaivaiset *kaksi havaintoa*, voi seurata varsin omituisia yllätyksiä, ellei tunne aineistoaan.

Puuttuvuuden monet syyt

On muistettava, että puuttuvuutta on monenlaista, joten muokkauksen ja paikkauksen kanssa pitää olla huolellinen. Kaikkia tietoja ei ole välttämättä ollut tarkoituskaan kysyä kaikilta, joten aineistossa on myös sellaisia ”reikiä”, joita ei pidäkään paikata. Rekisteritietojen avulla saattaa olla mahdollista paikata jopa taustamuuttujia, mutta tyypillisesti paikkaus kohdistetaan mittareihin, joihin jää aina jonkin verran puuttuvia tietoja.

Koska imputointimenetelmät olettavat puuttuvuuden olevan satunnaista, on aineistoon tutustuessa koetettava erottaa mahdolliset systemaattiset puuttuvuudet satunnaiselta vaikuttavista. Esimerkiksi vastaaja on saattanut jättää systemaattisesti vastaamatta jonkin mittarin kaikkiin väittämiin. Tällaisia ”reikiä” ei kannata mennä täyttämään keinotekoisilla luvuilla. Jos sen sijaan reikiä on jonkin verran ”siellä täällä”, voi korvaaminen olla kokonaisuutta ajatellen järkevämpää kuin koko vastauksen pois jättäminen.

Vastausten eli havaintojen lisäksi aineistoa on tutkittava samaan tapaan muuttujien suhteen. Jos johonkin kysymykseen ei jostain syystä ole juurikaan vastattu, ei kyseistä muuttujaa ole perusteltua paikata vaan se joudutaan kenties jättämään pois.

En osaa sanoa

Yksi hankala asia lisää ovat eos- eli ”en osaa sanoa” -vaihtoehdot, joista puhuttiin jo mittauksen yhteydessä, kohdassa 2.3.3 (s. 34). Joihinkin analyyseihin eos-vastaukset voidaan sisällyttää omana luokkanaan, mutta usein ne samastetaan puuttuviin tietoihin, jolloin niitä koskee kaikki edellä todettu. Numeeriselle asteikolle eos-vastaukset eivät lähtökohtaisesti kuulu, koska ne saattavat mitata aivan eri asiaa kuin varsinainen muuttuja. Joissain tapauksissa voi olla mahdollista pitää niitä ikään kuin neutraaleina vaihtoehtoina, mutta asiaa on tutkittava tapauskohtaisesti esimerkiksi taulukoiden avulla. Automaattisesti eos-vastaus ei ole neutraali vaihtoehto.

Ulkonäkötutkimuksen mittareiden paikkaus

Luvussa 2 alettiin pohtia ulkonäkötutkimuksen mittausmallia, jossa ulkonäköön liittyviä ulottuvuuksia mitataan kolmella mittarilla. Seuraavassa luvussa mittausmalliin ja sen analysointiin pureudutaan tarkemmin, mutta sitä ennen tarkastellaan puuttuvien tietojen paikkausta kyseisissä mittareissa. Todellisuudessa nämäkin tarkastelut olisivat laajempia, sillä tässähän on koko ajan rajoituttu keinotekoisesti vain kolmeen mittariin yrittämättäkään selostaa analysointia koko aineiston laajuudelta.

Ensimmäiseen mittariin sisältyy 22, toiseen 20 ja kolmanteen 11 osiota, siis yhteensä tullaan analysoimaan 53 muuttujaa yhtaikaa. Puuttuvia tietoja tarkastellaan saman tien molempien vuosien osalta, jolloin havaintoja on kokonaisuudessaan jo edellä tutuksi tullut määrä 496. Jos korrelaatiot laskettaisiin välittämättä puuttuvista tiedoista, täydellisiä havaintoja jäisi vain 418. Toisin sanoen 78 vastaajan, joista 40 vuodelta 1997 ja 38 vuodelta 2005, tiedot putoaisivat tässä vaiheessa kokonaan pois. Tämä ei tunnu alkuunkaan järkevältä, joten aineistoa pyritään paikkaamaan. Tilannetta on kuitenkin syytä sitä ennen katsastaa paljon tarkemmin.

Tässä yhteydessä on kuitenkin tyydyttävä tilan säästämiseksi hieman lyhyempään tarkasteluun. Korvaamiseen käytetään yksinkertaista mittarikohtaista päätössääntöä: jos vastaaja on vastannut yli puoleen mittarin väitteistä, niin loput mahdolliset puuttuviksi jääneet korvataan. Jos vastauksia on puolet tai alle, ei korvata vaan havainto jätetään kokonaan pois. Tämäntyyppisiä sääntöjä sovelletaan usein myös käytännössä. Toisinaan peruste voi olla lievempi, esimerkiksi 25 %. Mitään tilastollista perustetta tähän ei ole; tutkijan pitää osata arvioida, mikä kulloinkin on järkevää. Sääntö voi hyvin vaihdella mittareittainkin. Minkään sääntöjen sokea noudattaminen ei ole järkevää, vaan tilannetta pitää katsastaa monipuolisemmin kuin mitä tässä on mahdollista esittää.

Tulosteessa 3.12 on yhteenveto mittareittain. Edellä mainittu päätössääntö tarkoittaa tässä, että ensimmäisessä mittarissa pitää olla vähintään 12, toisessa 11 ja kolmannessa kuusi vastausta, jotta paikkaukseen turvaudutaan. Suurin osa havainnoista on täydellisiä, joten paikkaus on todella paikkausta eikä aineiston keksimistä. Ensimmäisessä mittarissa ei näytä olevan yhtään poistettavia havaintoja, mutta toisaalta siinä on eniten paikattavia.

Tuloste 3.12. Puuttuvien tietojen tarkastelu mittareittain.

```
             Vuosi  1997  2005   sum
Mittari1     *****
poistetaan             0     0     0
paikataan             23    20    43
jätetään             250   203   453
               sum   273   223   496


             Vuosi  1997  2005   sum
Mittari2     *****
poistetaan             3     1     4
paikataan             11    17    28
jätetään             259   205   464
               sum   273   223   496


             Vuosi  1997  2005   sum
Mittari3     *****
poistetaan             3     3     6
paikataan              7     3    10
jätetään             263   217   480
               sum   273   223   496
```

s. 207

Mittarikohtainen tarkastelu ei paljasta koko tilannetta. Sitä varten tiedot on taulukoitava sisäkkäin yhteen taulukkoon. Näin on tehty tulosteessa 3.13. Taulukosta on poistettu ensimmäiset yhdeksän riviä, koska ne ovat ensimmäisen mittarin ansiosta pelkkiä nollia, kuten edellä mainittiin. Tähdillä (*) on merkitty rivit, joissa ainakin yhden mittarin osalta ollaan ”poistoluokassa”. Nämä havainnot, joita on viisi vuodelta 1997 ja neljä vuodelta 2005, jäävät päätössäännön mukaisesti paikkauksen ulkopuolelle ja siten pois analyyseista. Tällaisen hävikin voi jotenkuten sietää. Tulosteesta voi päätellä, että paikattavia havaintoja on lopulta 69, joista 35 vuoden 1997 ja 34 vuoden 2005 aineistosta.

Tuloste 3.13. Puuttuvien tietojen tarkempi tarkastelu.

Mittari1	Mittari2	Mittari3	Vuosi *****	1997	2005	sum	
paikataan	poistetaan	poistetaan		1	0	1	*
		paikataan		0	0	0	
		jätetään		0	1	1	*
	paikataan	poistetaan		0	0	0	
		paikataan		2	0	2	
		jätetään		0	3	3	
	jätetään	poistetaan		0	1	1	*
		paikataan		0	1	1	
		jätetään		20	14	34	
jätetään	poistetaan	poistetaan		0	0	0	
		paikataan		0	0	0	
		jätetään		2	0	2	*
	paikataan	poistetaan		0	0	0	
		paikataan		1	0	1	
		jätetään		8	14	22	
	jätetään	poistetaan		2	2	4	*
		paikataan		4	2	6	
		jätetään		233	185	418	
			sum	273	223	496	

Millä puuttuva tieto paikataan?

Edellä saatiin kuva siitä, mitä paikattaisiin. Seuraava askel olisi päättää, miten paikkaus tapahtuisi. Yksinkertaisin tapa olisi korvata puuttuvat arvot esimerkiksi kunkin muuttujan keskiarvolla. Huono puoli siinä on se, ettei se ota millään lailla huomioon muita muuttujia. Toinen huono puoli on se, että keskiarvolla paikkaus pienentää muuttujan hajontaa, jos puuttuvia tietoja on vähänkään enemmän. Kehittyneemmät paikkausmenetelmät käyttävät apunaan muiden muuttujien arvoja ja muodostavat puuttuvan tiedon tilalle arvion olemassa olevien tietojen avulla. Näin paikatut arvot ovat todennäköisesti hieman oikeampia kuin pelkät yhden muuttujan perusteella lasketut keskiarvot.

Jos puuttuvia tietoja on todella vähän, voi keskiarvokorvauskin olla käytännössä ihan toimiva menetelmä. Tässä on kuitenkin sovellettu hieman kehittyneempää, niin sanottua *regressioimputointia*, jossa otetaan huomioon kaikkien kolmen mittarin tiedot samanaikaisesti. Nimi viittaa luvussa 5 käsiteltävään regressioanalyysiin, sillä puuttuvia tietoja yritetään tavallaan vuorollaan ”selittää” muilla muuttujilla, ja tätä jatketaan kunnes havaittavia eroja ei enää ilmene. Tällaiset menetelmät ovat itse asiassa niin tehokkaita, että aineistoon on lopuksi syytä lisätä hieman keinotekoista vaihtelua, jotteivät tulokset olisi ”liian hyviä”. Hyvä tapa on tehdä analyyseja sekä paikatulla että alkuperäisellä aineistolla ja vertailla tuloksia.

Tämä kaikki vaivannäkö on tietenkin vain kalpeaa yritystä jäljittää puuttuvien tietojen jättämiä aukkoja, mutta se on usein parasta, mitä voidaan käytännössä tehdä. Tässä kolmen mittarin tapauksessa aineiston ”reikiä” tulee paikatuksi yhteensä 144 kappaletta, joka on varsin vaatimaton määrä verrattuna 53 muuttujasta ja 487 havainnosta muodostuvaan lukujen kokonaismäärään, joka on 25 811. Paikattuja arvoja on siis noin puoli prosenttia. Niiden ansiosta saadaan käyttöön kymmeniä havaintoja, jotka muuten tulisivat hylätyiksi.

Aineiston esikäsittely on työläs vaihe, joka saattaa joissain tapauksissa viedä enemmän aikaa kuin varsinainen analysointi. Ajankäyttö kuitenkin kannattaa, sillä perusteellisen esikäsittelyn, muunnosten ja muokkausten myötä lähtökohdat aineiston tiivistämiseen ja tilastolliseen mallintamiseen ovat huomattavasti paremmat.

4 Aineiston tiivistäminen

Tulosten esittämiseen voi toisinaan riittää hyvä taulukko tai kuva, mutta sellaisen aikaansaaminen edellyttää yleensä aineiston tiivistämistä ja etenemistä perustarkasteluista pidemmälle. Näissä tehtävissä tarvitaan tilastollisia malleja ja menetelmiä.

Tässä luvussa kuvataan yleisiä mallintamisen tavoitteita ja käsitteitä, tarkennetaan kirjan alkupuolella hahmoteltua mittausmallia ja perehdytään sen avulla ulkonäkötutkimuksen ulottuvuuksiin. Mallintamisen pohjalta aineistoa tiivistetään lopuksi merkittävästi jatkoanalyyseja varten.

4.1 Tilastollinen malli

Tilastollisen mallin tavoitteena on ilmaista tiiviissä muodossa jotain kiinnostavaa tutkittavasta ilmiöstä. Mallin rakentamisessa eli *mallintamisessa* tarvitaan sekä ilmiön tuntemusta että tilastollisten menetelmien osaamista.

Tilastollisen päättelyn perusteet

Yleensä ajatuksena on, että aineisto muodostaa satunnaisotoksen jostakin hyvin määritellystä perusjoukosta (ks. kohta 2.5.1, s. 43). Otoksen avulla pyritään tekemään arvioita perusjoukkoa koskevista ominaisuuksista kuten odotusarvoista, hajonnoista, korrelaatioista tai todennäköisyyksistä. Tällaisia ominaisuuksia kutsutaan *parametreiksi* ja niiden arviointia *estimoinniksi*. Esimerkiksi perusjoukon tuntematonta odotusarvoa voidaan estimoida otoksesta lasketulla keskiarvolla.

Muita estimoitavia parametreja ovat muun muassa faktorilataukset ja regressiokertoimet, joihin perehdytään myöhemmin. Johtopäätösten tekemistä tällaisten arvioiden tai estimaattien perusteella kutsutaan *tilastolliseksi päättelyksi*.

Yksi päättelyn osa-alueista on *tilastollinen merkitsevyystestaus*. Siinä on ideana testata perusjoukkoa koskevia oletuksia eli *hypoteeseja* aineistoa vasten. Testaus muistuttaa logiikaltaan rikostutkintaa: aineisto edustaa ”todisteita”, joiden pohjalta tehdään johtopäätöksiä, ja oletuksia kumotaan vain, mikäli todisteet riittävät.

Hypoteesien testaus

Tilastolliseen testaukseen sisältyy kahdentyyppisiä hypoteeseja. Testattavaa oletusta kutsutaan *nollahypoteesiksi*. Jos todisteet riittävät sen kumoamiseen, voimaan astuu *vastahypoteesi*. Testaus eli todisteiden kokoaminen tapahtuu jollakin tilastollisella *merkitsevyystestillä*. Erityyppisiä testausasetelmia on lukuisia, samoin erilaisia testejä, mutta yhteistä niille on testauksen ja päättelyn periaate:

1. asetetaan nollahypoteesi ja vastahypoteesi,
2. kerätään ”todisteet” yhteen *testisuureeksi*,
3. tiivistetään testaustulos p*-arvoksi* ja
4. tehdään p-arvon perusteella *johtopäätökset*.

Testin p-arvo eli *havaittu merkitsevyystaso* kertoo, kuinka vahvat todisteet nollahypoteesia vastaan on esitetty. *Tilastollisesti merkitsevä* p-arvo tarkoittaa vahvoja todisteita, mutta se, onko tulos *sisällöllisesti merkittävä*, on tutkijan pääteltävä. Tilastollinen merkitsevyys ei sellaisenaan tarkoita paljoakaan eikä välttämättä takaa mitään oikeasti merkittävää. Se on vain yksi päättelyn apuneuvo, jota usein korostetaan aivan liikaa.

Monissa kyselytutkimusten testausasetelmissa voidaan pitää riittävänä todisteena nollahypoteesia vastaan noin 0.05:n suuruista p-arvoa. Se vastaa tällöin *viiden prosentin riskiä* tehdä päättelyssä väärä johtopäätös, hylätä paikkansa pitävä nollahypoteesi. Tätä ”riskirajaa” sovelletaan yleisesti. Päättely ei kuitenkaan saa olla liian mekaanista, esimerkiksi p-arvo 0.049 on käytännössä aivan sama kuin 0.051.

Dramaattisempia tulkintoja p-arvoille saadaan tilanteissa, joissa tilastollinen riski kytkeytyy johonkin todelliseen riskiin. Klassinen esimerkki on uuden lääkkeen testausasetelma, jossa p-arvo kytkeytyisi lääkkeen käyttäjän riskiin sairastua vakavasti. Lääke tuskin pääsisi markkinoille, jos riski olisi ”viiden prosentin luokkaa”, siis jos yksi 20:stä saattaisi sairastua vakavasti.

Uskottavuuspäättely ja mallintaminen

Keskeisellä sijalla tilastollisessa päättelyssä on niin kutsuttu *uskottavuuspäättely*, joka tarkoittaa, että kiinnostaville parametreille pyritään löytämään aineiston valossa uskottavimmat estimaatit. Vastaavaa estimointimenetelmää, jota kutsutaan *suurimman uskottavuuden (maximum likelihood)* menetelmäksi, sovelletaan monissa yhteyksissä, kuten faktorianalyysissa, kohdassa 4.3 (s. 93).

Mallintamista voidaan tyypitellä teoriapainotteisuuden perusteella. *Konfirmatorinen* mallintaminen edellyttää vankkaa sisällöllistä teoriapohjaa ja yksityiskohtaisempia hypoteeseja, joita voidaan testata tilastollisesti. Usein siinä tarvitaan myös erikoistuneempien menetelmien ja ohjelmistojen tuntemusta. *Eksploratiivinen* mallintaminen perustuu aineistolähtöiseen toimintaan, jossa tilastollisen testauksen sijaan korostuu aineiston kuvaaminen. Käytännön tutkimusasetelmat sijoittuvat yleensä jonnekin näiden ääripäiden väliin. Tämän kirjan lähestymistapa vastaa tyypillisiä kyselytutkimuksen tilanteita ja on siten melko eksploratiivinen.

Epävarmuuksia ja riskejä

Tilastollisilla malleilla on yksi yhteinen piirre: niiden avulla ei ilmaista tarkkoja totuuksia, ainoastaan todennäköisyyksiä. Mallinnetaan mitä tahansa ilmiöitä kuinka hyvin tahansa, niin mukana on myös liuta epävarmuustekijöitä. Niiden vaikutuksia voidaan arvioida ja kenties vähentää muttei täysin poistaa. Johtopäätöksiin sisältyy epävarmuuksia ja riskejä, joiden hallinta kuuluu tilastolliseen mallintamiseen ja päättelyyn.

Osa epävarmuuksista voi johtua ilmiön teorian tuntemuksen puutteesta: asiaan vaikuttavia tekijöitä ei ole osattu ottaa huomioon riittävästi tai teoria ei kaikilta osin päde oletetulla tavalla. Osa epävarmuuksista johtuu tiedonkeruusta: otos ei vastauskadon myötä edusta

perusjoukkoa tai näyteaineistoon ei saatu tavoiteltuja henkilöitä. Oli tiedonkeruun tapa ja teorian osuus mikä hyvänsä, niin osa epävarmuuksista johtuu mittauksesta: joidenkin käsitteiden operationalisointi on epäonnistunut, osaa kysymyksistä ei ole ymmärretty, tai ne on ymmärretty eri tavalla kuin tutkija on tarkoittanut, tai eri vastaajat ovat ymmärtäneet ne eri tavoin. Kenties kysymyksiä on ollut liikaa, ja vastausväsymyskin on alkanut vaikuttaa. Lisäksi vastauksia voi hämärtää *sosiaalinen suotavuus*, joka tarkoittaa sitä, että vastaaja pyrkii vastauksillaan välittämään itsestään myönteisen vaikutelman. Epävarmuuden lähteitä on siis lukuisia.

Joitakin epävarmuuksia voidaan hallita tekemällä sopivia oletuksia, mutta oletukset voivat myös aiheuttaa lisää epävarmuuksia, etenkin jos niitä on liikaa tai ne ovat epärealistisia. Oletuksia on helppo tehdä, mutta pitäisi myös tutkia, missä määrin ne pitävät paikkansa. Mallit ovat sitä yleiskäyttöisempiä ja tulokset sitä uskottavampia, mitä vähemmillä oletuksilla toimitaan.

Mallit ja menetelmät

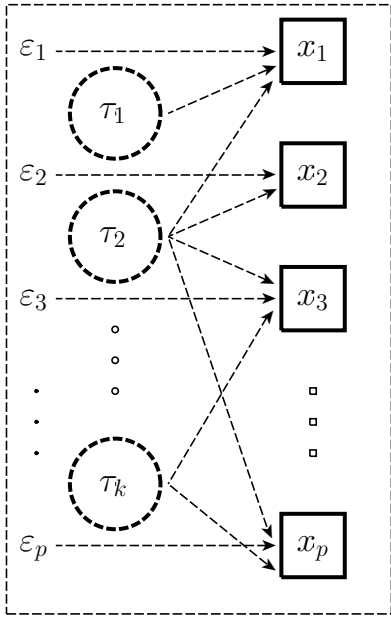
Tilastollisista malleista tehtyjen johtopäätösten yleistäminen on mahdollista, mikäli aineisto muodostaa edustavan otoksen jostakin perusjoukosta. Kiintoisia analyyseja voidaan tehdä myös aineistolähtöisesti, jolloin ei tarvita teoriasta johdettuja malleja vaan menetelmät riittävät. Analyysien perusteella tehdyt päätelmät rajoittuvat tällöin enemmän aineiston tasolle.

Tässä kirjassa käsitellään eräitä tyypillisimpiä malleja ja menetelmiä. Mallipohjaisempia menetelmiä edustavat etenkin perusmenetelmät *faktorianalyysi* ja *regressioanalyysi*, kun taas puhtaasti aineiston kuvaamisen menetelmiä ovat esimerkiksi *hierarkkinen ryhmittely* ja *korrespondenssianalyysi*. Yhteistä näille menetelmille on, että analysoitavana on yhtaikaa useita muuttujia. Siksi niitä kutsutaan monimuuttujamenetelmiksi.

Tässä kirjassa käsiteltävillä menetelmillä selviää käytännössä pitkälle. Turhan monimutkaisia menetelmiä on jopa syytä välttää. Vaikka itse osaisi niitä soveltaa, voivat tulokset jäädä muilta ymmärtämättä. ”Hienompia” malleja tai menetelmiä ei ole syytä edes kokeilla, ellei hallitse perusmenetelmiä kunnolla.

4.2 Mittausmalli

On aika palata luvussa 2 hahmoteltuun mittausmalliin ja tarkentaa sen määrittelyjä. Kuva 4.1 on muutoin täysin sama kuin kuva 2.1 (s. 21), mutta kysymysmerkit on nyt korvattu uusilla merkinnöillä. Näitä mittausmallin matemaattisesta esityksestä periytyviä merkintöjä käytetään vain käsitteiden tarkempaan määrittelyyn, eikä niitä vastaavia kaavoja tässä yhteydessä tarvita. Riittää tarkastella kaavojen sijasta kaaviota ja ymmärtää sen eri osien merkitykset.



Kuva 4.1. Mittausmalli.

Mittausmalli on esimerkki tilastollisesta mallista. Se täsmentää, mitä mitataan ja miten. Malli on ajatusrakennelma, jonka tutkija laatii työpöydällään. Mitä enemmän käytettävissä on ilmiötä koskevaa teoriaa tai aiempiin tutkimuksiin perustuvaa muuta tietoutta, sitä vankempi rakennelma tulee olemaan.

Käsitteet ja merkinnät

Mittausmalli sisältää kolmenlaisia käsitteitä (ks. kuva 4.1):

1. *Tosiarvoja*, jotka on piirretty ympyröinä ja merkitty kreikkalaisilla t-kirjaimilla *(tau)*. Tosiarvot vastaavat tutkittavan ilmiön ulottuvuuksia, joiden lukumäärä k on tärkeimpiä mittausmallin oletuksia. Sen määrääminen on ensisijaisesti sisällöllinen ja korkeintaan toissijaisesti tilastollinen haaste.

2. *Osioita*, jotka on piirretty neliöinä ja merkitty x-kirjaimin. Osiot ovat kysymyksiä tai väitteitä, joita on p kappaletta. Yleensä ne muodostavat yhden tai useampia mittareita, joilla tosiarvoja pyritään mittaamaan. Oletus on, että p on suurempi kuin k eli osioita on enemmän kuin tosiarvoja. Muistisääntönä voi ajatella, että tosiarvoja on ”kohtalaisesti” ja osioita ”paljon”. Määrä ei kuitenkaan kelpaa perusteeksi hyvälle mittaukselle.

3. *Mittausvirheitä*, joita on merkitty kreikan e-kirjaimin *(epsilon)* ja jotka väijyvät kunkin osion taustalla. Niitä ei voi kokonaan välttää, mutta mitä vähemmän mittausvirhettä, sitä parempi on mittauksen reliabiliteetti.

Tosiarvojen – siis esimerkiksi todellisten asenteiden tai arvojen – ajatellaan vaikuttavan siihen, miten ihminen lomaketta täyttäessään vastaa hänelle esitettyihin kysymyksiin tai väitteisiin. Siitä syystä mittausmallin nuolet osoittavat tosiarvoista osioihin. Kuten kuvasta 4.1 ilmenee, kutakin tosiarvoa kohti voi ja on hyväkin olla useita osioita. Vastaavasti yksi osio voi mitata useampaa tosiarvoa; se on moniulotteisessa mittaamisessa aivan luonnollista.

Tosiarvoja ja mittausvirheitä on tapana merkitä kreikkalaisilla kirjaimilla sen vuoksi, että ne ovat teoreettisia käsitteitä, joita ei voi suoraan havaita. Ne ovat osa ajatusrakennelmaa samoin kuin katkoviivoin piirretyt nuolet, tosiarvojen ympyrät ja koko mittausmallin kehys. Osiot ovat ainoat käytännössä vastaajalle näkyvät mittausmallin osat. Pyrkimys on mitata tosiarvoja, mutta se on mahdollista vain epäsuorasti, osioiden välityksellä.

Ulkonäkötutkimuksen mittausmalli

Luvussa 2 hahmoteltiin ulkonäkötutkimuksesta kysymyksenasettelu, jossa suomalaisten naisten ulkonäkökäsityksiä tarkastellaan kolmen ulottuvuuden kannalta. Ulottuvuudet ovat 1) *itsetunto ulkonäköasioissa*, 2) *panostaminen ulkonäköön* ja 3) *sosiaaliset ulkonäköpaineet*. Näitä mitataan kolmella mittarilla, joista ensimmäiseen sisältyy 22, toiseen 20 ja kolmanteen 11 osiota.

Aineiston tiivistämistä ajatellen hyötysuhde vaikuttaisi erinomaiselta. Mallia on hahmoteltava paljon tarkemmin pohtimalla muun muassa, minkä osioiden oletetaan mittaavan mitäkin tosiarvoja. Näitä asioita on syytä miettiä perusteellisesti jo mittareita laatiessa. Tässä yhteydessä yksityiskohtaisempi pohdiskelu sivuutetaan.

Kun malli on tällä tarkkuudella laadittu, voidaan alkaa koetella sen toimivuutta. Tilastollisen mallintamisen seuraavassa vaiheessa valitaan tarkoituksenmukainen menetelmä, jolla mallin parametrit estimoidaan aineiston perusteella. Kiinnostavia ovat muun muassa seuraavat ulottuvuuksia ja osioita koskevat kysymykset:

- Tukeeko aineisto oletusta kolmesta ulottuvuudesta?
- Ovatko ulottuvuudet nimettävissä kuten ajateltiin?
- Mitkä mallin osioista toimivat parhaiten, mitkä huonoiten?
- Millaisia yhteyksiä tosiarvojen ja osioiden välillä vallitsee?

Vastauksia saadaan tutkimalla ja tulkitsemalla menetelmän, tässä tapauksessa faktorianalyysin, tulostuksia.

4.3 Faktorianalyysi

Faktorianalyysi on tilastollinen menetelmä, jolla mitä erilaisimpia mittausmallirakennelmia voidaan tarkastella havaintoaineistojen valossa, niin kyselytutkimuksessa kuin muissakin yhteyksissä. Kyseessä on yksi perinteisimmistä menetelmistä, jota on tutkittu ja sovellettu laajalti jo yli vuosisadan ajan. Faktorianalyysia pidetään yhtenä tilastotieteen menestystarinana erityisesti yhteiskuntatieteissä (Cudeck & MacCallum, 2007).

Sana ”faktori” tarkoittaa mittauskehikossa samaa kuin tosiarvo tai ulottuvuus. Kirjallisuudessa käytetään myös nimitystä *yhteisfaktori* erotukseksi *ominaisfaktoreista*, jotka tarkoittavat osiokohtaisia, satunnaisia häiriötekijöitä. Mittauskehikossa ominaisfaktorit tulkitaan mittausvirheiksi.

Faktorianalyysi koostuu useasta vaiheesta lähtien mittausmallin parametrien estimoinnista ja päätyen varsinaiseen aineiston tiivistämiseen mitta-asteikoiksi. Tässä luvussa näitä vaiheita selvennetään sekä yleisellä tasolla että ulkonäkötutkimuksen esimerkkien avulla.

4.3.1 Oletukset

Kuten tilastolliset menetelmät yleensä, myös faktorianalyysi toimii tiettyjen oletusten varassa. Mitä paremmin oletukset pätevät, sitä luotettavampia analyysin tulokset ovat, ainakin tilastolliselta kannalta. Sisällöllisesti luotettavuuteen vaikuttavat muutkin seikat.

Osa faktorianalyysin oletuksista liittyy lähinnä menetelmän teoreettisiin tarkasteluihin. Tilastotieteen teoria ja sovellusalojen käytäntö ovat kuitenkin kaksi eri maailmaa, ja käytännössä faktorianalyysi, kuten muutkin monimuuttujamenetelmät, sietävät melko hyvin poikkeamia teoreettisista oletuksista. Tietyt keskeiset oletukset on kuitenkin syytä tiedostaa, jotta osaa soveltaa faktorianalyysia oikealla tavalla ja oikeanlaisissa tilanteissa.

Havaintojen riippumattomuus

Havaintojen oletetaan olevan toisistaan riippumattomia. Kyselytutkimuksessa tämä tarkoittaa, etteivät eri vastaajien antamat vastaukset saisi riippua toisistaan ajan tai paikan suhteen. Tämä oletus rajaa käytännössä pois muun muassa aikasarja-aineistot, pitkittäistutkimusaineistot ja hierarkkiset koesuunnitteluaineistot. Kaikissa näissä tutkimusasetelmat ovat sellaisia, että havainnot riippuvat enemmän tai vähemmän toisistaan. Faktorianalyysin kannalta riippuvuutta on tällöin ”väärässä suunnassa”.

Silloin kun havaintojen riippumattomuusoletus pätee, aineiston havaintojen järjestyksellä ei ole väliä. Edellä mainittuja aineistotyyppejä ei sen sijaan voi järjestää mielivaltaiseen järjestykseen aineiston sisällön kärsimättä. Esimerkiksi aikasarja-aineistossa havaintojen ajallinen järjestys on aineiston olennainen ominaisuus. Tällaisten aineis-

tojen analysointiin tarvitaan pidemmälle meneviä faktorianalyysin muunnelmia ja muita menetelmiä, joita ei tässä kirjassa käsitellä.

Tavanomaisessa kyselytutkimusaineistossa ei haittaa, että osa vastaajista on samalta alueelta – sehän on päinvastoin aivan tyypillistä. Mahdollisia vastaajien välisiä alueellisia riippuvuuksia voidaan tutkia tavanomaisten taustamuuttujien avulla. Parhaiten tällaisia riippuvuuksia hallitaan tarkoituksenmukaisella otanta-asetelmalla. Tutkimusajankohdan selkeä määrittely ja dokumentointi auttaa arvioimaan havaintojen mahdollisia riippuvuuksia aineiston ulkopuolisista tekijöistä kuten yhteiskunnallisista tilanteista ja tapahtumista.

Muuttujien riippuvuus ja jakauma

Muuttujien välillä oletetaan ilmenevän selkeitä riippuvuuksia. Tämä on siinä mielessä selvää, että mittareita konstruoitaessa pyritään kehittämään useampia, samaa ulottuvuutta mittaavia osioita. Riippuvuuksien oletetaan olevan lineaarisia, toisin sanoen faktorianalyysi perustuu muuttujien välisiin korrelaatioihin (ks. kohdat 3.4.2, s. 71 ja 3.4.3, s. 77). Selvimmin näistä tarkasteluista pois jäävät luokittelutasoiset muuttujat. Niille on käyttöä myöhemmin muun muassa jatkoanalyysien taustamuuttujina.

Faktorianalyysin taustalla on periaatteessa oletus normaalijakaumasta, jonka pätiessä kohdassa 3.4.3 (s. 77) mainitut tyhjentävät tunnusluvut, siis keskiarvot, keskihajonnat ja korrelaatiot, riittävät aineiston olennaisen informaation kuvaamiseen. Tarkemmin sanottuna oletus koskee muuttujien yhteisjakaumaa, jota kutsutaan *multinormaalijakaumaksi*. Oletus ei oikeastaan päde käytännössä, sillä tyypilliset kyselytutkimuksen mittaukset eivät ole edes yksittäin normaalisia. Faktorianalyysi sietää kuitenkin tällaiset poikkeamat hyvin.

Aineiston koko

Eräs yleisimpiä tilastotieteilijälle esitettyjä kysymyksiä niin faktorianalyysin kuin muidenkin analyysien yhteydessä koskee aineiston kokoa, etenkin havaintojen lukumäärää. Kysymykseen ei ole yksiselitteisiä vastauksia, sillä asiaan vaikuttavat monet seikat, joista vain osa on tilastollisia. Aineiston tavoitekoko voi määräytyä tutkimusasetelman perusteella, mutta lopullisen koon ratkaisevat usein tiedonkeruun kustannukset ja aikataulu, vastausprosentti ja vastausten laatu. On

helppo neuvoa keräämään niin paljon havaintoja kuin mahdollista, mutta laatua ei voi korvata määrällä. Myös pienempien aineistojen analysointi on mahdollista.

Mittausmallin perusteella on jo selvää, että muuttujia on oltava enemmän kuin faktoreita. Havaintojen osalta vaatimus ei ole näin yksiselitteinen. Jonkinlaisena miniminä voidaan pitää muuttujien ja faktoreiden määrien tuloa, sillä se vastaa tärkeimpien analyysilla estimoitavien parametrien määrää. Muiden parametrien takia siihen on vielä lisättävä muuttujien määrä kertaalleen.

Vaikka faktoreiden määrä olisi kohtuullinen eikä muuttujia olisi paljonkaan, on estimoitavia parametreja pian useita kymmeniä tai satoja. Estimoinnin vakauden ja tulosten luotettavuuden kannalta olisi hyvä, että havaintoja olisi kutakin parametria kohden useita, eikä niin, että jonkin parametrin estimointi on yhden havainnon varassa.

Ulkonäkötutkimuksen esimerkkitilanteessa, jota seuraavaksi käsitellään, rajoitutaan aineiston ensimmäiseen keruuvuoteen 1997, koska ei voida olettaa, että tilanne olisi rakenteellisesti samanlainen vuonna 2005. Tällöin ollaan jo aineiston koon suhteen hieman arveluttavilla rajoilla, sillä havaintoja jää käyttöön vain 268 senkin jälkeen, kun puuttuvat tiedot on paikattu (ks. kohta 3.5, s. 81). Kolmen faktorin ja 53 muuttujan mallissa estimoitavia parametreja on kaikkiaan 212. Kun asetelmaa lavennetaan neliulotteiseksi, parametrien lukumäärä nousee jo 265:een, siis käytännössä samaan kuin osa-aineiston koko. Johtopäätöksiä tehdessä on viisasta välttää ylitulkintoja.

4.3.2 Faktoreiden tulkinta

Mittausmallin parametrien estimointia faktorianalyysilla kutsutaan *faktoroinniksi*. Lähtökohtana on kohdassa 3.4.3 (s. 77) käsitelty korrelaatiomatriisi, johon sisältyvä informaatio pyritään tiivistämään määritellyn mittausmallin mukaisesti.

Tärkeimpiä estimoitavia parametreja kutsutaan *faktorilatauksiksi*. Ne heijastelevat faktoreiden ja osioiden välisiä yhteyksiä, joita kuvassa 4.1 (s. 91) havainnollistivat näiden väliset nuolet. Koska yhteydet oletetaan lineaarisiksi, ovat faktorilatauksetkin korrelaatioita ja siten periaatteessa helposti tulkittavissa. Muita parametreja ovat mittausvirheiden varianssit, joita on yhtä paljon kuin osioita.

Tehdään nyt faktorianalyysi ulkonäkötutkimuksen mittausmallista, jonka pohtiminen aloitettiin jo luvussa 2. Korrelaatiomatriisi

53 muuttujasta perustuu siis 268 havaintoon, ja tiedot halutaan tiivistää kolmeksi faktoriksi. Tältä pohjalta parametrien estimointi suurimman uskottavuuden faktorointimenetelmällä tuottaa seuraavaksi tarkasteltavasta tulosteesta 4.1 löytyvät 216 lukua. Nyt pitäisi vain saada selville, mitä nuo luvut tarkoittavat ja mitä niistä voi päätellä.

Tuloste 4.1. Kolme faktoria vuoden 1997 aineistosta.

```
          F1    F2    F3  Comm
k26.1   0.71 -0.27  0.15  0.60 Olen tyytyväinen ulkonäköön.
k26.2   0.73 -0.10  0.05  0.54 Vaatteet näyttävät hyvältä päälläni.
k26.3   0.74 -0.28  0.14  0.64 Pidän ulkonäöstäni juuri sellaisenaan.
k26.4  -0.33  0.21  0.22  0.20 En pidä kehostani.
k26.5   0.59  0.18 -0.03  0.38 Olen naisellinen.
k26.6  -0.46  0.07  0.37  0.35 En ole fyysisesti viehättävä.
k26.7   0.53 -0.16  0.41  0.47 Olen aina hyvännäköinen ajankohdasta...
k26.8   0.56 -0.18  0.25  0.40 Laittautumattakin näytän hyvälle.
k26.9   0.70  0.10 -0.16  0.52 Kehoni on seksuaalisesti viehättävä.
k26.10  0.54 -0.26  0.41  0.53 Olen aina ollut tyytyväinen ulkonäköön.
k26.11  0.40  0.06  0.28  0.25 Ulkonäkö kertoo millainen ihminen olen.
k26.12 -0.58  0.02  0.37  0.48 Olen ruma.
k26.13  0.54 -0.11  0.21  0.35 Ulkonäköni vastaa sisäistä minääni.
k26.14 -0.19  0.28 -0.18  0.15 Suhtautumiseni ulkonäköön vaihtelee.
k26.15  0.08  0.36 -0.27  0.21 Meikattuna olen tyytyväisempi ulkonäköön.
k26.16  0.35  0.49  0.07  0.36 Ulkonäköni on tärkeä osa minua.
k26.17  0.66  0.13 -0.02  0.45 Olen kaunis nainen.
k26.18 -0.68  0.13  0.26  0.55 En pidä ulkonäöstäni.
k26.19  0.52 -0.05  0.15  0.30 Olen fyysisesti hyvässä kunnossa.
k26.20  0.43  0.45  0.31  0.48 Minulle on tärkeätä, että näytän hyvälle.
k26.21  0.18  0.33 -0.07  0.15 Tiedän, jos olen "huonosti laitettu".
k26.22  0.65 -0.16  0.01  0.44 Pidän siitä mille näytän ilman vaatteita.
k30.1   0.25  0.42  0.10  0.25 Katson aina miltä näytän kun lähden...
k30.2   0.24  0.53  0.26  0.40 Tarkastan ulkonäköni peilistä aina,...
k30.3   0.18  0.62  0.13  0.44 Käytän aikaa itseni "laittamiseen"...
k30.4   0.11  0.61  0.15  0.41 Yritän aina parantaa ulkonääköäni.
k30.5   0.42  0.51 -0.16  0.46 Nautin kun ihmiset katsovat minua.
k30.6  -0.46  0.11  0.30  0.31 En mielelläni käy yleisillä rannoilla...
k30.7  -0.14  0.46 -0.05  0.24 Pukeudun niin, ettei "heikot kohtani"...
k30.8   0.33  0.44 -0.23  0.36 Ostan vaatteita, joissa näytän hyvälle.
k30.9  -0.19 -0.36  0.38  0.31 En välitä miltä vaatteeni näyttävät...
k30.10  0.26  0.48 -0.14  0.31 Pukeudun mielelläni seksikkäästi.
k30.11  0.25  0.31  0.20  0.20 Liikun pitääkseni vartaloni "kunnossa".
k30.12 -0.25 -0.16  0.54  0.38 Yritän olla huomaamattoman näköinen.
k30.13 -0.13  0.30  0.00  0.11 Olen aikonut mennä plastiikkakirurgille.
k30.14  0.05  0.46 -0.24  0.27 Viehättävänä olen myös halukkaampi...
k30.15  0.29  0.51  0.03  0.34 Pyrin herättämään huomiota ulkonäölläni.
k30.16 -0.32  0.19  0.37  0.28 Ulkonäköni takia jätän osallistumatta...
k30.17 -0.38  0.12  0.38  0.30 En osallistu iltamenoihin ulkonäköni...
k30.18 -0.20 -0.41  0.19  0.24 Käytän vähän kauneudenhoitotuotteita.
k30.19  0.14  0.27  0.13  0.11 Kiinnitän erityistä huomiota hiuksiini.
k30.20  0.14  0.60  0.19  0.41 Käytän aikaa ulkonäköni tutkimiseen.
k71.1  -0.13  0.14  0.09  0.05 Ulkonäkö on yliarvostetussa asemassa.
k71.2  -0.04  0.37  0.03  0.14 Miellyttävästä ulkonäöstä on hyötyä.
k71.3  -0.08  0.36  0.30  0.23 Hyvännäköiset pärjäävät elämässään...
k71.4  -0.04  0.36  0.21  0.17 Hyvännäköiset ihmiset ovat suositumpia.
k71.5  -0.20  0.21  0.01  0.08 Hoikkuuden ihannointi asettaa paineita.
k71.6  -0.07  0.23 -0.04  0.06 Ulkonäkövaatimukset naisille ovat kovia.
k71.7  -0.31  0.43  0.03  0.29 Median naiskuva vähentää tyytyväisyyttä...
k71.8  -0.39  0.42  0.08  0.34 Koetan olla kauneusihanteiden mukainen.
k71.9   0.05 -0.29 -0.08  0.09 Elämässä pärjää ulkonäöstä riippumatta.
k71.10 -0.11  0.21 -0.00  0.06 Nuoria ja hoikkia naisia ihannoidaan.
k71.11 -0.07  0.20 -0.06  0.05 Naisten ulkonäkö merkitsee enemmän...
Sumsqr  8.05  5.85  2.59 16.49
```

Tulosteen tulkintaa

Tässä vaiheessa tehdään nimenomaan *tulosteen* eikä *tulosten* tulkintaa. Parhaimmillaan ahertaminen voi johtaa tuloksiin ja syvällisempiin tulkintoihin, mutta se vaatii yleensä useita vaiheita, joista tämä on vasta ensimmäinen. Mikään ei takaa, että mallin parametrien estimointi osuisi heti kohdalleen.

Faktorianalyysista paljon olennaista tiivistyy tulosteeseen 4.1, mutta kuten myöhemmin havaitaan, tiedot voidaan esittää selvemminkin. Tulosteessa on yksi rivi kutakin analyysiin valittua muuttujaa kohden sekä otsikko- ja yhteenvetorivi. Vasemmalla olevista muuttujien nimistä voi päätellä, että kyseessä on kolme mittaria, jotka on numeroitu alkuperäisen lomakkeen mukaisesti `k26`, `k30` ja `k71`. Oikealla on näihin kuuluvien osioiden sanalliset sisällöt. Alkuperäisiä kyselylomakkeen sanamuotoja on tilankäytön takia tiivistetty, mutta kokonaiskuvan saamiseen ne soveltuvat hyvin ja ovat helpompia hahmottaa. Osioiden yksityiskohtainen tutkiskelu kuuluu aineiston perustarkasteluihin (ks. luku 3), joten todellisuudessa osiot tunnetaan tässä vaiheessa jo hyvin.

Pystyrivien otsikoissa ovat faktoreiden ”nimet” `F1`, `F2` ja `F3`. Sen kummempia eivät ohjelmistot tietenkään kykene tarjoamaan. Eräs tulkinnan tärkeimpiä asioita on faktoreiden nimeäminen. Sitä varten pitää perehtyä tulosteen lukuihin.

Luvuista suurin osa on faktorilatauksia, ja niihin myös tulkinta enimmiltä osin perustuu. Niistä muodostuu niin sanottu *faktorimatriisi*. Lataukset, jotka siis kuvaavat faktoreiden ja osioiden välisiä yhteyksiä korrelaatioina, voivat olla positiivisia tai negatiivisia, yhteyden luonteesta riippuen. Tulosteessa 4.1 valtaosa latauksista on positiivisia, mutta negatiivisiakin on jonkin verran. Esimerkiksi faktorin `F1` ja osion `k26.3` välinen lataus on 0.74, kun taas osion `k26.18` lataus samalle faktorille on –0.68. Vastakkaisuus selittyy helposti tutkimalla osioiden sanamuotoja (vrt. kohta 3.3, s. 64).

Faktorilatauksia tarkastelemalla voi tuntua siltä, että jokainen osio latautuu jokaiselle faktorille. Se ei vaikuta miellyttävältä, sillä mallintamisen idea olisi ilmaista asioita tiiviimmin. Tarkemmin katsottuna lukujen suuruudet vaihtelevat. Mitä suurempi lataus osiolla on jollekin faktorille, sitä pienempiä sen lataukset näyttävät olevan muille faktoreille. Tämä pätee yleisesti: sama osio ei voi latautua voimakkaasti usealle faktorille. Se on myös luontevaa – pyritäänhän

osiot alun alkaen laatimaan niin, että ne mittaisivat vain yhtä asiaa. Puhtaasti niin yksinkertainen rakenne ei ole realistinen, koska tutkittavat ilmiöt ovat moniulotteisia. Luonnostaan mukana on myös osioita, jotka latautuvat useammalle kuin yhdelle faktorille.

Tulosteen 4.1 neljännen pystyrivin (`Comm`) lukuja tutkimalla saadaan käsitys mittausvirheiden vaikutuksista. Luvut lähestyvät asiaa positiivisessa valossa, sillä ne ilmaisevat, kuinka suuri osa kunkin osion vaihtelusta tulee tiivistetyksi faktoreilla. Nämä luvut, joita kutsutaan nimellä *kommunaliteetti*, eivät siis ole korrelaatioita vaan suhteellisia osuuksia. Mitä lähempänä kommunaliteetti on ykköstä, sitä paremmin osio mittausmallissa toimii. Parhaisiin osioihin näyttäisivät lukeutuvan ainakin edellä mainitut `k26.3` ja `k26.18`. Jos kommunaliteetti on lähellä nollaa, ei osiosta saada juuri hyötyä irti mallissa. Silloin joudutaan tulkitsemaan, että osion vaihtelusta suurin osa johtuu mittausvirheistä. Tällaisia osioita näyttäisi tulosteessa olevan jonkin verran, erityisesti `k71`-mittarissa.

Jäljellä on enää tulosteen 4.1 alimman, yhteenvetorivin tulkinta. Sen luvut, jotka ovat aivan eri suuruusluokkaa kuin faktorilatausket tai kommunaliteetit, kuvaavat faktoreiden *voimakkuuksia*, eli sitä, miten voimakkaasti osiot kaikkiaan ovat niihin yhteyksissä. Faktorit esiintyvät tulosteessa voimakkuusjärjestyksessä. `Sumsqr`-otsikko viittaa siihen, että voimakkuudet ovat faktorilatausten pystyriveittäin laskettujen neliöiden summia. Kommunaliteetit ovat vastaavia lukuja vaakariveittäin laskettuina. Lukuja ei tarvitse itse laskea, vaan ne sisältyvät faktorianalyysin tulostuksiin.

Aivan viimeinen luku tulosteessa 4.1 on voimakkuuksien summa 16.5, joka on samalla myös kommunaliteettien summa. Siitä voi päätellä, miten hyvin malli kokonaisuudessaan tiivistää faktoreiden ja osioiden välisiä yhteyksiä ja miten suuri osuus on mittausvirheillä. Korrelaatioiden myötä muuttujien varianssit ovat ykkösiä. *Kokonaisvaihtelusta*, jota on alun perin 53 yksikön verran, on faktorianalyysi tiivistänyt *yhteisvaihteluksi* 16.5 yksikköä, siis noin 30 %, josta ensimmäisen faktorin osuus on noin puolet. Mittausvirheiden tiliin menee suurin osa, lähes 70 %, joten ei olisi kovin järkevää käyttää osioita jatkoanalyyseissa sellaisenaan. Faktorianalyysilla saadaan parhaiten käyttöön jatkon kannalta olennainen informaatio. Tärkein tavoite ei ole vaihtelun määrällinen tiivistäminen, vaan faktoreiden ja osioiden välisten yhteyksien kuvaaminen.

4.3.3 Mittausmallin rakennevaliditeetti

Vaikka faktorianalyysin tulostuksen tärkeimmät osat ovat edellä tulleet selvitetyiksi, on kokonaiskäsityksen muodostaminen tulosteen 4.1 perusteella hankalaa. Se johtuu tulosteen *surkeasta esitystavasta*: siinä on kyllä kaikki tarvittavat ainekset, mutta ne ovat tulkintaa ajatellen väärässä järjestyksessä. Osioiden listaaminen numerojärjestyksessä ei ole kovin hyödyllistä.

Kuten tulosten esittämiseen, johon tässä kirjassa ei juuri puututa, myös tulosteiden esittämiseen on kiinnitettävä huomiota. Vaikka tulosteita lukisi vain tutkija itse analyyseja tehdessään, on tätä yksinäistä lukijaa autettava. Satojen lukujen ja numeroiden perusteella muodostettava tulkinta ei ole muutenkaan helppoa, ei etenkään, jos tulosteet ovat sellaisia kuin edellä käsitelty tuloste 4.1.

Tavoitteena on arvioida *mittausmallin rakennevaliditeettia*. Sitä varten pitää saada parempi käsitys faktorianalyysin onnistumisesta. Todellisuudessa arviointiin kuuluu myös sisällöllisiä näkökohtia, joihin ei tässä yhteydessä puututa. Rakennevaliditeetin arviointi nojaa siis seuraavassa vain tilastollisiin näkökohtiin.

Järjestelemällä tulostusta vuoroin faktoreiden voimakkuuksien, vuoroin kommunaliteettien perusteella sekä korostamalla voimakkaimpia latauksia ja heikoimpia kommunaliteetteja alkaa syntyä selvempi kuva siitä, mitä tuloste koettaa kertoa. Nämä dynaamiset vaiheet joudutaan tässä esityksessä sivuuttamaan, mutta johtopäätös niiden perusteella on, että mallin rakennevaliditeetti on ilmeisesti aika heikko: analyysi ei anna riittävää näyttöä mallin ajatellusta, kolmiulotteisesta rakenteesta. Erityisesti ulkonäköä koskevien sosiaalisten paineiden ulottuvuus ei tule esiin, vaan liian moni sitä mittaavan `k71`-mittarin osioista on käytännössä pelkkää mittausvirhettä. Osioita ei pidä kuitenkaan lähteä poistamaan; sitä olisi vaikea perustella sisällöllisesti. Sen sijaan herää kysymys, onko ennalta ajateltu faktoreiden määrä sittenkään oikea.

Tulosteen tulkintaa, uusi yritys

Luvun alussa viitattiin eksploratiiviseen ja konfirmatoriseen mallintamistapaan. Jos jälkimmäistä noudatettaisiin tiukasti, voitaisiin lopettaa äskeiseen johtopäätökseen ja todeta, ettei aineisto antanut tukea mallin pohjalta laadituille hypoteeseille. Tässä ei kuitenkaan olla niin

tiukkoja, vaan hypoteesien testauksen sijaan pyritään saamaan kokonaiskäsitys ilmiöstä. Analyysi on siis melko eksploratiivista, mutta siinä mielessä konfirmatorista, että se perustuu etukäteen pohdittuun mittausmalliin. Silloin on sallittua kokeilla muitakin vaihtoehtoja.

Tuloste 4.2. Neljä faktoria vuoden 1997 aineistosta.

	F1	F2	F3	F4	Comm	
k26.3	**0.82**	-0.02	-0.05	0.01	0.68	Pidän ulkonäöstäni juuri sellaisenaan.
k26.1	**0.78**	-0.01	-0.08	0.02	0.62	Olen tyytyväinen ulkonäköön.
k26.2	**0.70**	0.14	-0.11	-0.12	0.54	Vaatteet näyttävät hyvältä päälläni.
k26.22	**0.64**	0.05	-0.11	-0.12	0.44	Pidän siitä mille näytän ilman vaatteita.
k26.10	**0.63**	0.06	-0.17	**0.30**	0.52	Olen aina ollut tyytyväinen ulkonääköön.
k26.8	**0.63**	0.06	-0.03	0.13	0.41	Laittautumattakin näytän hyvälle.
k26.7	**0.60**	0.13	-0.12	0.28	0.47	Olen aina hyvännäköinen ajankohdasta...
k26.9	**0.58**	0.25	-0.05	**-0.34**	0.52	Kehoni on seksuaalisesti viehättävä.
k26.17	**0.57**	0.29	0.02	-0.21	0.46	Olen kaunis nainen.
k26.13	**0.57**	0.11	-0.09	0.09	0.35	Ulkonäköni vastaa sisäistä minääni.
k26.19	**0.50**	0.18	-0.15	0.02	0.30	Olen fyysisesti hyvässä kunnossa.
k26.5	**0.49**	**0.32**	0.03	-0.21	0.39	Olen naisellinen.
k26.11	**0.39**	0.26	-0.07	0.15	0.25	Ulkonäkö kertoo millainen ihminen olen.
k26.14	-0.26	0.11	0.21	-0.17	0.16	Suhtautumiseni ulkonääköön vaihtelee.
k26.4	**-0.34**	0.15	0.06	0.25	0.20	En pidä kehostani.
k30.6	**-0.39**	0.03	0.11	**0.37**	0.30	En mielelläni käy yleisillä rannoilla...
k71.8	**-0.42**	0.23	**0.37**	0.10	0.38	Koetan olla kauneusihanteiden mukainen.
k26.18	**-0.64**	0.00	0.04	**0.40**	0.57	En pidä ulkonäöstäni.
k30.3	-0.04	**0.68**	0.04	-0.05	0.46	Käytän aikaa itseni "laittamiseen"...
k30.20	-0.07	**0.67**	-0.02	0.02	0.46	Käytän aikaa ulkonäköni tutkimiseen.
k30.4	-0.09	**0.65**	0.04	-0.02	0.43	Yritän aina parantaa ulkonääöäni.
k26.20	0.29	**0.62**	0.03	0.09	0.48	Minulle on tärkeätä, että näytän hyvälle.
k30.2	0.09	**0.62**	0.09	0.08	0.40	Tarkastan ulkonäköni peilistä aina,...
k26.16	0.17	**0.56**	0.05	-0.12	0.36	Ulkonäköni on tärkeä osa minua.
k30.5	0.18	**0.55**	0.02	**-0.36**	0.47	Nautin kun ihmiset katsovat minua.
k30.15	0.11	**0.54**	0.10	-0.15	0.34	Pyrin herättämään huomiota ulkonäölläni.
k30.10	0.03	**0.49**	-0.01	**-0.30**	0.33	Pukeudun mielelläni seksikkäästi.
k30.1	0.11	**0.49**	0.04	-0.05	0.25	Katson aina miltä näytän kun lähden...
k30.11	0.13	**0.44**	-0.07	0.06	0.22	Liikun pitääkseni vartaloni "kunnossa".
k30.8	0.12	**0.43**	0.05	**-0.40**	0.36	Ostan vaatteita, joissa näytän hyvälle.
k71.3	-0.12	**0.36**	0.18	0.23	0.23	Hyvännäköiset pärjäävät elämässään...
k30.14	-0.12	**0.34**	0.18	**-0.33**	0.27	Viehättävänä olen myös halukkaampi...
k26.21	0.04	**0.33**	0.01	-0.18	0.15	Tiedän, jos olen "huonosti laitettu".
k71.4	-0.08	**0.33**	0.24	0.14	0.19	Hyvännäköiset ihmiset ovat suositumpia.
k30.19	0.06	**0.32**	0.02	0.03	0.11	Kiinnitän erityistä huomiota hiuksiini.
k30.7	-0.25	**0.32**	0.25	-0.10	0.24	Pukeudun niin, ettei "heikot kohtani"...
k71.2	-0.11	0.29	0.22	-0.03	0.14	Miellyttävästä ulkonäöstä on hyötyä.
k30.13	-0.21	0.23	0.10	-0.03	0.11	Olen aikonut mennä plastiikkakirurgille.
k71.9	0.13	-0.28	-0.03	-0.03	0.10	Elämässä pärjää ulkonäöstä riippumatta.
k30.18	0.02	**-0.41**	0.04	**0.32**	0.27	Käytän vähän kauneudenhoitotuotteita.
k71.5	-0.08	-0.03	**0.77**	0.03	0.60	Hoikkuuden ihannointi asettaa paineita.
k71.6	0.02	0.02	**0.75**	-0.06	0.57	Ulkonäkövaatimukset naisille ovat kovia.
k71.10	-0.00	0.01	**0.74**	-0.01	0.55	Nuoria ja hoikkia naisia ihannoidaan.
k71.1	-0.00	-0.03	**0.65**	0.12	0.43	Ulkonäkö on yliarvostetussa asemassa.
k71.11	-0.02	0.03	**0.55**	-0.08	0.31	Naisten ulkonäkö merkitsee enemmän...
k71.7	**-0.35**	0.23	**0.43**	0.02	0.36	Median naiskuva vähentää tyytyväisyyttä...
k30.12	-0.08	-0.07	-0.00	**0.61**	0.38	Yritän olla huomaamattoman näköinen.
k26.12	**-0.49**	-0.04	0.02	**0.49**	0.49	Olen ruma.
k30.9	0.01	-0.27	-0.07	**0.49**	0.31	En välitä miltä vaatteeni näyttävät...
k26.6	**-0.38**	0.03	0.06	**0.45**	0.35	En ole fyysisesti viehättävä.
k30.17	**-0.31**	0.09	0.12	**0.43**	0.30	En osallistu iltamenoihin ulkonäköni...
k30.16	-0.27	0.15	0.16	**0.40**	0.28	Ulkonäköni takia jätän osallistumatta...
k26.15	-0.10	0.29	0.02	**-0.35**	0.22	Meikattuna olen tyytyväisempi ulkonääköön.
Sumsqr	6.98	5.68	3.26	3.17	19.09	

s. 207

Samalla kun palataan tulosteen tulkintaan, kohennetaan edellä kritisoitua esitystapaa. Tulosteessa 4.2 on neljän faktorin ratkaisu, jonka perusainekset ovat samat kuin edellä, mutta esitettyinä niin, että tulkinnan muodostaminen sujuisi kätevämmin. Itse luvut eivät ole samat, joten tulkinta alkaa alusta.

Kun katsoo tulostetta 4.2, ei huomio ensimmäiseksi kiinnity osioihin, vaan faktoreihin niin kuin pitääkin. Tarkoitushan ei ole tehdä *osioanalyysia* vaan faktorianalyysia. Lataukset on järjestetty faktoreittain positiivisista negatiivisiin korostaen suurimpia, jolloin hahmottuu välittömästi neljä faktoria. Kolme niistä on mittausmallin mukaisia, voimakkuusjärjestyksessä 1) *itsetunto ulkonäköasioissa* (tulosteen `F1`), 2) *panostaminen ulkonäköön* (`F2`) ja 3) *sosiaaliset ulkonäköpaineet* (`F3`). Tältä osin rakennevaliditeetti alkaa vaikuttaa selvästi paremmalta. Faktorit voidaan ainakin näiden kolmen osalta nimetä, ja unohtaa tulosteen väliaikaiset, paremminkin formula-kilpailuihin viittaavat nimet.

Pienen lisähaasteen tulkinnalle aiheuttaa faktori `F4`, joka näyttää kokoavan yhteen ulkonäköä koskevia, enimmäkseen negatiivisluonteisia väittämiä kahdestakin eri mittarista (`k26` ja `k30`). Tämä faktori on voimakkuudestaan huolimatta kaikista heikoin, koska sillä on vähiten omia osioita eivätkä niiden lataukset ole yhtä korkeita kuin muiden faktoreiden vastaavat. Lähinnä se näyttäisi tuovan esiin passiivisen suhtautumisen ulkonäköön (esim. osiot `k30.12`, `k30.9` tai `k30.17`) tai karun realistisen näkemyksen siitä (esim. osiot `k26.12`, `k26.6` tai `k26.18`). Sille latautuvat osiot latautuvat myös kahdelle ensimmäiselle faktorille päinvastaisilla etumerkeillä. Kaikesta huolimatta negatiivissävyinen faktorikin lienee sisällöllisesti tulkittavissa, joten sekin pidetään jatkossa mukana tarkasteluissa.

Aina kun faktoreita lisätään, kommunaliteetit ja niiden summa kasvavat, koska suurempi osuus muuttujista tulee tiivistetyksi faktoreihin. Taikatempuksi tästä keinosta ei ole. Liian monessa faktorissa ei ole mieltä, koska rakenne hajoaa liiaksi. Sitäpaitsi, jos faktoreiden määrä ei enää ole sinne päinkään sama kuin alun perin, voi kysyä, mitä mittausmallia pohtiessa on tehty ja mitä aineiston osioiden on luultu mittaavan. Ellei mittausmallia ole lainkaan pohdittu, faktorianalyysi saattaa mennä tyystin arvailuksi.

Tulkinnan selkiyttäminen rotaatiolla

Edellä nähtiin, että tarkasteltu neljän faktorin analyysi johti kohtalaisen selkeään tulkintaan. Asiaa auttoi faktorimatriisin esittäminen niin, että tulkinta tuli mahdollisimman helpoksi. Esittäminen perustui faktorilatausten järjestämiseen ja korostamiseen, mutta myös uuden faktoroinnin jälkeen suoritettuun välivaiheeseen, jota kutsutaan *faktorirotaatioksi*.

Faktorointi ei tuota tuloksenaan yksikäsitteistä ratkaisua. Se kiinnittää muuttujien kommunaliteetit, mutta faktorien tulkintaan jää useita mahdollisuuksia, koska samat kommunaliteetit voivat muodostua eri latauksista lukemattomilla tavoilla. Koska faktorointimenetelmien kriteerit ovat laskennallisia, ei lopputulos yleensä ole sisällöllisesti ”parhaassa asennossa” vaan sitä on syytä rotatoida eli muuntaa niin että selkein tulkintasuunta tulee esiin. Toisinaan nähdyt arvostelut faktorianalyysin ”subjektiivisuudesta” voi jättää omaan arvoonsa. Tutkijan on joka tapauksessa tehtävä erilaisia valintoja ja perusteltava niitä. Mikään laskennallinen kriteeri ei tee analyysista sen objektiivisempaa. Sama koskee muitakin menetelmiä.

Rotaatiomenetelmät ovat olleet tutkimuksen kohteina moniulotteisen faktorianalyysin syntyajoista 1930-luvulta. Sitä ennen rotaatiota ei tarvittu, koska analyysin varhaisimmat muodot perustuivat vain yhteen faktoriin. Sittemmin rotaatiosta on tullut käytännössä kiinteä osa faktorianalyysia, usein pelkkä rutiininomainen vaihe. Siinä on kuitenkin omaa mielenkiintoa, jossa ei pitäisi luottaa pelkkiin rutiineihin.

Graafinen rotaatio

Faktorimatriisin järjestelyn ja latausten korostamisen tapaan rotaatio on keino nähdä syvemmälle aineistoon. Kirjallisessa esityksessä nämä työvaiheet eivät pääse oikeuksiinsa, ja siksi ne on tässä kuitattu vain lyhyesti. Osaavan tutkijan käsissä paras rotaatiomenetelmä on interaktiivinen *graafinen rotaatio*, josta kaikki aikanaan alkoikin. Menetelmä syrjäytyi 1960-luvulla laskemiseen rajoittuneiden tietokoneiden takia, mutta nykyiselle visuaalisuuden aikakaudelle graafinen rotaatio sopii kerrassaan hyvin. Toistaiseksi ainoa graafisen rotaation mahdollistava ohjelmisto lienee Survo (Mustonen, 1995, 81–82).

Edellä tutkittu ja tulkittu neljän faktorin ratkaisu on muodostettu graafisella rotaatiolla. Täysin samaa lopputulosta ei saa muilla rotaatiomenetelmillä. Käytännössä eniten sovellettu, 1960-luvulta periytyvä *varimax*-rotaatio mahdollistaa suurin piirtein saman tulkinnan, muttei anna kaikilta osin yhtä selvää kuvaa tilanteesta.

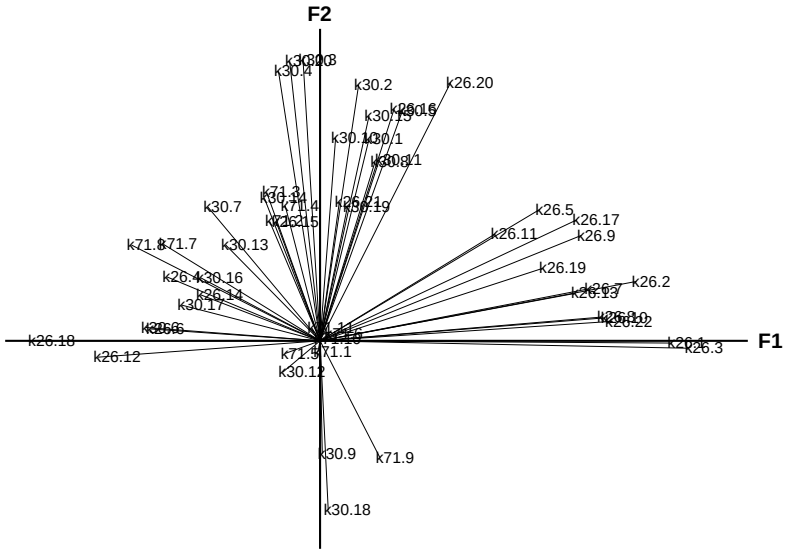
Suorakulmaiset ja vinot rotaatiot

Edellä mainitut rotaatiot ovat tyypiltään *ortogonaalisia*, mikä tarkoittaa, että faktorit ovat kohtisuorassa toisiaan vastaan, ne eivät korreloi keskenään, vaan edustavat selkeästi eri ulottuvuuksia. Vielä eri tavalla sanottuna niiden muodostama *faktoriavaruus*, johon muuttujat analyysissa sijoittuvat, on tavallinen suorakulmainen koordinaatisto.

On siis mahdollista piirtää esimerkiksi kaksi ensimmäistä faktoria vastakkain ikään kuin koordinaattiakseleiksi ja sijoittaa kuvaan muuttujat sen mukaan, miten hyvin ne latautuvat näille faktoreille. Kuvassa 4.2 muuttujat on piirretty origosta lähtevinä vektoreina. Mitä korkeampi lataus, sitä pidempi vektori ja sitä lähempänä muuttuja on vastaavaa faktoriakselia (vrt. tuloste 4.2, s. 101). Juuri tällaisten kuvien parissa vuorovaikutteinen graafinen rotaatio tapahtuu, mikä tekee siitä hyvin konkreettisen ja aineistoläheisen menetelmän.

Suorakulmainen rotaatio on suoraviivaisen tulkintansa vuoksi suositeltavampi ainakin tavanomaisissa, melko eksploratiivisissa tutkimusasetelmissa. Konfirmatorisissa asetelmissa ilmiön teoria saattaa tukea myös niin sanottua *vinorotaatiota*. Siirtyminen tutusta suorakulmaisesta koordinaatistosta vinokulmaiseen koordinaatistoon tarkoittaa, että faktoreiden annetaan korreloida keskenään. Voidaan tietysti ajatella, että esimerkiksi ”itsetunto ulkonäköasioissa” ja ”panostaminen ulkonäköön” korreloivat keskenään. Silti ei kannata ihan helposti lähteä tähän houkutukseen, sillä vinokulmaisen faktoriavaruuden tulkinta on vaikeaa.

On hyödyllisempää analysoida ulottuvuuksia toisistaan riippumattomina ja soveltaa suorakulmaista rotaatiota. Osaamattoman käsissä vinorotaatiot johtavat harhaan. Jos tuijottaa vain vinorotaation faktorilatauksia, voi jäädä siihen käsitykseen, että osiot latautuvat todella selkeästi eri faktoreille. Todellisuudessa tilanne saattaa olla aivan toinen, kun faktorit korreloivat keskenään. Tuloksia pitäisi silloin tulkita niin sanotusta vinorotaation rakennematriisista. Vinorotaatioita ei ole syytä pitää tasaveroisina vaihtoehtoina suorakulmaisille rotaatioille.



Kuva 4.2. Muuttujat faktoriavaruudessa.

Faktorirakenteiden vertailu

Rotaatiota voidaan soveltaa myös käyttäen kriteerinä toista faktorimatriisia. Rotaatiossa pyritään tällöin siihen, että tulos muistuttaisi mahdollisimman paljon kohteena olevaa, esimerkiksi hypoteesin mukaista tai aiemmassa tutkimuksessa löydettyä faktorirakennetta. *Kohderotaatio* on esimerkki konfirmatorisesta faktorianalyysista.

Faktorirakenteita vertailemalla voidaan tutkia, millä tavoin eri ajankohtien, eri organisaatioiden, eri maiden tai kaikkien näiden suhteen tehdyt kyselytutkimukset eroavat rakenteellisesti toisistaan.

Edellä rajoituttiin vain vuoteen 1997, koska olisi ollut uskaliasta olettaa tilanne rakenteellisesti samanlaiseksi vuonna 2005. Mikään ei takaa, että löydettäisiin samat faktorit, vaikka kysyttäisiin samoja kysymyksiä. Yhteiskunnalliset ilmiöt muuttavat muotoaan lyhyemmässäkin ajassa.

Kahden eri faktorianalyysin, esimerkiksi vuosien 1997 ja 2005, silmämääräinen vertailu ei auta, koska kummankaan analyysin tulokset eivät ole yksikäsitteisiä. Edes sama rotaatio ei takaa vertailukelpoisuutta. Tilan säästämiseksi yksityiskohtaiset tarkastelut on tässä yhteydessä sivuutettu, mutta tehdyn kohderotaation perusteella samat ulottuvuudet voidaan tunnistaa myös vuoden 2005 aineistosta. Jos ei voitaisi, ei olisi perusteltua tehdä ajankohtien välisiä keskiarvotai muita vertailuja; tällöin vertailtaisiin keskenään eri asioita.

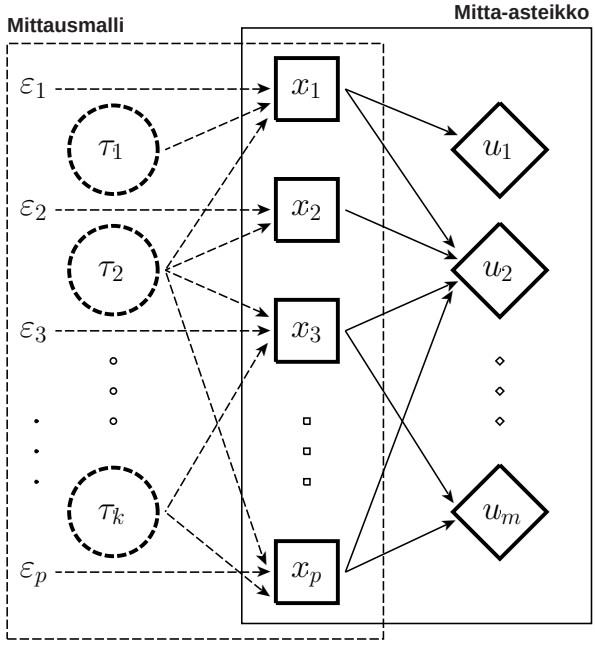
Rakennevertailuun soveltuva kohderotaatiomenetelmä tunnetaan Suomessa nimellä *transformaatioanalyysi*. Tavoitteiltaan vastaava menetelmä kantaa nimeä *Prokrustes-analyysi*. Tässä kirjassa ei syvennytä näihin menetelmiin. Asiaan perehdyttää Mustonen (1995, 95–105). Konfirmatorisesta faktorianalyysista kertoo tarkemmin muun muassa Nummenmaa ym. (1997, 263–290).

Palataan takaisin faktorianalyysiin ja sen seuraavaan vaiheeseen, jossa varsinainen aineiston tiivistäminen tapahtuu.

4.4 Mitta-asteikko

Mitta-asteikko tarkoittaa valituista muuttujista muodostettua yhdistelmää. Sillä esitetään tiiviisti asioita, joita alun perin mitataan väljästi. Tyypillisiä asteikkoja ovat muuttujien painotetut summat kuten *faktoripisteet* ja *summamuuttujat*. Mitta-asteikkojen myötä saatetaan loppuun faktorianalyysilla aloitettu aineiston tiivistäminen, jolloin tarkasteltavien muuttujien määrä tulee vähenemään merkittävästi. Samalla huomio alkaa siirtyä faktoreista ja muuttujista kohti havaintoja, joiden vertailuun ja ryhmittelyyn perehdytään seuraavissa luvuissa.

Kuva 4.3 esittää mitta-asteikkoa, jonka taustalle on asetettu aiemmin läpikäyty mittausmalli. Yhdessä ne muodostavat tärkeimmän osan tätä kirjaa hallitsevasta mittauskehikosta; loput osat tulevat esiin luvussa 5. Kuvasta nähdään, että x-muuttujat eli osiot kuuluvat sekä mittausmalliin että mitta-asteikkoon. Mallissa ne edustavat ainoita havaittavissa olevia asioita; kaikki muu on katkoviivoja ja kreikkaa. Asteikon puolella kaikki on havaittavissa, joten kirjaimet ovat latinalaisia ja kehys yhtenäinen. Asteikkoja on merkitty u-kirjaimilla. Muistisäännöistä pitävä voi mieltää ne ”uusiksi muuttujiksi”. Koska asteikot ovat osioiden yhdistelmiä, niin nuolet osoittavat osioista eteenpäin. Kaaviossa edetään siis koko ajan vasemmalta oikealle.



Kuva 4.3. Mitta-asteikko; taustalla mittausmalli.

Teoriasta käytäntöön

Asteikkojen myötä palataan takaisin havaintojen eli kyselyaineiston vastausten pariin. Havainnothan jäivät sivuosaan, kun aineistoa ryhdyttiin tutkimaan tunnuslukujen avulla, ennen kaikkea muuttujien välisillä korrelaatioilla. Faktorianalyysilla päästiin tarkastelemaan muuttujien keskinäisiä suhteita mielenkiintoisissa faktoriavaruuksissa. Ellei sieltä tulla takaisin, jäädään tavallaan pyörittelemään pelkkää teoriaa. Sitäkin voidaan tehdä, mutta tässä kirjassa halutaan päästä teoriasta taas käytäntöön.

Osioiden merkitystä ei voi vähätellä, sillä ne kytkevät teorian ja käytännön toisiinsa: ainakin karkeasti malli edustaa teoriaa ja asteikko käytäntöä. Aluksi mallia voidaan pohtia puhtaasti teoriapohjalta, mutta aineiston ja faktorianalyysin avulla teoriaa koetellaan käytännössä. Lopulta tämän työn tulokset tiivistetään mitta-asteikoiksi.

Myös mitta-asteikko on moniulotteinen. Asteikkojen lukumääränä kuvassa 4.3 esiintyy m, joka on ”mitä vain” – se ei riipu faktoreiden eikä muuttujien määristä. Tyypillisesti halutaan tehdä ainakin sellaisia asteikkoja, jotka sisällöltään vastaisivat tosiarvoja eli faktoreita, siis teoreettisia käsitteitä kuten asenteita tai arvoja, joita ei suoraan pystytä mittaamaan.

Tarvitaanko mittausmallia?

On paikallaan pohtia, tarvitaanko mittausmallia mihinkään. Jos kuvasta 4.3 häivyttää katkoviivat ja kreikkalaiset kirjaimet, jäljelle jää pelkkä mitta-asteikon kehys. Kaikki siinä on käytännössä havaittavissa: joukko muuttujia, joista voidaan tehdä erilaisia yhdistelmiä. Mihin tässä mittausmallia tarvitaan?

Mittausmallia tarvitaan, jotta voidaan arvioida, miten hyvin asteikot vastaavat ilmiön ulottuvuuksia. Jos mittausmallia ei lainkaan pohdita, ajaudutaan toimimaan kokonaan aineiston ehdoilla.

Täysin aineistolähtöistä toimintatapaa vastaa menetelmä nimeltään *pääkomponenttianalyysi*, joka historiallisista syistä johtuen sekoitetaan usein faktorianalyysiin. Menetelmien olennainen ero on, että faktorianalyysi perustuu tilastolliseen malliin, pääkomponenttianalyysi ei. Mittausmallin ja faktorianalyysin perusteella saadaan selville, miten suuri osuus muuttujien vaihtelusta johtuu mittausvirheistä. Tietoa pystytään hyödyntämään muuttujien yhdistelyssä ja siten vähentämään mittausvirheiden vaikutuksia jatkoanalyyseissa. Pääkomponenttianalyysilla ei mittausvirheistä päästä eroon eikä selville, vaan kaikki vaihtelu seuraa jatkoanalyyseihin samanarvoisena.

Käytännössä menetelmien sekoittaminen ilmenee niin, että faktorianalyysia tehdään vähän kuin pääkomponenttianalyysia: annetaan menetelmän itse ”keksiä” faktorien lukumäärä korrelaatiomatriisin niin sanottujen *ominaisarvojen* perusteella ja tyydytään siihen. Tällä tavoin ”löydetään” useimmiten aivan liian suuri määrä enemmän tai vähemmän keinotekoisia ulottuvuuksia. Ulkonäkötutkimuksen edellä käytetyistä 53 osiosta tämä lähestymistapa tuottaisi peräti 14 faktoria! Se on liikaa, minkä voisi todeta näkemättä koko aineistoa. Tutkittavan ilmiön ulottuvuuksien hahmottamista ei pidä antaa menetelmän tehtäväksi. Vaikkei toimittaisi edes kovin vankalta teoriapohjalta, pitäisi parhaan käsityksen faktorien lukumäärästä löytyä tutkijan päästä eikä jonkin matriisin ominaisarvoista.

Mittausmallin ja faktorianalyysin yhdistelmä on paras keino peilata aineiston perusteella nähtyä teorian perusteella pohdittuun. Myös seuraavaksi tarkasteltavista vaihtoehdoista luontevin tapa siirtyä mallista asteikkoon pohjautuu faktorianalyysiin. Näin ollen faktorianalyysi kattaa kaikki kuvan 4.3 (s. 107) esittämät vaiheet, joten sitä voi perustellusti pitää kyselytutkimuksen keskeisimpänä menetelmänä.

4.4.1 Faktoripisteet

Kaikki muuttujat eivät ole mittaustarkkuudeltaan yhtä hyviä, vaan joissain on enemmän mittausvirhettä kuin toisissa. Jatkoa ajatellen on tarkimmin mitatuille muuttujille syytä antaa enemmän painoarvoa ja jättää vähemmälle huomiolle muuttujat, joiden vaihtelu johtuu lähinnä mittausvirheistä.

Tavoitteeseen päästään parhaiten *faktoripisteillä*, joilla tarkoitetaan faktoreita kuvaavia havaintokohtaisia lukuja. Yksittäisinä luvut eivät ehkä puhuttele, mutta yhdessä ne kuvaavat faktoreiden edustamien ulottuvuuksien kuten asenteiden jakautumista aineistossa – tai perusjoukossa, jos aineisto on otos. Faktoripisteiden perusteella havaintoja voidaan ryhmitellä, järjestellä tai muuten tiivistetysti tarkastella tutkimuksen kannalta kiintoisista näkökulmista. Alkuperäisten muuttujien perusteella se ei onnistu yhtä hyvin.

Faktoripisteiden nimeäminen

Faktoripisteet muodostetaan laskemalla osioita yhteen ja painottamalla niitä faktorianalyysin perusteella. Painotus syntyy kokonaisuudesta, jossa huomioidaan sekä faktorilataukset että kommunaliteetit. Tarvittavat painokertoimet saadaan helposti, mutta automaattisesti ei faktoripisteitä voi tehdä, sillä uusille faktoripistemuuttujille pitää saada kunnolliset nimet.

Nimeämistä pohdittiin jo faktorimatriisin äärellä kohdassa 4.3.2 (s. 96). Ellei faktoreita pystytä nimeämään, ei faktoripisteisiin siirtymiselle ole edellytyksiä. Ohjelmistojen ”keksimät” nimet eivät kelpaa, sillä analyyseja ei voi selostaa tyyliin *”… selitettiin faktorilla 2 ja faktorilla 4 …”* tai että *”… ryhmiteltiin F3:n arvojen perusteella …”*. Kuvaavien nimien keksiminen on tulosten tulkinnan, aineiston hallinnan ja jatkoanalyysien kannalta olennaista, ja se on osa dokumentointia, jota käsitellään liitteessä A.

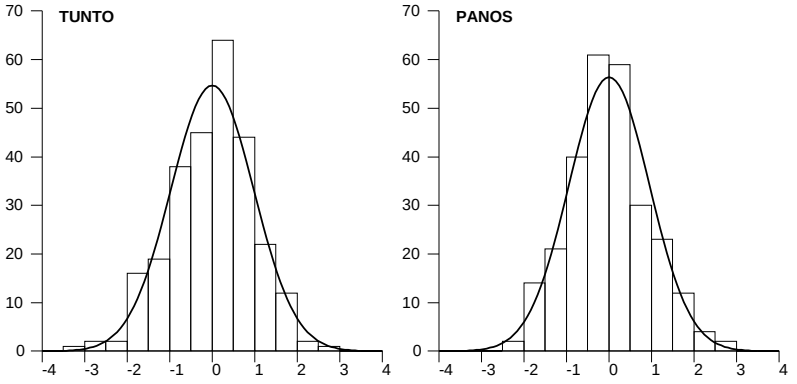
Edellä muodostetuista faktoripistemuuttujista tullaan jatkossa käyttämään taulukossa 4.1 esiintyviä nimiä ja kuvauksia.

Nimi	Kuvaus
TUNTO	Itsetunto ulkonäköasioissa
PANOS	Panostaminen ulkonäköön
PAINE	Sosiaaliset ulkonäköpaineet
NEGAT	Negatiivinen suhtautuminen

Taulukko 4.1. Faktoripistemuuttujien nimet ja kuvaukset.

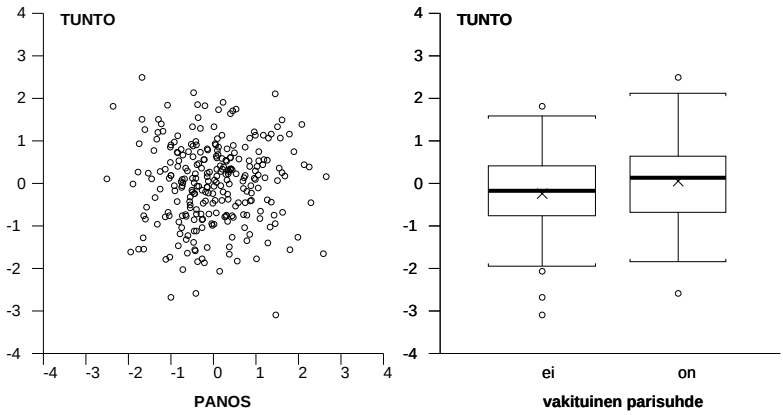
Faktoripisteiden ominaisuuksia

Kuvassa 4.4 on kahden ensimmäisen faktoripistemuuttujan jakaumia kuvattu histogrammeilla, joihin on liitetty myös normaalijakauman kuvaajat. Vaikka alkuperäisten osioiden jakaumat olisivat suunnilleen millaisia tahansa, niin niiden kaikenlaiset summat lähestyvät nopeasti normaalijakaumaa. Tämä tilastotieteen teoriassa jo varhain osoitettu tulos on käytännössäkin tärkeä ja kuvastaa normaalijakauman keskeistä asemaa tilastollisessa työskentelyssä.



Kuva 4.4. Kahden faktoripistemuuttujan histogrammit.

Monissa jatkoanalyyseissa oletetaan muuttujien noudattavan normaalijakaumaa, joten tässä suhteessa faktoripisteet ovat otollisia. Toinen jatkoanalyysien kannalta mukava ominaisuus on se, että faktoripistemuuttujat eivät korreloi keskenään, jos faktorianalyysissa on käytetty suorakulmaista rotaatiota (ks. kohta 4.3.3, s. 100). Tällöin esimerkiksi regressioanalyysin tekeminen ja tulkinta on selkeämpää. Kuvassa 4.5 itsetunnon ja ulkonäköön panostamisen hajontakuva osoittaa, ettei kyseisten faktoripisteiden välillä ilmene riippuvuutta.



Kuva 4.5. Faktoripisteet hajonta- ja laatikkokuvana.

Faktoripisteet on skaalattu siten, että niiden keskiarvot ovat nollia, jolloin positiiviset ja negatiiviset arvot kuvastavat ulottuvuuksien eri suuntia. Keskiarvoeroja ilmenee kuitenkin tarkasteltaessa faktoripisteitä muiden muuttujien, esimerkiksi taustamuuttujien, suhteen. Kuvan 4.5 laatikkokuva vertailee itsetuntoa parisuhteen vakituisuuden suhteen. Vakituisessa parisuhteessa elävien itsetunto ulkonäköasioissa näyttäisi kuvan perusteella jonkin verran paremmalta.

4.4.2 Summamuuttujat

Toinen tapa muodostaa mitta-asteikkoja on laskea vain valitut osiot yhteen ja unohtaa muut. Tällaisia asteikkoja on tapana kutsua *summamuuttujiksi*. Usein summamuuttuja skaalataan vielä jakamalla se summauksessa käytettyjen muuttujien lukumäärällä, jolloin saatu asteikko vastaa yleensä alkuperäisten osioiden keskiarvoa.

Seuraavassa esitettyjen näkemysten hieman kriittinen sävy selittyy sillä, että tilastotieteilijän näkökulmasta faktoripisteet ovat monella muotoa parempi tapa tiivistää aineistoa kuin summamuuttujat. Osittain erilaisia näkemyksiä esittää muun muassa Alkula ym. (1994, 100–103, 277–278).

Perusteluja summamuuttujien käytölle

Summamuuttujia perustellaan yleensä tulkinnalla: on helpompi ymmärtää, kun mitta-asteikon luvut vaihtelevat samalla välillä kuin alkuperäiset osiotkin. Tässä kuitenkin oletetaan, että mittaukset on alun perin tehty samanlaisilla, tyypillisesti viisiportaisilla osioilla. Jos osiot ovat erilaisia, tulkinta ontuu. Osioista muodostettujen mitta-asteikkojen ei kuitenkaan tarvitse muistuttaa lukuarvoiltaan alkuperäisiä osioita. Pikemminkin olisi parempi, ettei liikaa jututtaisi ajattelemaan osioita. Nehän ovat vain mittausvälineitä, joista pitäisi analyysien myötä päästä ylemmän tason muuttujiin, mitta-asteikkoihin.

Toinen peruste summamuuttujille tulee ilmiön teoriasta, joka saattaa sanella, mitä osioita pitää painottaa ja millä tavalla. Periaatteessa tällaisessa tilanteessa ei koko faktorianalyysille ole käyttöä. Ennen summamuuttujien muodostamista olisi kuitenkin hyvä edes tarkistaa faktorianalyysilla, vastaako aineiston perusteella havaittu rakenne sitä, mitä teoria väittää. Jos vain ”sokeasti” lasketaan mitattuja osioita yhteen, lopputulos voi olla mitä sattuu, vaikka mittari olisi kuinka ”validoitu” tahansa. Kansainvälisesti tunnettujen mittareidenkaan toimivuus ei ole itsestäänselvyys, sillä kielelliset ja kulttuuriset erot aiheuttavat helposti yllätyksiä.

Faktorianalyysin käyttö pelkkänä summamuuttujien muodostamisen perusteena ei myöskään ole ongelmatonta, koska osioita valitessa huomio kiinnittyy useimmiten pelkkiin faktorilatauksiin. Faktoripisteissä huomioidaan sekä lataukset että kommunaliteetit, toisin sanoen koko moniulotteinen faktorirakenne.

Esimerkki summamuuttujien muodostamisesta

Tarkastellaan asiaa esimerkin avulla tekemällä summamuuttujia vuoden 1997 faktorirakenteen pohjalta (ks. tuloste 4.2, s. 101). Otetaan mukaan vain osiot, joilla on vähintään 0.6:n suuruinen lataus jollakin faktorilla. Tämä on täysin mielivaltainen päätössääntö ja perustuu vain tässä analyysissa saatuihin faktorilatauksiin, joita on tulosteessa korostettu. Neljäs faktori on parasta unohtaa kokonaan, koska sitä jäisi tällä säännöllä kuvaamaan ainoastaan yksi osio. Samalla on käännettävä osion `k26.18` suunta, koska sen suurin lataus on negatiivinen. Faktoripisteissä mahdolliset eri suunnat eli latausten etumerkit otetaan huomioon automaattisesti; summamuuttujissa niistä pitää itse huolehtia.

Tuloste 4.3. Faktoripisteiden ja summamuuttujien tunnuslukuja.

```
Means, std.devs and correlations of UNSUMMAT  N=273
# of missing observations =5
Variable Mean Stddev        TUNTO tunto PANOS panos PAINE paine
  TUNTO  0.00  0.96  TUNTO  1.00  0.97  0.01  0.06 -0.03 -0.02
  tunto  3.26  0.77  tunto  0.97  1.00  0.07  0.11 -0.13 -0.13
  PANOS  0.00  0.94  PANOS  0.01  0.07  1.00  0.92  0.02 -0.01
  panos  2.67  0.85  panos  0.06  0.11  0.92  1.00  0.06  0.05
  PAINE  0.00  0.92  PAINE -0.03 -0.13  0.02  0.06  1.00  0.97
  paine  4.09  0.79  paine -0.02 -0.13 -0.01  0.05  0.97  1.00
```

Näin saadaan kolme summamuuttujaa, joihin tiivistyy kahdeksan, viiden ja neljän, yhteensä 17 muuttujan tiedot. Uudet muuttujat on nimetty muuten samoin kuin faktoripisteet, mutta pienillä kirjaimilla. Tulosteesta 4.3 nähdään, että sisällölliset vastaavuudet ovat yhtä hyvät kuin ”nimelliset”: muuttujat korreloivat lähes täydellisesti.

Kaikkiaan 36 osiota jää kuitenkin tällä tavoin kokonaan tarkastelujen ulkopuolelle. Miksi suotta vaivata vastaajia turhilla kysymyksillä, jos niillä ei analyyseissa tehdä mitään? Vastauksetkin voisivat olla luotettavampia, jos kysymyksiä olisi vähemmän. Tarkemmin mietittynä asiaan kätkeytyy kuitenkin potentiaalinen ongelma: Entä jos tutkimus toistetaan ja siinä nyt valitut 17 osiota eivät toimikaan niin hyvin kuin tässä? Olisivatko pois jätetyt osiot toimineet sittenkin paremmin?

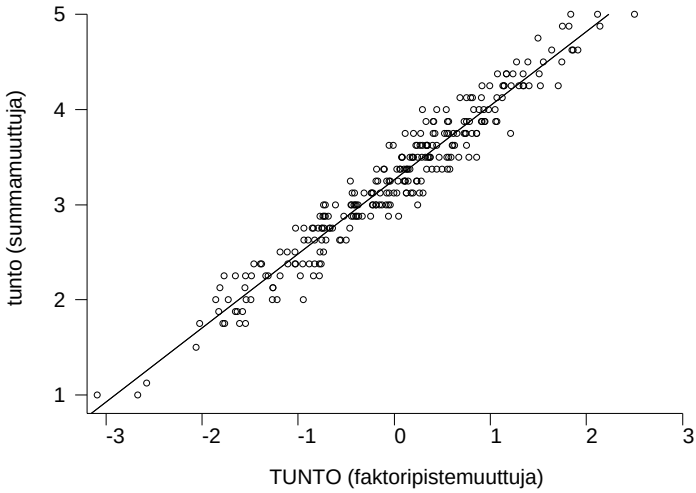
Vuosien 1997 ja 2005 faktorirakenteiden vertailu, jonka yksityiskohdat edellä sivuutettiin, todistaa rakenteiden vastaavan hyvin toisiaan. Faktoreille voi siis antaa samat nimet, ja vastaavat faktoripisteet voidaan muodostaa myös vuoden 2005 havainnoille. Summamuuttujissa ilmenee kuitenkin eroja, kun sovelletaan samaa, vähintään 0.6:n suuruisten latausten päätössääntöä kuin edellä. Selvin yksittäinen poikkeama on ulkonäkövaatimuksia luotaavan osion `k71.6` (*"Ulkonäkövaatimukset naisille ovat kovia"*) faktorilataus, joka vuonna 2005 on noussut koko analyysin korkeimmaksi. Tällä kiinnostavalla yksityiskohdalla voi hyvin olla sisällöllistä merkitystä, mutta "tasapaksusti" painotetut summamuuttujat eivät ota sitä lainkaan huomioon, toisin kuin faktoripisteet.

Kysymys ei siis ole pelkästään uusien muuttujien muodostamisesta, vaan jo kohdassa 4.3.3 (s. 100) pohditusta rakennevaliditeetista – tarkemmin sanottuna sen puutteesta. Tiedot jotka kerran on kerätty, pitäisi pyrkiä hyödyntämään, eikä heittää niitä menemään mielivaltaisin perustein.

Mittareita pitää kehittää, mutta sitä voidaan tehdä vasta, kun tutkimus toistetaan ja mittauksia pohditaan uudelleen. Tällöin on aika hyödyntää aiemman tutkimuksen faktorianalyyseja, jättää huonoimpia osioita pois ja korvata niitä mahdollisesti paremmilla. Rakennevaliditeetin kannalta ei tunnu perustellulta ratkaisulta, että kesken analyysin kelpuutetaan jatkoon vain parhailta vaikuttavat osiot. Faktoripisteissä käytetään kaikkia osioita painottamalla niitä mittaustarkkuuden mukaisesti.

Summamuuttujien ja faktoripisteiden vertailua

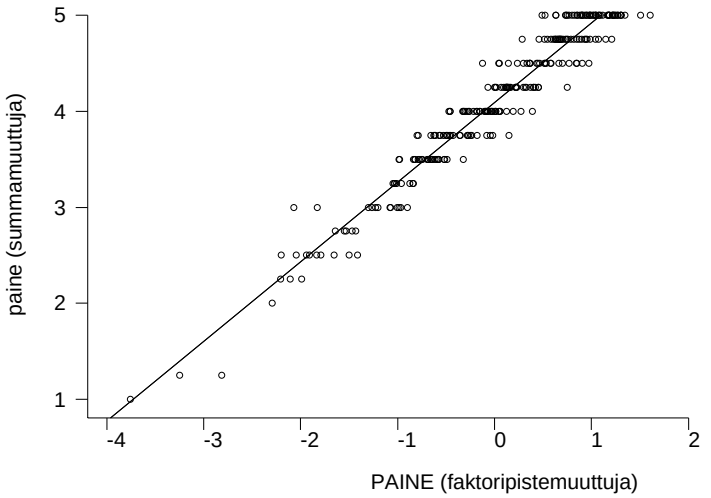
Summamuuttujien käyttöön houkuttelevat siis tulkinnan näennäinen helppous sekä ajateltu vastaavuus alkuperäisiin osioihin. Tulosteessa 4.3 (s. 113) korrelaatiot viittasivat summamuuttujien ja faktoripisteiden yhteneväisyyteen, ja samaa todistavat kaksi niiden välistä hajontakuvaa, kuvat 4.6 (s. 115) ja 4.7 (s. 116). Hajontakuvista ja tulosteesta 4.3 havaitaan myös, että summamuuttujien ja faktoripisteiden välillä on eroja sekä keskiarvoissa että hajonnoissa.



Kuva 4.6. Itsetunto faktoripiste- ja summamuuttujina.

Faktoripisteiden keskiarvot ovat nollia, mutta summamuuttujien keskiarvot määräytyvät osioiden perusteella. Niitä voi tulkita samalla tavalla kuin osioidenkin keskiarvoja, kunhan kaikki osiot ovat samalla tavalla mitattuja. Mikäli samassa yhteydessä esiintyisi muunkin tyyppisiä kuin viisiportaisia osioita, ei näin lasketuissa summamuuttujien keskiarvoissa olisi mitään mieltä. Faktoripisteet toimivat, vaikka alkuperäiset osiot olisivat erilaisia. Osioiden keskiarvoihin on yhtä turha jumittua kuin osioihin muutenkaan.

Toinen summamuuttujien ja faktoripisteiden ero koskee niiden hajontoja ja jatkuvuutta. Faktoripisteiden hajonnat ovat yleensä suurempia, kuten tässäkin, joten niihin sisältyy enemmän jatkoanalyyseissa tarvittavaa tilastollista informaatiota. Summamuuttujat ovat sidottuja osioiden vaihteluvälille 1–5. Ne eivät myöskään ole niin jatkuvia kuin faktoripisteet, vaan osioiden diskreetti luonne ”paistaa läpi” kuten kuvassa 4.7. Pisteet muodostavat vaakasuoria pisteviivoja summamuuttujan diskreettien arvojen tasoille.



Kuva 4.7. Ulkonäköpaineet faktoripiste- ja summamuuttujina.

Todellisuudessa summamuuttujien ja faktoripisteiden erot voivat olla selvästi suurempia kuin tässä esitetyt. Seuraavassa eroja tarkastellaan vielä asteikkojen mittaustarkkuuden kannalta.

4.4.3 Mitta-asteikon reliabiliteetti

Aineiston tiivistäminen koostuu monesta eri vaiheesta. joiden luotettavuuden arviointi on jatkon kannalta olennaista. Validiteettikysymykset ovat paljolti sisällöllisiä. Mikäli niihin saadaan hyväksyttäviä vastauksia, kiinnostaa myös *reliabiliteetti* eli mittauksen tarkkuus. Reliabiliteetti on tärkeä, mutta validiteettiin verrattuna vasta toissijainen peruste mittauksen luotettavuudelle.

Reliabiliteettia arvioidaan tilastollisesti tutkimalla mittauksen vaihtelun määrää ja laatua. Täsmällisemmin määriteltynä reliabiliteetti ilmaisee mittauksen *todellisen vaihtelun osuuden*. Jäännösosuus aiheutuu satunnaisesti vaihtelevasta mittausvirheestä.

Mittauksen todelliseen vaihteluun kiinni pääseminen edellyttää mittauksen mallintamista, sillä satunnaisia mittausvirheitä ei voi suoraan havaita. Kohdassa 4.2 (s. 91) esitetty mittausmalli luo perus-

tan reliabiliteetin arvioinnille, mutta itse reliabiliteetti on vain mitta-asteikon ominaisuus. Niinpä kuvan 4.3 (s. 107) havainnollistamilla, mitta-asteikon kehykseen sisältyvillä x- ja u-muuttujilla on reliabiliteetti, joka määräytyy siitä, missä suhteessa tosiarvot ja mittausvirheet vaikuttavat niiden vaihteluun. Tosiarvoilla ja mittausvirheillä ei sen sijaan ole reliabiliteettia, koska ne kuuluvat vain mittausmalliin.

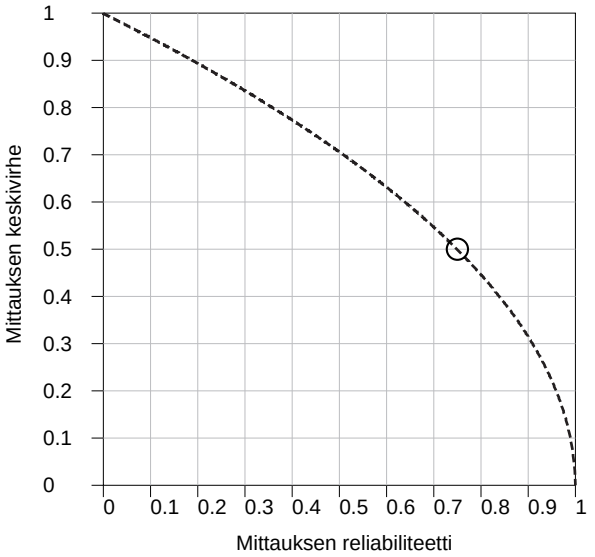
Mittauksen keskivirhe

Reliabiliteetti raportoidaan usein rutiininomaisesti lukuina, jotka eivät anna mittaustarkkuudesta selvää käsitystä. Luvut voivat olla avuksi vertailtaessa eri tutkimusten reliabiliteetteja, mutta käytännön mittaustarkkuus on parempi ilmaista *mittauksen keskivirheenä*.

Mittauksen keskivirheen tulkinta perustuu tilastollisen päättelyn periaatteeseen, jossa *kahden keskivirheen eroa* voidaan pitää tilastollisesti merkitsevänä ja siten vähän painavampana perusteena (vrt. kohta 3.2.2, s. 54). Vastaavasti voidaan muodostaa myös mittaustarkkuuteen perustuvia *luottamusvälejä* ja verrata esimerkiksi, menevätkö kahden ryhmän mittausten luottamusvälit päällekkäin. Tulkintojen taustalla on oletus normaalijakaumasta, joka yleensä pätee varsin hyvin useammista osioista muodostetuille mitta-asteikoille, kuten edelläkin havaittiin (ks. kuva 4.4, s. 110). Yksittäisten osioiden reliabiliteetit eivät enää tässä vaiheessa kiinnosta; niiden tarkastelu kuuluu mittausmallin rakennevaliditeetin arviointiin.

Jos raportoidaan vain, että ”*reliabiliteetti on 0.75*”, voidaan päätellä, että neljännes mitta-asteikon vaihtelusta johtuu mittausvirheestä. Mittaustarkkuudesta saadaan parempi käsitys, kun todetaan, että ”*mittauksen keskivirhe on 0.5*”. Tällöin on perusteltua puhua noin ykkösen suuruisista eroista. Pienemmät erot mahtuvat virhemarginaaliin, joten niistä ei kannata tehdä numeroa.

Mittauksen keskivirhe tuo reliabiliteettitarkastelut suhdeluvuista käytännöllisemmälle, asteikon vaihtelun tasolle. Kuva 4.8 näyttää reliabiliteetin ja mittauksen keskivirheen yhteyden, kun hajonta on ykkösen suuruista. Edellä mainittu esimerkki luvuista 0.75 ja 0.5 on ympyröity. Kuva osoittaa selvästi, että korkeampi reliabiliteetti vastaa pienempää mittauksen keskivirhettä. Yhteyden epälineaarisuuden vuoksi mittauksen keskivirhe pienenee yhä jyrkemmin, kun reliabiliteetti lähestyy ykköstä. On siis eroa, onko reliabiliteetti 0.9 vai 0.95.



Kuva 4.8. Kaavio mittauksen reliabiliteetista ja keskivirheestä.

Ulkonäkötutkimuksen mitta-asteikkojen tarkkuus

Tarkastellaan esimerkkinä ulkonäkötutkimuksen mitta-asteikkojen tarkkuutta. Laskelmat on tehty mittauskehikkoon sisältyvällä reliabiliteetin arviointimenetelmällä, joka huomioi mittausmallin ja mitta-asteikon moniulotteisuuden.

Taulukosta 4.2 nähdään, että faktoripisteiden reliabiliteetit ovat parhaimmillaan 0.92 ja huonoimmillaan 0.82. Hajonnat ovat vähän alle ykkösen, kuten jo aiemmin todettiin. Mittauksen puolesta tarkin asteikko on ”itsetunto ulkonäköasioissa” ja karkein on odotetusti tulkinnaltaan vaikein ”negatiivinen suhtautuminen”. Mittauksen keskivirheiden perusteella tarkimmalla asteikolla voidaan puhua noin 0.5:n kokoisista eroista.

Mittaustarkkuus on syytä ottaa huomioon esimerkiksi muuttujien luokittelussa. Aiemmin piirretyissä histogrammeissa (ks. kuva 4.4, s. 110) itsetuntoasteikon luokittelu vastaa hyvin sen mittaustarkkuutta. Kauimmaisia luokkia tosin jouduttaisiin yhdistämään vähäisten havaintomäärien vuoksi, mutta nähtävästi olisi mahdollista muodos-

Faktoripistemuuttuja	Hajonta	Reliabiliteetti	Mittauksen keskivirhe
Itsetunto ulkonäköasioissa	0.96	0.92	0.26
Panostaminen ulkonäköön	0.94	0.89	0.31
Sosiaaliset ulkonäköpaineet	0.92	0.85	0.35
Negatiivinen suhtautuminen	0.91	0.82	0.38

Taulukko 4.2. Faktoripisteiden vaihtelu ja mittaustarkkuus.

taa jopa 7–9 luokkaa, joiden väliset erot eivät johtuisi ainakaan kokonaan mittausvirheistä. Luottamusvälitulkinnoissa huomioon otettaisiin 0.5:n pituinen vaihtelu molempiin suuntiin, joten tyypillisen 95 %:n luottamusvälin pituus olisi noin yhden yksikön mittainen. Tiiviimpi luokitus, esimerkiksi neljä luokkaa, kuvastaisi jo varsin luotettavasti todellisia eroja vastaajien itsetunnon suhteen. Tähän palataan luvun 7 lopussa.

Summamuuttuja	Hajonta	Reliabiliteetti	Mittauksen keskivirhe
Itsetunto ulkonäköasioissa	0.77	0.83	0.32
Panostaminen ulkonäköön	0.86	0.79	0.39
Sosiaaliset ulkonäköpaineet	0.80	0.82	0.34

Taulukko 4.3. Summamuuttujien vaihtelu ja mittaustarkkuus.

Taulukosta 4.3 nähdään, että kolmen summamuuttujan mittaustarkkuudet eivät ole yhtä hyviä kuin faktoripisteiden. Kuten aiemmin tuotiin esille, niiden hajonnat ovat pienempiä, toisin sanoen informaatiota on vähemmän. Kun reliabiliteetitkin ovat heikompia kuin faktoripisteiden, on mittauksen keskivirhe suurempi, etenkin suhteessa vastaavaan hajontaan.

Pohdintaa reliabiliteetista

Reliabiliteetin arviointiin on aikojen saatossa kehitetty lukuisia menetelmiä. Pohjimmiltaan kyse on siitä, miten paljon ja millaisia oletuksia tehdään mittausvirheistä. Varhaisissa menetelmissä jouduttiin tekemään kohtuuttoman paljon oletuksia, koska vähemmillä oletuksilla oltaisiin jouduttu laskemaan enemmän. Laskeminen oli vaikeaa ja vei aikaa, joten sitä haluttiin välttää. Enää laskeminen ei ole ongelma, joten voidaan soveltaa realistisempia menetelmiä.

Valitettavasti reliabiliteetin arvioinnissa tyydytään useimmiten *Cronbachin alfa* -nimisen, 1930-luvulta periytyvän laskukaavan rutiininomaiseen käyttöön. Kaavaa sovelletaan kritiikittömästi tiedostamatta sen taustalle hautautuneita oletuksia. Psykologian alan vuosikatsauksessa todettiin jo 1980-luvulla, että reliabiliteetin arvioinnin tarkoitus on jossain vaiheessa hämärtynyt (Weiss & Davison, 1981).

Tavaksi on muun muassa tullut jättää osioita pois summamuuttujasta, jotta saadaan sen alfa-arvo korkeammaksi. Epärealistisista oletuksista johtuen alfan kaava ei siedä sitä itsestäänselvyyttä, että osioiden mittaustarkkuus vaihtelee. Tällainen reliabiliteetin maksimointi aiheuttaa vakavia seurauksia, sillä se huonontaa validiteettia. Osioiden karsintaa mielivaltaisten alfa-rajojen avulla on vaikea perustella sisällöllisesti kestävällä tavalla.

Kehittyneemmän menetelmän reliabiliteetin arviointiin osana mittauskehikkoa on esittänyt Lauri Tarkkonen väitöskirjassaan (Tarkkonen, 1987). Menetelmän tunnetuksi tekeminen on haasteellista, mutta toisaalta sen soveltaminen on palkitsevaa. Tarkkosen menetelmällä reliabiliteettia voidaan arvioida monipuolisemmin eikä osioita tarvitse keinotekoisesti jättää pois. Menetelmällä saadut arviot ovat myös korkeampia kuin Cronbachin alfat niissä tilanteissa, joissa menetelmiä on perusteltua vertailla (Vehkalahti ym., 2008; Vehkalahti, 2000).

5 Havaintojen vertailu

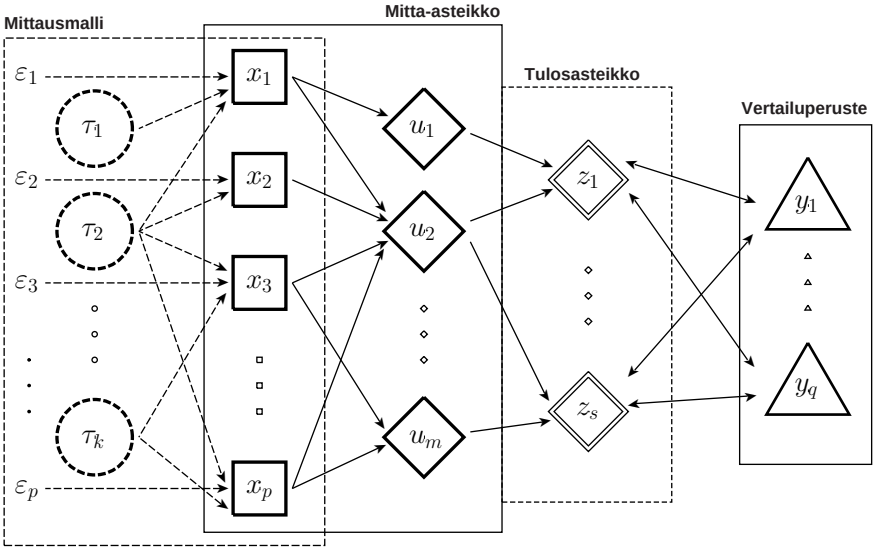
Kun aineistoon on tutustuttu ja sitä on muokattu ja tiivistetty, huomio kääntyy takaisin yksittäisiin havaintoihin. Tällöin päästään yleensä kiinni tutkimuksen kannalta mielenkiintoisimpiin kysymyksiin. Jotta kysymyksiin voidaan vastata, tarvitaan monenlaista työstämistä: *järjestelyä*, *ryhmittelyä*, *selittämistä*, *sovittamista*, *valikointia*, *erottelua*, *luokittelua* ja *ennustamista*. Näitä toimintoja yhdistää havaintojen tai havaintoryhmien vertailu.

Tässä luvussa havaintoja vertaillaan yleisimmin käytetyllä tilastollisella menetelmällä, regressioanalyysilla. Menetelmää tarkastellaan mittauskehikossa, joka nyt esitetään kokonaisuudessaan.

5.1 Mittauskehikko

Lyhyesti sanottuna mittauskehikko kytkee toisiinsa mittarit ja menetelmät. Pidemmin ilmaistuna se yhdistää tutkimuskysymyksen, ilmiön ulottuvuudet, niitä mittaavat osiot, mittausvirheet, mitta-asteikot ja näihin liittyvät tilastolliset menetelmät. Kehikon ydinosat, mittausmalli ja mitta-asteikko, esiteltiin luvussa 4. Kokonaisuuden, joka näkyy kuvassa 5.1, täydentävät *tulosasteikko* ja *vertailuperuste*.

Mittauksen haasteet alkavat tutkimuksen suunnittelusta ja seuraavat läpi analyysien. Koska mittausta ei voi parantaa jälkikäteen millään menetelmällä, on tärkeää laatia mittausmalli, jonka pohjalta asteikot luodaan. Mittauskehikon kaksi viimeistä osaa, tulosasteikko ja vertailuperuste, täsmentävät, miten asteikkoja käytetään tilastollisissa analyyseissa.



Kuva 5.1. Mittauskehikko kokonaisuudessaan.

Seuraavassa vertailuperustetta ja tulosasteikkoa kuvaillaan yleisellä tasolla. Konkreettisempi käsitys niistä muodostuu regressioanalyysin yhteydessä, kohdassa 5.2 (s. 124).

5.1.1 Vertailuperuste

Vertailuperuste muodostaa pohjan havaintojen vertailulle. Se määritellään yleensä erillään mallista ja asteikoista. Tästä syystä sen kehyskin kuvassa 5.1 on erillään kehikon muista osista. Havaintojen vertailun perusteet ovat sisällöllisiä ja tulevat usein ilmiön teoriasta tai aiemmasta tutkimuksesta. Vertailu voi perustua esimerkiksi kansallisesti tai kansainvälisesti määriteltyihin kriteereihin. Toisinaan tutkijan pitää myös löytää uusia vertailuperusteita. Aineistolähtöisemmät vertailut perustuvat yksinkertaisesti aineiston muihin muuttujiin.

Kohdassa 5.2 perehdytään regressioanalyysiin, jossa vertailuperuste on regressiomallin selitettävä muuttuja, mutta mittauskehikon muiden osien tapaan myös vertailuperuste voi olla moniulotteinen.

Kuvassa 5.1 vertailuperustetta merkitään kirjaimella y, joka on tilastollisissa malleissa perinteinen selitettävän muuttujan tunnus. Vastaavasti perinteinen selittävän muuttujan symboli on x. Mittauskehikossa x edustaa vain mittauksia, joista lähdetään liikkeelle. Varsinaiset analyysit tehdään pääosin mitta-asteikoilla.

Kokonaisuudessaan kuva 5.1 esittää vain yhteen mittausmalliin pohjautuvaa asetelmaa. Tässä kirjassa riittää yksi mittausmalli, mutta tyypillisesti kyselyaineistojen tiivistämiseen tarvitaan useampia. Vertailuperuste voi silloin olla jonkin toisen mittausmallin pohjalta luotu mitta-asteikko, jolloin päästään analysoimaan samanaikaisesti useampia moniulotteisia ilmiöitä. Tässä suhteessa mittauskehikko muistuttaa *rakenneyhtälömalleja*, joihin perehdyttää muun muassa Nummenmaa ym. (1997, 302–375). Lähestymistavoilla on kuitenkin olennainen ero: rakenneyhtälömalleissa huomio kohdistuu mittausmalleihin, mittauskehikossa mitta-asteikkoihin.

5.1.2 Tulosasteikko

Tulosasteikko, joka kuvassa 5.1 sijoittuu mitta-asteikon ja vertailuperusteen väliin, tiivistää havaintojen vertailun tulokset asettamalla havainnot vertailuperusteen mukaiseen järjestykseen. Tulosasteikko voi kuvata esimerkiksi havainnoista muodostettua ryhmitystä, luokitusta tai ennustetta, kuten kohdassa 5.2 käsiteltävässä regressioanalyysissa. Tulosasteikkoja synnyttävät muutkin tilastolliset menetelmät, joihin perehdytään kirjan viimeisissä luvuissa.

Kuvassa 5.1 vasemmalta oikealle osoittavat nuolet näyttävät, että z-kirjaimella merkityt tulosasteikot ovat u:lla merkittyjen ”ensimmäisen tason” mitta-asteikkojen yhdistelmiä. Tulosasteikkoja kutsutaankin tästä syystä myös *toisen tason mitta-asteikoiksi*. Niidenkin perustana ovat alkuperäiset osiot, mutta suodatettuina mittausmallin ja mitta-asteikon läpi.

Vertailuperusteen ja tulosasteikon väliset kaksisuuntaiset nuolet kuvassa 5.1 viittaavat havaintojen vertailussa tarkasteltaviin yhteyksiin. Yhteyksien tarkempi luonne ja nuolien suunta määräytyvät kysymyksenasetteluista; yleensä ne eivät ole tilastotieteen asioita.

5.2 Regressioanalyysi

Regressioanalyysi on perinteinen, faktorianalyysiakin vanhempi tilastollinen menetelmä, jonka avulla voidaan monin tavoin analysoida havaintoja, laatia selitysmalleja sekä tehdä ennusteita ja vertailuja. Analyysista on erilaisia muunnelmia, mutta tässä kirjassa pitäydytään perusmuodossa, *lineaarisessa regressioanalyysissa*. Analysoitavat riippuvuudet oletetaan lineaarisiksi, joten tarkastelut perustuvat korrelaatioihin kuten faktorianalyysissakin.

Faktorianalyysin tapaan regressioanalyysi pohjautuu tilastolliseen malliin, jonka parametrit estimoidaan aineistosta. Mallia kutsutaan *lineaariseksi regressiomalliksi*, lyhyemmin regressiomalliksi, toisinaan vain lineaariseksi malliksi. Luvun 4 alussa läpikäydyt, tilastollista mallia koskevat periaatteet pätevät myös regressiomalliin.

Koska regressioanalyysi sisältyy lähes kaikkiin menetelmäkirjoihin ja moniin tilastotieteen perusteiden oppimateriaaleihin, sitä ei tässä yhteydessä käsitellä yksityiskohtaisesti. Menetelmän perusteisiin johdattaa esimerkiksi Alkula ym. (1994, 244–257), Nummenmaa ym. (1997, 307–324) sekä Nummenmaa (2004, 297–317).

Seuraavassa perehdytään regressiomallin oletuksiin sekä katsotaan, miten menetelmä asettuu mittauskehikkoon. Regressioanalyysin tärkeimmät vaiheet käydään läpi esimerkkien avulla, minkä jälkeen oletusten paikkansapitävyyttä arvioidaan tutustumalla lyhyesti *regressiodiagnostiikkaan*.

5.2.1 Oletukset

Regressiomallin oletuksia voidaan lähteä purkamaan vertailemalla sitä edellisessä luvussa käsiteltyyn mittausmalliin. Molemmat ovat esimerkkejä tilastollisista malleista, joten niillä on jo sen vuoksi paljon yhteistä. Malleilla on kuitenkin kaksi selvää eroa.

Ensimmäinen ero on näkyvämpi ja liittyy muuttujien asemaan mallissa. Mittausmallissa kaikki muuttujat ovat samassa asemassa mitaten taustalla olevia tosiarvoja, mutta regressiomallin muuttujista yksi on *selitettävä muuttuja* ja muut *selittäjiä*. Muuttujille on myös muita nimityksiä, esimerkiksi selitettävää muuttujaa voidaan kutsua y-muuttujaksi, tai kuten jatkossa lyhyesti ilmaistaan, *vastemuuttujaksi* tai pelkästään *vasteeksi*.

Toinen ero on näkymättömämpi ja liittyy mallin satunnaisvaihteluun. Mittausmallissa satunnaisvaihtelun oletetaan johtuvan mittausvirheestä, regressiomallissa otantavirheestä. Tilastollisen tutkimuksen keskeiset epävarmuuden lähteet, mittaus ja tiedonkeruu (ks. luku 2), vaikuttavat siis selvästi taustalla, mutta malleissa niihin otetaan kantaa valikoidusti, toisiaan täydentäen. Mittausmallissa mittausvirheitä oletetaan sisältyvän kaikkiin muuttujiin, kun taas regressiomallissa oletus otantavirheestä liittyy ainoastaan vastemuuttujaan. Mittausvirheen ja otantavirheen yhtaikainen huomiointi samassa mallissa johtaa monimutkaisiin asetelmiin, joita ei tässä kirjassa käsitellä.

Regressiomalli mittauskehikossa

Mittauskehikossa edetään vaihe kerrallaan. Ensin huolehditaan mittausepävarmuuksista mittausmallin ja mitta-asteikon avulla. Seuraavassa vaiheessa otetaan kantaa muihin mahdollisiin epävarmuuksiin. Mittauskehikon käsitteistössä vastemuuttuja on vertailuperuste ja selittäjät ovat asteikkoja. Analyysi synnyttää tulosasteikon, jota regressiomallissa kutsutaan *sovitteeksi* tai *ennusteeksi*. Sovite on selittäjistä muodostettu yhdistelmä, jonka painokertoimet analyysi määrää siten, että tulos muistuttaa mahdollisimman paljon vastemuuttujaa. Ennustesanaa saatetaan käyttää sovitteen synonyyminä, vaikkei aineistoon sisältyisi ajallisia tarkasteluja.

Mittauskehikko osoittaa, kuinka sekä mallit että menetelmät muistuttavat monessa suhteessa toisiaan. Niin faktori- kuin regressioanalyysissa tehdään laaditun mallin perusteella uusia asteikkoja, joiden painokertoimet heijastavat analyysien tavoitteita. Faktorianalyysissa tavoitteena on kuvata taustalla olevia ulottuvuuksia, regressioanalyysissa mielenkiinnon kohteena olevaa vastetta.

Muuttujia koskevat oletukset

Regressiomalli muistuttaa mittausmallia myös siinä suhteessa, että sen taustalla on oletus normaalijakaumasta. Se koskee vain vastemuuttujaa, joten tavallinen yksiulotteinen normaalijakauma riittää. Perusmuodossaan regressioanalyysi soveltuu vain sellaisten ilmiöiden mallintamiseen, joissa vastemuuttujan jakauma on normaalijakauman tyyppinen, siis kohtalaisen symmetrinen ja yksihuippuinen.

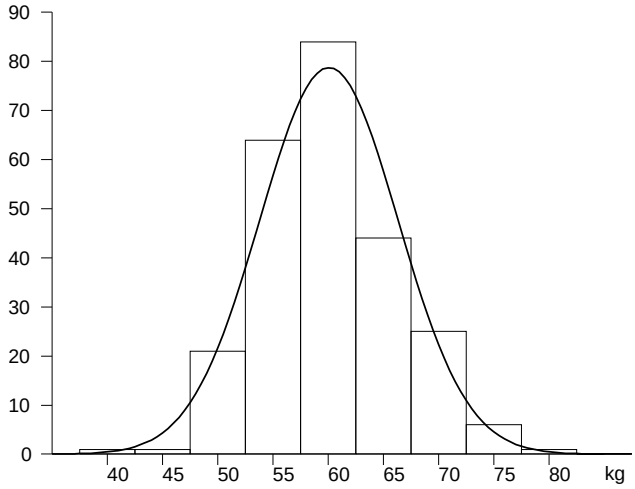
Ellei vastemuuttujan normaalisuusoletus edes suurin piirtein päde, niin regressiomalli ei toimi oikein. Mallintamisen vaiheisiin sisältyy tilastollisia testejä, joiden luotettavuus on paljolti normaalisuusoletuksen varassa. Mikäli oletus ei lainkaan päde, voi olla parempi siirtyä *yleistettyihin lineaarisiin malleihin*, joissa vasteen todennäköisyysjakaumaksi voidaan valita jokin muu kuin normaalijakauma. Näitä malleja ei tässä kirjassa käsitellä, mutta mallintamisen periaatteet ovat niissä hyvin samanlaiset. Johdattelua mallien perusteisiin sisältyy esimerkiksi teoksiin Alkula ym. (1994) ja Nummenmaa ym. (1997). Syvällisemmin niihin perehdyttää muun muassa kirja *Regression with Social Data: Modeling Continuous and Limited Response Variables* (DeMaris, 2004).

Toisinaan vastemuuttuja voidaan muuntaa paremmin normaaliseksi esimerkiksi *logaritmoimalla*. Kyseinen muunnos on yleisesti käytetty monissa yhteiskunnallisissa tutkimuksissa, sillä sen avulla päästään lukumäärinä mitatuista muuttujista monesti kiinnostavampiin suhteellisiin osuuksiin. Usein muuttujan jakauma samalla *normalisoituu* ja sen yhteydet muihin muuttujiin *linearisoituvat*. Muitakin muunnoksia on olemassa, mutta niiden tulkinta on vaikeampaa. Tässä kirjassa näihin muunnoksiin ei perehdytä.

Selittäjistä tarvitsee tehdä vähemmän oletuksia. Jopa luokittelutason selittäjiä voidaan käyttää niin sanottujen *osoitinmuuttujien*, toiselta nimeltään *dummy*-muuttujien, välityksellä. Niihin perehdytään tarkemmin kohdassa 5.2.4 (s. 134). Osoitinmuuttujat ovat tyypillisesti dikotomisia muuttujia, jotka edustavat mallissa selittäjien, usein taustamuuttujien, eri luokkia. Sellaisenaan ei luokittelu- tai edes järjestystason muuttujia pidä malliin laittaa, ellei sitten kyse ole alun perin dikotomisesta muuttujasta. Sehän täyttää periaatteessa kaikki mittaustasovaatimukset, kuten jo kohdassa 2.3.3 (s. 39) todettiin.

Ulkonäkötutkimuksen regressiomalli

Seuraavissa esimerkeissä mallinnetaan vastaajilta tiedusteltua ihannepainoa edellä muodostetuilla faktoripisteillä ja eräillä taustamuuttujilla vuoden 1997 aineistossa. Aluksi katsotaan, miltä ihannepainon jakauma vaikuttaa. Kuvan 5.2 histogrammi näyttää jakauman olevan hieman vino: ihannepaino on vähän kallellaan keskimääräistä kevyempään suuntaan. Kuvassa on mukana myös oletuksen mukainen normaalijakauma.



Kuva 5.2. Ihannepainon histogrammi ja normaalijakauma.

Silmämääräisesti yhteensopivuus vaikuttaa melko hyvältä. Käsitystä tukee tulosteen 5.1 *yhteensopivuustesti*, jonka nollahypoteesi olettaa ihannepainon olevan normaalisti jakautunut. Testi osoittaa *havaittujen* (`f`) ja *odotettujen* (`e`) luokkafrekvenssien poikkeamien olevan sen verran pieniä, että voidaan katsoa niiden johtuvan sattumasta ($p = 0.097$). Jos siis nollahypoteesi hylättäisiin ja todettaisiin, että ihannepaino ei noudata normaalijakaumaa, olisi väärän johtopäätöksen tekemisen riski noin 10 %. Sitä pidetään yleensä liian suurena, joten nollahypoteesi jää voimaan. Johtopäätös on, että ihannepaino noudattaa normaalijakaumaa ja soveltuu näin regressiomallin vastemuuttujaksi.

Tulosteesta 5.1 nähdään, että aineiston havaintomäärä on jälleen eri kuin aikaisemmin. Faktorianalyysia tehtiin 268 havainnon turvin; nyt ihannepainosta on tiedossa vain 247. Osa vastaajista ei ole ilmaissut ihannepainoaan, mutta samalla on otettu huomioon puuttuvuus faktoripisteissä, joita tullaan käyttämään mallin selittäjinä. Kohdassa 3.5 (s. 81) suoritettu aineiston paikkaus ulotettiin ainoastaan kolmeen mittariin, joten lisää paikkaustarpeita seuraisi välittömästi.

Tuloste 5.1. Ihannepainon normaalisuuden testaus.

```
Frequency distribution of ipaino in UN2007FA: N=247

Class midpoint  f     %   Sum     %     e      e    f     X^2
        40.0    1   0.4     1   0.4   0.7
        45.0    1   0.4     2   0.8   5.2    5.9    2     2.6
        50.0   21   8.5    23   9.3  22.9   22.9   21     0.2
        55.0   64  25.9    87  35.2  56.0   56.0   64     1.2
        60.0   84  34.0   171  69.2  75.7   75.7   84     0.9
        65.0   44  17.8   215  87.0  56.8   56.8   44     2.9
        70.0   25  10.1   240  97.2  23.6   23.6   25     0.1
        75.0    6   2.4   246  99.6   5.4
        80.0    1   0.4   247 100.0   0.7
        85.0    0   0.0   247 100.0   0.0    6.1    7     0.1
Mean=60.06073 Std.dev.=6.258303
Fitted by NORMAL(60.061,39.166) distribution
Chi-square=7.858 df=4 P=0.0969
```

Tässä yhteydessä ei puuttuviin tietoihin voida kiinnittää huomiota niin paljon kuin todellisuudessa olisi syytä. Paikkaukset tulisi tehdä laajemmin, jotta aineisto olisi eri analyyseissa yhdenmukaisempi.

5.2.2 Selittäjien valinta

Tilastollisessa kirjallisuudessa annetaan paljon neuvoja regressiomallin selittäjien valintaan. Tarjolla on automaattisia valintamenetelmiä, jotka tilastollisin perustein poimivat tarjolla olevista selittäjistä ”parhaat”. Selittäjien valinta ei silti saa olla arpapeliä. Sisällöllisten perusteiden tulee olla ensisijaisia, ja tutkijalla pitää olla käsitys siitä, mikä vaikuttaa mihinkin ja millä tutkittavaa ilmiötä voidaan selittää.

Tässäkin asiassa näkyy yhtenevyys mittausmalliin, sillä keskeisten selittäjien valinta tapahtuu parhaiten jo tutkimuksen suunnitteluvaiheessa, samalla kun mietitään, mitä ilmiöitä tutkimuksessa pyritään selittämään. Nämä ovat usein tutkimussuunnitelman ydinasioita.

Selittäjinä faktoripisteet

Esimerkkinä tarkasteltavassa ulkonäkötutkimuksen regressiomallissa on tarkoituksena tutkia, miten hyvin ihannepainoa voi selittää aiemmin muodostetuilla faktoripisteillä, joihin tiivistyivät kolmen mittarin, yhteensä 53 osion mittaukset. Näistä muodostettiin kohdassa 4.4.1

(s. 109) neljä faktoripistemuuttujaa (ks. taulukko 4.1, s. 110). Valitaan faktoripisteet selittäjiksi ja estimoidaan regressiomallin parametrit.

Tuloste 5.2 sisältää regressioanalyysin olennaisimmat selittäjäkohtaiset tiedot: estimoidut *regressiokertoimet* (otsikolla `Regr.coeff`), niiden estimointitarkkuutta kuvaavat *keskivirheet* (`Std.dev.`) ja *t-arvot*, joilla kertoimien tilastollista merkitsevyyttä testataan niin kutsutulla *t-testillä*, sekä selittäjien keskinäiseen vertailuun soveltuvat *standardoidut regressiokertoimet* (`beta`). Sana `constant` tarkoittaa *vakiotermiä*, joka on käytännössä kiinteä, mutta melko tekninen osa regressiomallia. Vakion kuulumista malliin ei ole tapana testata tilastollisesti, vaikka tulosteissa senkin t-arvo raportoidaan. Tässä yhteydessä vakio kertoo ihannepainon keskiarvon, sillä faktoripistemuuttujien keskiarvot ovat aiemmin todetun perusteella nollia.

Tulosteessa ei ole testien p-arvoja, mutta tavanomaista viiden prosentin riskiä vastaa itseisarvoltaan noin kakkosen kokoinen t-testisuure. Testit koskevat nollahypoteeseja, joiden mukaan regressiokertoimet ovat nollia, siis että vastaavilla selittäjillä ei ole mallissa mitään virkaa. Tulkinnoissa keskitytään enimmäkseen selittäjiin ja niiden regressiokertoimiin. Esimerkiksi kun panostaminen ulkonäköön kasvaa yhden yksikön, niin sen regressiokertoimen mukaisesti ihannepaino vähenee yhden kilon.

Tuloste 5.2. Ihannepainon ja faktoripisteiden regressiomalli.

```
Linear regression analysis: Data UN2007RA, Regressand ipaino    N=247
N(missing)=26
Variable Regr.coeff Std.dev.    t     beta
TUNTO    -0.597318  0.401790 -1.487 -0.092 itsetunto ulkonäköasioissa
PANOS    -1.008377  0.418080 -2.412 -0.150 panostaminen ulkonäköön
PAINE     0.254721  0.410328  0.621  0.039 sosiaaliset ulkonäköpaineet
NEGAT     1.247659  0.429892  2.902  0.180 negatiivinen suhtautuminen
constant  60.11931  0.366771  163.9
Variance of regressand ipaino=38.59918205 df=246
Residual variance=36.55122054 df=242
R=0.2616 R^2=0.0685
```

Tulosteessa 5.2 on myös muita lukuja, jotka kertovat vastemuuttujan ja mallin *jäännösten* eli *residuaalien* vaihtelusta. Mallin *sovite*, joka mittauskehikon käsittein on tulosasteikko, syntyy muodostamalla selittäjistä painotettu summa, jossa painokertoimet ovat juuri estimoidut

regressiokertoimet. Näiden tarkasteluun palataan regressiodiagnostiikan yhteydessä, kohdassa 5.3 (s. 141).

Eniten mallissa näyttäisivät painottuvan lukujen etumerkeistä päätellen toisilleen vastakkaiset ”panostaminen ulkonäköön” ja ”negatiivinen suhtautuminen”. Enemmän ulkonäköönsä panostavien ihannepaino olisi näin keskimääräistä kevyempi, negatiivisesti ulkonäköön suhtautuvien painavampi. Itsetunto ja sosiaaliset paineet eivät näytä vaikuttavan, ainakaan niiden regressiokertoimet eivät t-testin perusteella poikkea merkitsevästi nollasta. Koska mallin *selitysaste* (`R^2`) on vain alle 7 %, ihannepainoa selittävät ilmeisesti suurelta osin aivan muut kuin nyt mallissa mukana olleet tekijät.

Miksei suoraan alkuperäisillä muuttujilla?

Vaikkeivät faktoripisteet näytä sellaisenaan paljoa selittävän ihannepainon vaihtelua, edellä laadittu malli näyttää kuitenkin, miten mittausmallin ja mitta-asteikon läpi suodatetut tiedot tulevat regressiomallissa käyttöön.

Tavallisesti regressioanalyysin kirjallisuudessa y-muuttujaa selitetään yhdellä tai useammalla x-muuttujalla. Tämäkin perinteinen mallintamisasetelma sisältyy kuvan 5.1 (s. 122) mittauskehikkoon – tarkemmin sanottuna se jää jäljelle, jos mittausmalli ja kaikki asteikot unohdetaan ja tehdään vain suoraan regressioanalyysia alkuperäisillä x-muuttujilla. Etenkään kyselytutkimuksessa siihen ei kuitenkaan kannata ryhtyä. Tähän on kolme painavaa syytä:

- *x-muuttujissa on mittausvirheitä*, joita regressiomalli ei millään tavoin huomioi
- *x-muuttujat korreloivat keskenään*, jolloin niiden vaikutuksia on hankala erottaa toisistaan
- *x-muuttujia on liikaa*, mikä tekee mallintamisesta epäselvää.

Pahinta jälkeä tulee, jos selittäjät valitaan automaattisilla keinoilla. Edellä luetelluista syistä ja niiden yhdistelmistä johtuen lopputulokset saattavat olla, mitä sattuu. Näytteenä siitä, miten voi käydä, vilkaistaan tulostetta 5.3. Siinä esitetty malli on aikaansaatu *askeltavalla menettelyllä*, jossa selittäjiä lisäillään ja poistellaan vuorotellen mallista tilastollisten kriteerien nojalla. Liikkeelle on lähdetty alkuperäisistä 53 muuttujasta, ja proseduurin päätteeksi jäljelle on jäänyt neljä. On

siis päästy samaan selittäjämäärään kuin edellä, mutta monivaiheisen tiivistämisen sijaan – jonka selostamiseen edellä käytettiin kokonainen luku – on valittu nämä neljä muuttujaa ja heitetty loput 49 sivuun. Saattaa vaikuttaa tehokkaalta, mutta sekä mallin luotettavuus että yleistettävyys jäävät kyseenalaiseksi.

Tuloste 5.3. Ihannepainon askeltamalla valittu regressiomalli.

```
Linear regression analysis: Data STEPREG, Regressand ipaino    N=249
N(missing)=24
Variable Rcoeff Stddev   t
k26.6     1.001 0.366  2.763 En ole fyysisesti viehättävä.
k26.15   -0.849 0.336 -2.523 Meikattuna olen tyytyväisempi ulkonäköön.
k26.21   -1.088 0.445 -2.446 Tiedän, jos olen "huonosti laitettu".
k71.9     0.908 0.353  2.571 Elämässä pärjää ulkonäöstä riippumatta.
constant  61.86 2.785  22.21
Variance of regressand ipaino=38.39067561 df=248
Residual variance=34.82447443 df=244
R=0.3279 R^2=0.1075
```

Erilaisilla valintamenettelyillä päädytään erilaisiin malleihin, mikä jättää paljon arvailujen varaan. Jos käytetään kaikkia 53:a muuttujaa karsimatta yhtäkään, eniten esiin sattuvat nousemaan tulosteen 5.3 kolme viimeistä muuttujaa. Siitä huolimatta on kyseenalaista, edustavatko juuri nämä muuttujat parhaalla tavalla tutkittavien ulkonäkökäsitysten ulottuvuuksia. Faktorianalyysissa yksikään niistä ei ollut parhaiden osioiden joukossa (vrt. tuloste 4.2, s. 101), päinvastoin. Mallintamisen tavoite on tiivistäminen, mutta se on tehtävä faktorianalyysilla eikä regressioanalyysilla.

5.2.3 Taustamuuttujat ja ennustevaliditeetti

Palataan faktoripisteisiin, joilla regressiomallintaminen aloitettiin. Ensimmäisen mallin perusteella (ks. tuloste 5.2, s. 129) ei kannata tehdä lopullisia johtopäätöksiä, muttei myöskään karsia selittäjiä. Sen sijaan olisi pohdittava mallin sisältöä paremmin, sillä siitä näyttää puuttuvan jotain olennaista. Malliin olisi syytä sisällyttää myös muuttujia, joilla vastaajien erilaiset taustat otetaan huomioon. Tällainen *taustamuuttuja* voisi olla esimerkiksi vastaajan oma paino, sillä ihannepaino riippuu varmasti nykyisestä painosta. Toinen asia, joka olisi hyvä ot-

taa mallissa huomioon, on ikä. Muitakin keskeisiä tekijöitä voi olla, mutta tässä tyydytään näihin kahteen.

Tulosteessa 5.4 malli on estimoitu uudelleen, kun paino ja ikä on lisätty selittäjiksi. Niiden myötä häviää taas muutamia havaintoja. Molempien taustamuuttujien regressiokertoimet ovat t-arvojen perusteella tilastollisesti merkitseviä. Painolla on odotetusti suurin ”painoarvo”. Sen voi päätellä standardoidusta regressiokertoimesta 0.74, joka on mallin suurin. Koska selittäjät eivät korreloi juurikaan keskenään, sama luku kertoo myös painon ja ihannepainon välisen korrelaation. Luvun neliö on noin 0.55, josta voi päätellä, että paino yksin vastaa suurinta osaa mallin selitysasteesta, joka on 0.59.

Tuloste 5.4. Taustamuuttujilla täydennetty regressiomalli.

```
Linear regression analysis: Data UN2007RA, Regressand ipaino    N=243
N(missing)=30
Variable Regr.coeff Std.dev.    t     beta
paino     0.376200  0.023887  15.75  0.740 paino (kg)
ika       0.065179  0.021557  3.024  0.148 ikä (vuosina)
TUNTO     0.833334  0.286837  2.905  0.131 itsetunto ulkonäköasioissa
PANOS    -0.457873  0.278438 -1.644 -0.069 panostaminen ulkonäköön
PAINE    -0.156410  0.271834 -0.575 -0.024 sosiaaliset ulkonäköpaineet
NEGAT    -0.071129  0.318041 -0.224 -0.011 negatiivinen suhtautuminen
constant  32.33364  1.562751  20.69
Variance of regressand ipaino=37.07904806 df=242
Residual variance=15.55229076 df=236
R=0.7687 R^2=0.5910
```

Painoa ja ikää hyödynnettiin tulosteen 5.4 mallissa taustamuuttujina. Ne ovat silloin hieman eri asemassa kuin muut selittäjät. Tämän asian ilmaisemiseen on useita tapoja: voidaan muun muassa sanoa, että paino ja ikä on *otettu huomioon* tai että ne on *vakioitu*. Regressiomalli ei selittäjiä erottele; vakiointi on enemmän tulkintakysymys. Tavallaan kaikki selittäjät vakioivat toisiaan, mutta osaan vain suhtaudutaan taustamuuttujina. Tällaisia muuttujia ei ole yleensä syytä poistaa mallista tilastollisin perustein, vaan päinvastoin ne pidetään mukana sisällöllisin perustein. Periaatteessa voitaisiin yhtä hyvin ajatella, että varsinainen mielenkiinto kohdistuisikin painoon ja ikään, ja vakiointi tehtäisiin faktoripisteiden edustamien ulottuvuuksien suhteen. Tähän ajatukseen palataan kohdassa 5.2.4 (s. 134).

Tulosteesta 5.4 voidaan tulkita esimerkiksi, että itsetunnon merkitys korostuu, kun paino ja ikä on vakioitu: mitä parempi itsetunto, sitä korkeampi ihannepaino. Negatiivisen suhtautumisen merkitys hälvenee aiempaan verrattuna täysin, ja ulkonäköön panostamisen merkitys vähenee, joskin se edelleen osoittaa kevyemmän ihannepainon suuntaan. Sosiaalisten paineiden vaikutussuunta muuttuu aiempaan verrattuna, mutta yhteys on melko heikko. Kaikkiaan tässä esitetyt tulkinnat ovat aika pinnallisia, mutta vakioinnin merkitys on kuitenkin selvä. Samalla mallin selitysaste nousee 7 %:sta lähes 60 %:iin.

Todellisuudessa, kun selittäjiä on yleensä enemmän, voi olla hyvä tarkastella yksittäisten selittäjien ohella niistä koostuvia ryhmiä. Ulkonäkötutkimuksessa selittäjäryhmiä voisivat taustamuuttujien lisäksi olla esimerkiksi fyysiset tekijät, psyykkiset tekijät ja asenteet. Tässä käsitellyt mallit ovat sen verran yksinkertaisia, ettei tällaisia tarkastelutapoja tarvita.

Tulosasteikon ennustevaliditeetti

Selitysaste kertoo siis, kuinka suuren osan mallin selittäjät pystyvät vastemuuttujan vaihtelusta selittämään. Edellä olevissa tulosteissa esiintyy myös symboli R. Se tarkoittaa *yhteiskorrelaatiokerrointa* eli vasteen ja sovitteen välistä korrelaatiokerrointa, jonka neliö selitysaste on. Yhden selittäjän mallissa yhteiskorrelaatio olisi sama asia kuin vasteen ja selittäjän välinen tavallinen korrelaatio.

Mittauskehikossa yhteiskorrelaatiokerroin tulkitaan *tulosasteikon ennustevaliditeetiksi*. Mitä parempi ennustevaliditeetti, sitä luotettavammin mallin synnyttämällä tulosasteikolla voidaan ennustaa tai selittää eroja vertailuperusteena käytetyssä vastemuuttujassa. Tämä on kuitenkin vain suuntaa-antavaa tietoa, joka kertoo mallin todellisesta luotettavuudesta aika vähän. Selitysasteella, tai yhtä hyvin ennustevaliditeetilla, on perustavaa laatua oleva heikkous: se kasvaa, kun malliin lisätään selittäjiä, oli näillä selitysvoimaa tai ei. Sen takia pelkästä selitysasteesta ei kannata tehdä liikoja tulkintoja.

Käytännössä selitysastetta enemmän kiinnostavat selittäjät, joihin seuraavaksi perehdytään tarkemmin. Tulosteen 5.4 mallin luotettavuuteen palataan vielä kohdassa 5.3 (s. 141), jossa tutkitaan, miten hyvin mallista tehdyt oletukset pitävät paikkansa.

5.2.4 Luokitellut selittäjät

Edellä laadituissa malleissa selittäjät olivat jatkuvia – joko asteikkoja kuten faktoripisteet tai jo alun perin tasoltaan numeerisia mittauksia kuten paino ja ikä. Näiden selittäjien tulkinta voi olla vaikeaa, riippuen näkökulmasta. Jos halutaan tutkia tarkemmin painon ja iän vaikutuksia ihannepainoon, eivät yksittäiset regressiokertoimet paljoa auta. Niistä voi päätellä vaikutuksen suunnan ja voiman, mutta tulkinnat jäävät vaisuiksi. Tulkinta helpottuu luokittelemalla jatkuvat selittäjät uusiksi, diskreeteiksi muuttujiksi. Informaatiota hukataan, mutta vastapainoksi saavutetaan parempi tulkittavuus.

Neljään luokkaan jaettu paino ja kuuteen luokkaan jaettu ikä on tulosteessa 5.5 taulukoitu vastakkain. Frekvenssien lisäksi taulukon soluissa on ihannepainon keskiarvot ja hajonnat. Esimerkiksi 26–35-vuotiaita, 46–60-kiloisia naisia on vuoden 1997 aineistossa 18. Heidän ihannepainonsa on keskimäärin 55 kiloa ja hajonta noin neljä kiloa.

Tuloste 5.5. Ihannepaino ikä- ja painoluokissa.

```
         LIKA6   18-25   26-35   36-45   46-55   56-65   66-74
LPAINO4  *****
46-60                17      18      33      14       5       3
         Mean     55.29   55.00   55.86   54.50   58.00   53.67
         SD        3.55    3.94    3.33    3.41    2.35    2.08

61-70                15      19      20      13      16       6
         Mean     58.00   59.89   59.50   60.38   62.38   60.67
         SD        1.85    2.71    4.52    4.07    3.16    5.89

71-80                 4      10      10      11       6       3
         Mean     60.25   64.20   64.10   66.09   66.17   68.67
         SD        7.41    4.29    4.75    3.86    5.78    7.77

81-115               0       4       2      10       6       2
         Mean        -   61.25   70.00   72.60   69.50   68.50
         SD          -    2.50    0.00    4.01    6.28    4.95
```

Tarkempi näkymä edellä mainittuun aineiston osajoukkoon on tulosteessa 5.6, josta selviää, että minimi on sama kuin kyseisen luokan painon minimi, 46 kiloa, mutta maksimi on hieman sen yli. Viiden vastaajan tiedot näyttävät puuttuvan.

Luokittelut on tehtävä harkiten, sillä ne vaikuttavat olennaisesti tulkintoihin. Monesti luokitukset kannattaa tehdä niin, että ne ovat hel-

Tuloste 5.6. Ihannepaino, kun paino 46–60 kg ja ikä 26–35 v.

```
Basic statistics of data UN2007RA N=23
Variable   mean      stddev          N  minimum   maximum
 ipaino    55.00000  3.940737       18  46.00000  62.00000
```

pommin vertailtavissa aiempiin tutkimuksiin. Esimerkiksi ikäryhmien muodostamisessa voi olla syytä noudattaa virallisissa tilastoissa sovellettuja luokituksia. Tässä on tyydytty aineistolähtöiseen luokitteluun, jossa olisi varmasti parantamisen varaa, sillä joissakin luokissa on varsin vähän havaintoja, yhdessä ei lainkaan. Näistä aiheutuu mallintamisessa ongelmia, joita voi koettaa ratkaista joko yhdistämällä luokkia tai luokittelemalla muuttujat uudelleen.

Osoittimet ja vertailuluokka

Luokitellut versiot painosta ja iästä eivät paranna regressiomallia lainkaan, oikeastaan päinvastoin, koska ne ovat vain järjestystasoisia muuttujia, koodattuina numeroin 1–4 ja 1–6. Jos ne laitetaan malliin alkuperäisten tilalle, siirrytään karkeampiin selittäjiin ja luovutaan kiloista ja vuosista. Näin saadaan kyllä suurin piirtein samat tulokset, mutta vielä huonommin tulkittavassa muodossa.

Kätevin tapa edetä regressiomallintamisessa on luoda jokaiselle luokalle oma *osoitinmuuttuja*, lyhyemmin *osoitin*, joka ilmaisee, kuuluuko havainto kyseiseen luokkaan vai ei. Selvintä on soveltaa dikotomista koodausta, jossa osoittimen arvo on ykkönen, jos havainto kuuluu luokkaan, ja nolla, jollei. Osoittimet on helppo tehdä, kun luokitukset on päätetty. Painoa tarvitaan edellä tehdyssä luokituksessa kuvaamaan neljä osoitinta, ikää kuusi. Valmiiksi kaksiluokkaisia selittäjiä voisi käyttää sellaisenaan, mutta nekin on usein selvempää koodata nollilla ja ykkösillä, jos ne alun perin on koodattu esimerkiksi ykkösillä ja kakkosilla, kuten tulosteessa 3.4 (s. 59) esiintynyt muuttuja ”vakituinen parisuhde”.

Kun osoittimet on luotu, ne lisätään malliin selittäjiksi alkuperäisen jatkuvan selittäjän tilalle. Yksi jatkuva selittäjä korvautuu siis useilla dikotomisilla selittäjillä, joista koostuvaa selittäjäryhmää voidaan nimittää *osoitinselittäjäksi*.

Ennen kuin katsotaan, miltä malli osoitinselittäjien lisäyksen myötä näyttää, on jäljellä vielä yksi tulkinnan kannalta tärkeä asia. Kaikkia yksittäisiä osoittimia ei voi nimittäin sisällyttää malliin yhtaikaa. Se johtuu siitä, että kunkin luokitellun selittäjän osoittimet ovat toisensa poissulkevia, jolloin ne riippuvat täydellisesti toisistaan. Parametrien estimointi ei onnistu, jos selittäjien välillä on täydellisiä sidoksia. Sama ongelma esiintyisi, jos vahingossa laittaisi selittäjäksi saman muuttujan kahteen kertaan.

Jos tiedetään painon osoittimista kolme, tiedetään neljäskin, koska kukin vastaaja kuuluu vain yhteen painoluokkaan. Siis vain yksi painon osoittimista on ykkönen, muut nollia. Sama koskee jokaista osoitinselittäjää.

Sidokset vältetään valitsemalla jokaisesta osoitinselittäjästä yksi luokka *vertailuluokaksi* tai *vertailuryhmäksi* ja jättämällä sitä vastaava osoitin pois mallista. Nimensä mukaan vertailuluokka on sellainen, johon muita osoitinselittäjän luokkia vertaillaan. Sen valinta perustuu puhtaasti tulkintaan; tilastollisesti on samantekevää, mikä luokista on vertailuluokka. Valintaa voidaan muuttaa mallinnuksen edetessä, samoin selittäjien luokituksia. Tavoitteena on mahdollisimman selkeästi tulkittava malli.

Regressiomalli osoitinselittäjillä

Tuloste 5.7 esittää estimoidut regressiokertoimet ja muita tietoja mallista, jossa paino ja ikä ovat osoitinselittäjinä. Tulosteen ulkoasu on erilainen kuin aiempien mallitulosteiden, mutta siitä näkyvät vastaavat tiedot. Viimeisen sarakkeen `Sig.` (*significance*, merkitsevyys) tarkoittaa p-arvoa. Lukuja on hyödyllistä vertailla tulosteeseen 5.4 (s. 132). Faktoripisteiden tiedot ovat hyvin samanlaiset molemmissa malleissa, koska selittäjät eivät juuri korreloi keskenään. Olennainen ero on taustamuuttujissa. Sekä painon että iän vertailuluokaksi on valittu ensimmäinen luokka (ks. tuloste 5.5, s. 134), mistä johtuen osoittimet `paino1` ja `ika1` eivät sisälly mallin tulosteeseen. Muille luokille saadaan kullekin oma regressiokerroin ja muut tiedot.

Tulosteesta 5.7 painon ja iän yhteydet ihannepainoon on aiempaa helpompi tulkita. Painon osalta yhteys on selvä, ja luokat eroavat 46–60 kilon vertailuluokasta sitä enemmän, mitä painavammasta luokasta

Tuloste 5.7. Osoitinselittäjillä täydennetty regressiomalli.

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
	(Constant)	54.506	.732		74.414	.000
	TUNTO	.780	.281	.123	2.776	.006
	PANOS	-.663	.280	-.100	-2.364	.019
	PAINE	-.059	.274	-.009	-.214	.831
	NEGAT	.152	.323	.022	.470	.639
	paino2	4.528	.602	.358	7.519	.000
	paino3	9.549	.772	.594	12.375	.000
	paino4	13.570	.987	.666	13.756	.000
	ika2	.166	.868	.011	.191	.849
	ika3	.725	.843	.053	.860	.391
	ika4	1.685	.927	.110	1.818	.070
	ika5	3.188	1.023	.177	3.116	.002
	ika6	.341	1.394	.013	.245	.807

on kyse. Kaikki kolme painon osoitinta ovat merkitseviä. Sen sijaan ikä näyttäytyy tyystin toisenlaisena. Vasta ikäluokat 46–55 ja 56–65 vuotta eroavat merkitsevästi 18–25-vuotiaiden vertailuluokasta. Muiden luokkien erot eivät ole merkitseviä.

Kumpaakaan taustamuuttujaa ei voi tulosteen 5.7 perusteella poistaa mallista, sillä ainakin tilastollisesti merkitseviä eroja näyttää ilmenevän. Etenkin jos painoon ja ikään suhtaudutaan taustamuuttujina, ne pidetään mallissa p-arvoista riippumatta, kuten aiemmin todettiin. Jos osoitinselittäjä poistetaan mallista, on poistettava kaikki sen osoittimet. Yksittäistä osoitinta ei saa poistaa, sillä tällöin sen osoittaman luokan havainnot yhdistyvät vastaavaan vertailuluokkaan. Mahdolliset luokkien yhdistelyt on tehtävä erikseen.

Yhden selittäjän mallissa, jossa voidaan piirtää regressiosuora, osoitinselittäjiin siirtyminen tarkoittaa, että malli tuottaa useita samansuuntaisia regressiosuoria eri tasoille. Mallintamisessa kiinnostaa silloin, ovatko *tasoerot* niin suuria, että tarvitaan luokittaisia regressiosuoria, vai riittääkö yksinkertaisempi malli. Tasoerojen lisäksi saattaa ilmetä *rakenne-eroja*, joiden analysointia tarkastellaan seuraavassa.

Varianssianalyysi

Tulosteessa 5.7 ei mainittu selitysastetta, mutta se on suunnilleen sama kuin vastaavassa jatkuvien selittäjien mallissa (ks. tuloste 5.4, s. 132). Itse asiassa se on jopa hivenen korkeampi, mikä johtuu siitä, että selittäjiä on enemmän. Erot selitysasteissa ovat joka tapauksessa epäolennaisia, sillä tärkeintä on selittäjien tulkinnan helppous.

Selitysasteen ohella regressiomallien yhteydessä voidaan tarkastella myös *varianssitaulua*, joka osoitinselittäjämallista on esitetty tulosteessa 5.8. Siitä nähdään, miten vastemuuttujan vaihtelu on hajotettu mallin selittämään (`Regression`) ja selittämättä jääneeseen (`Residual`) osaan. Selitysaste on mallin selittämä osuus kokonaisvaihtelusta (`Total`). Taulun luvut vastaavat aiempien tulosteiden alaosissa olleita lukuja, mutta niihin ei tässä syvennytä sen tarkemmin kuin että eri osien varianssien (`Mean Square`) suhteesta muodostuu taulun oikean reunan F*-testisuure*. Sillä testataan nollahypoteesia, jonka mukaan kaikki mallin varsinaiset regressiokertoimet ovat nollia. Jos tämä hypoteesi jää voimaan, ei malliin ole kyetty valitsemaan yhtään järkevää selittäjää.

Tuloste 5.8. Osoitinselittäjämallin varianssitaulu.

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5524.692	12	460.391	30.707	.000
	Residual	3448.437	230	14.993		
	Total	8973.130	242			

Jos tutkija on osannut valita malliin oikeita selittäjiä, regressiomallin F-testin p-arvo osoittautuu yleensä merkitseväksi niin kuin tässäkin, onhan tulosteen 5.8 p-arvo kolmella desimaalilla nolla. Testin informaatioarvo on kuitenkin vähäinen; korkeintaan se todistaa, että mallissa on jotain ideaa. Regressiokertoimien tulkinta on paljon olennaisempaa, kuten edellä on todettu. Regressioanalyysin varianssitaulu on sisällytetty tähän siksi, että se luo yhteyden *varianssianalyysiin*, jota tyypillisesti pidetään kokonaan eri menetelmänä. Todellisuudessa se on läheistä sukua regressioanalyysille, sillä molemmat perustuvat samaan lineaariseen malliin.

Varianssianalyysin nimi johtuu edellä selostetusta varianssin hajotelmasta, mutta varsinaisesti siinä mielenkiinto kohdistuu eri luokkien tai ryhmien keskiarvojen vertailuun, jota regressiomallissakin osoittimien myötä tehdään. Mitä enemmän mallissa on luokiteltuja selittäjiä, sitä enemmän erilaisia havaintojen ryhmittäisiä keskiarvovertailuja se mahdollistaa. Tällöin varianssianalyysi voi olla regressioanalyysia kätevämpi menetelmä.

Yhdysvaikutukset

Varianssianalyysin edut korostuvat, jos tutkitaan selittäjien mahdollisia *yhdysvaikutuksia* eli *interaktioita*. Tähänastiset tarkastelut ovat rajoittuneet *päävaikutuksiin*, siis esimerkiksi iän ja painon yksittäisiin vaikutuksiin ihannepainoa selitettäessä. Tällöin tullaan olettaneeksi painon vaikutus samanlaiseksi kaikissa ikäluokissa ja toisinpäin. Jos oletus pätee, malli on yksinkertaisempi, mutta jollei se päde, joudutaan malliin sisällyttämään yhdysvaikutuksen huomioivat lisäselittäjät. Sisällöllisesti se tässä tilanteessa tarkoittaisi, että painon vaikutus ihannepainoon olisi erilainen eri ikäluokissa.

Tuloste 5.9. Yhdysvaikutusmallin varianssitaulu.

Tests of Between-Subjects Effects

Dependent Variable: ipaino

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	5965.483	26	229.442	16.478	.000
Intercept	451083.194	1	451083.194	32395.421	.000
TUNTO	114.035	1	114.035	8.190	.005
PANOS	68.780	1	68.780	4.940	.027
PAINE	1.779	1	1.779	.128	.721
NEGAT	2.845	1	2.845	.204	.652
LPAINO4	2601.388	3	867.129	62.275	.000
LIKA6	313.528	5	62.706	4.503	.001
LPAINO4 * LIKA6	440.791	14	31.485	2.261	.007
Error	3007.646	216	13.924		
Total	885754.250	243			
Corrected Total	8973.130	242			

Katsotaan, miltä tilanne näyttää aineiston valossa. Laajennetaan edellä tehtyä osoitinselittäjämallia ottamalla mukaan painon ja iän keskinäinen yhdysvaikutus, mutta käytetään nyt regressioanalyysin si-

jasta varianssianalyysia. Koska mallissa ovat mukana myös jatkuvat faktoripisteselittäjät, kyseessä on oikeastaan regressioanalyysin ja varianssianalyysin välimuoto, *kovarianssianalyysi*. Tuloste 5.9 on laajennettu versio tulosteen 5.8 varianssitaulusta. Samalla siitä nähdään kaikkien selittäjien merkitsevyydet. Paino ja ikä esiintyvät nyt aiemmin tehdyissä luokissa. Tulosteen neljänneksi alin rivi ilmaisee painon ja iän yhdysvaikutuksen.

Tulosteen 5.9 mukaan yhdysvaikutus on tilastollisesti merkitsevä ($p = 0.007$), joten se jätetään malliin. Tulkinta perustuu enimmäkseen yhdysvaikutuksiin; painon ja iän päävaikutuksia ei tule tarkastella. Niitä ei kuitenkaan saa poistaa, sillä muuten malli menettää *hierarkkisuutensa*; siltä ikään kuin putoaa pohja pois. Yhdysvaikutusmallin regressiokertoimia tulkitaan samaan tapaan kuin aiemmin esitetyssä osoitinselittäjämallissa, kertoimia vain on enemmän. Tilan säästämiseksi niitä ei ole esitetty tässä. Helpoiten tulkinta tapahtuu luokittaisia keskiarvoja esittävästä kuvasta 5.3, johon seuraavassa tutustutaan.

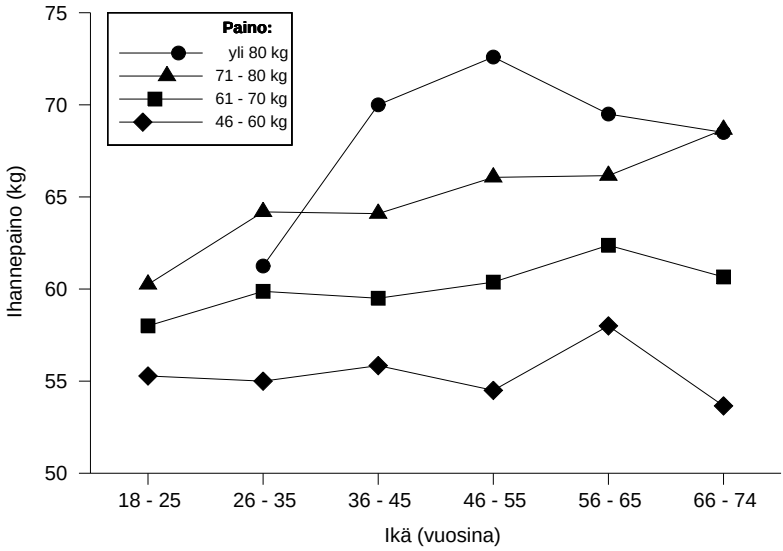
Keskiarvoprofiilit

Varianssianalyysi on nimestään huolimatta keskiarvoerojen vertailumenetelmä. Siihen sisältyy käytännössä vielä enemmän tilastollisia testejä kuin mitä tässä on esitetty. Paras keino luokkien tasoerojen ja yhdysvaikutusten tarkasteluun on piirtää *keskiarvoprofiileja*. Todelliset erot näkyvät niistä välittömästi. Liiallista tilastollista testausta on syytä välttää.

Kuvaan 5.3 on piirretty ihannepainon keskiarvoprofiilit paino- ja ikäluokittain. Jos profiileja kuvaavat viivat ovat samansuuntaisia, yhdysvaikutusta ei ole. Kuvasta ilmenee, että osa viivoista leikkaa toisensa. Selvimmät erot ilmenevät yli 80 kiloa painavilla. Myös vanhimmassa ikäluokassa painon vaikutus ihannepainoon on vähän erilainen kuin muissa ikäluokissa. Kolmen alimman painoluokan profiilit muistuttavat toisiaan melko paljon.

Kuvan 5.3 esittämät luvut ovat tämän kohdan alussa esitetyn tulosteen 5.5 (s. 134) ristiintaulukkoon sisältyviä keskiarvoja. Yksi luokista, yli 80-kiloiset 18–25-vuotiaat, oli tyhjä, joten sen piste puuttuu myös kuvasta.

Varianssianalyysia ja sen eräitä laajennuksia käsittelee yksityiskohtaisemmin muun muassa Nummenmaa (2004, 173–262).



Kuva 5.3. Ihannepainon keskiarvo paino- ja ikäluokittain.

5.3 Regressiodiagnostiikka

Tilastollisia malleja laadittaessa joudutaan tekemään erilaisia oletuksia. Jotta mallit eivät jäisi pelkkien oletusten varaan, on syytä tarkistaa, miten hyvin oletukset pitävät paikkansa. Tällaista mallin toimivuuden tutkimista kutsutaan *mallidiagnostiikaksi* ja erityisesti regressiomallin tapauksessa *regressiodiagnostiikaksi*. Seuraavassa keskitytään diagnostiikan olennaisimpiin osa-alueisiin enimmäkseen visuaalisin keinoin. Syvällisemmin regressiodiagnostiikkaan perehdyttää esimerkiksi teos *Applied Regression Including Computing and Graphics* (Cook & Weisberg, 1999).

Diagnosoitavana mallina seuraavissa tarkasteluissa on edellä laadittu, taustamuuttujien suhteen vakioitu regressiomalli, jossa vastemuuttuja on ihannepaino, ja selittäjinä ovat faktoripisteet (ks. tuloste 5.4, s. 132).

5.3.1 Jäännösvaihtelu

Valtaosa regressiodiagnostiikasta koskee mallin *jäännösvaihtelua*: sitä osaa vastemuuttujan vaihtelusta, jota malli ei selitä. Jäännösvaihtelua tutkimalla arvioidaan mallista tehtyjen oletusten pätevyyttä.

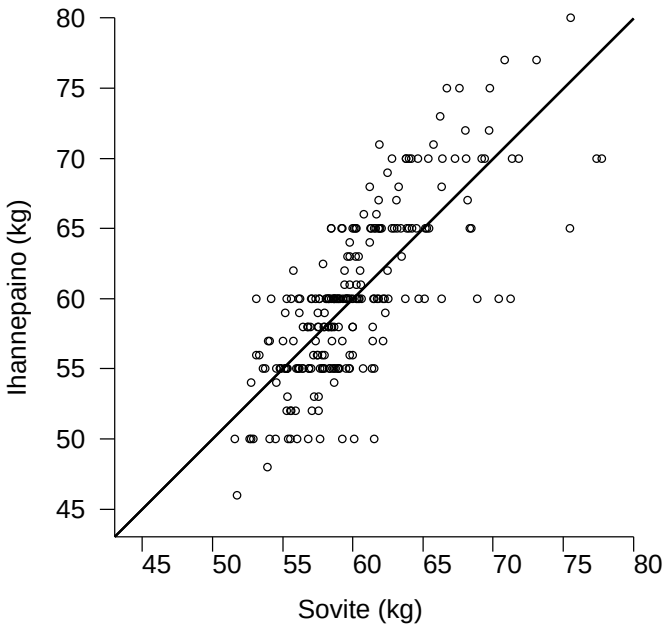
Sovite, vaste ja jäännös

Kohdan 5.2.2 (s. 128) alussa todettiin, että sovite on selittäjien regressiokertoimilla painotettu summa. Mitä paremmin sovite korreloi vastemuuttujan kanssa, sitä parempia ovat mallin selitysaste, yhteiskorrelaatiokerroin tai ennustevaliditeetti. Kaikki nämä siis kuvaavat karkeasti mallin hyvyyttä.

Kuvassa 5.4 sovite ja vaste on piirretty vastakkain. Koska kuvan vaaka- ja pystyakselit ovat samat, jakaa kuvaan piirretty regressiosuora koordinaatiston kahteen yhtä suureen osaan. Suoran yläpuolella ihannepaino on suurempi, ja alapuolella pienempi, kuin mallin antama tulos. Toisin sanottuna suoran yläpuolella mallin jäännökset ovat positiivisia ja alapuolella negatiivisia. Suoran kohdalla havainnot ovat parhaiten mallin mukaisia; malli siis selittää kyseisten vastaajien ihannepainon tarkalleen heidän painonsa, ikänsä ja ulkonäkökäsitystensä perusteella.

Samalla kuva 5.4 paljastaa vastemuuttujaan sisältyvän lievän, sinänsä merkityksettömän diskreettiyden, joka ei noussut esiin sen histogrammista (kuva 5.2, s. 127). Pisteet muodostavat vaakasuoria rivejä, sillä tasalukemat 50, 55, 60, 65 ja 70 ovat ymmärrettävästi suositumpia ihannepainoja kuin niiden väliin jäävät lukemat. Silti melkein kaikkia kokonaislukuja 45:n ja 80 kilon väliltä näyttää esiintyvän. Mallin sovite ei ulotu aivan yhtä laajalle alueelle.

Kuvassa 5.4 vain pieni osa pisteistä on täsmälleen kuvaan piirretyllä suoralla; suurin osa on ainakin jonkin verran sen ulkopuolella. Jos kaikki pisteet olisivat regressiosuoralla, selitysaste olisi 100 %, mutta niin täydellinen selitys olisi jo epäilyttävää. Erityisesti yhteiskuntatieteissä selitysasteet ovat selvästi alhaisempia, sillä tutkittaviin ilmiöihin vaikuttavat tyypillisesti monet aineiston ulkopuolelle rajatut tekijät. Toisaalta verrattain pienetkin selitysasteet voivat kätkeä taakseen sisällöllisesti kiinnostavia yhteyksiä. Nyt tarkasteltavan mallin selitysasteessa päästiin 60 %:iin (ks. tuloste 5.4, s. 132), joten jäännösvaihtelun osuudeksi jäi noin 40 %.



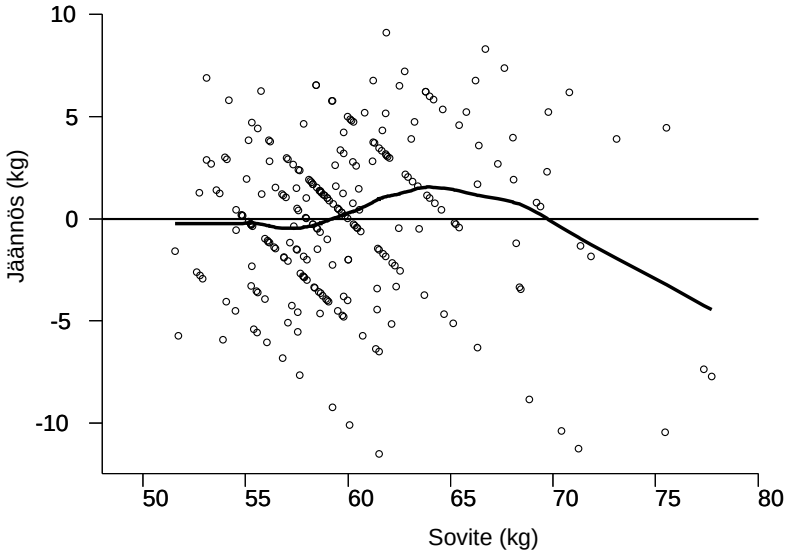
Kuva 5.4. Regressiomallin sovite ja vaste.

Mallin harhattomuus

Mallin toimivuuden kannalta tärkein ominaisuus on sen *harhattomuus*. Harhaton malli antaa keskimäärin oikeita tuloksia kaikilla selittäjien arvoilla. Mikäli malli antaa joillain selittäjien arvoilla esimerkiksi säännönmukaisesti korkeampia tuloksia, malli on harhainen. Harhattomuutta arvioidaan jäännösvaihtelun perusteella.

Jäännösvaihtelua on tutkittava vastemuuttujan suuntaisesti, siis tarkastellen havaintojen pystysuuntaisia etäisyyksiä regressiosuorasta. Kuvasta 5.4 tämä on hankalaa, mutta tehtävä helpottuu, kun vasteen tilalle kuvaan sijoitetaan jäännös. Kuvaa 5.5, jossa ovat vastakkain sovite ja jäännös, kutsutaan jäännösvaihtelukuvaksi. Siinä regressiosuora kulkee vaakasuunnassa, mikä tarkoittaa, että sovite ja jäännös eivät korreloi keskenään.

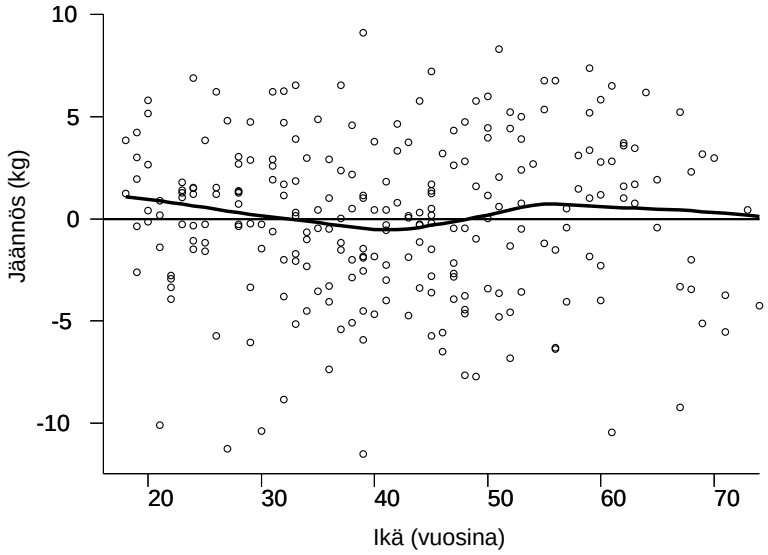
Kuvasta 5.5 on helpompi arvioida mallin harhattomuutta. Sitä auttaa myös kuvaan piirretty *regressiotasoitus*, joka on ikään kuin paloittain muodostettu regressiomalli. Mitä paremmin se seurailee yhdestä palasta muodostettua, tavallista regressiosuoraa, sitä harhattomampi malli on. Aineiston keskivaiheilla tasoitus näyttää tekevän pienen kaarroksen, mutta se johtunee vain satunnaisvaihtelusta. Noin 70 kilon kohdalla tasoitus sen sijaan kääntyy alas osoittaen lievää harhaisuutta aineiston loppupäässä. Se puolestaan selittyy sillä, että loppupäässä on vain vähän havaintoja ja viimeisillä on suurehkot negatiiviset jäännökset.



Kuva 5.5. Regressiomallin tasoitettu jäännösvaihtelu.

Harhattomuuden lisäksi jäännösvaihtelun tulisi olla samansuuruista sovitteen koko alueella. Kuvan 5.5 mukaan tämäkin pätee melko hyvin; vain edellä havaitut viiden vuoden piikit ihannepainon jakaumassa aiheuttavat pientä aaltoilua jäännöksen negatiiviselle puolelle. Regressiosuoran yläpuolella vaihtelu on varsin tasaista.

Vastaavia kuvia kannattaa piirtää myös jokaisesta mallin selittäjästä jäännöstä vasten, sillä ne voivat paljastaa systemaattisuuksia, jotka eivät näy sovitteen jäännösvaihtelukuvasta. Tarkasteltavassa mallissa kuvia tarvittaisiin kuusi lisää, mutta tilan säästämiseksi tässä on piirretty niistä vain yksi, iän ja jäännöksen välinen kuva 5.6. Sen perusteella ei ole mitään syytä epäillä mallin harhattomuutta, sillä tasoitus ei mutkittele käytännössä lainkaan. Jäännösvaihtelu on myös varsin samansuuruista kaikenikäisillä vastaajilla.



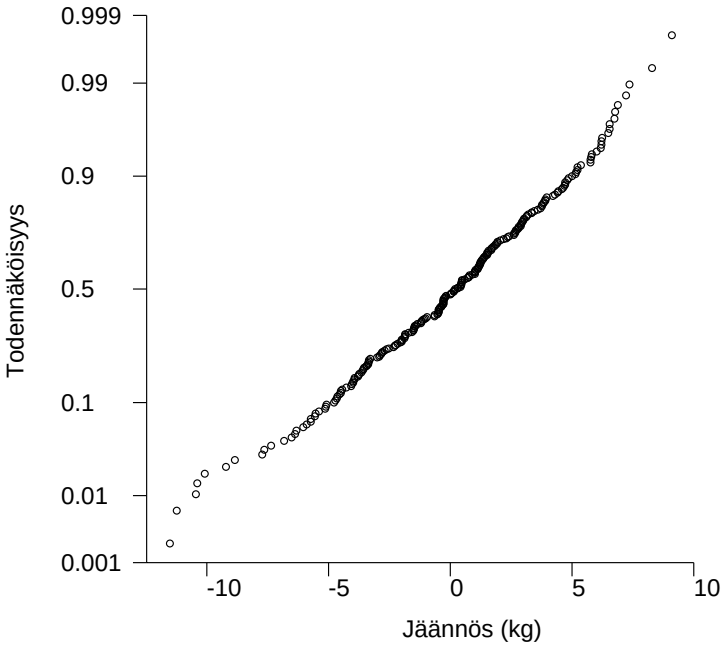
Kuva 5.6. Regressiomallin jäännösvaihtelu iän suhteen.

Tähänastisen diagnostiikkatarkastelun perusteella voidaan jo todeta, että mallin toimivuuden kannalta kaksi tärkeintä ominaisuutta, harhattomuus ja samansuuruinen jäännösvaihtelu, ovat kunnossa. Seuraava tarkastelu koskee jäännösten normaalisuutta.

Jäännösten normaalisuus

Mallintamisen alkajaisiksi tarkistettiin, että ihannepaino soveltuu jakaumaltaan regressiomallin vastemuuttujaksi. Mallin oletuksen mukainen normaalisuus pääteltiin histogrammista (kuva 5.2, s. 127) ja siihen liittyvästä yhteensopivuustestistä (tuloste 5.1, s. 128). Mallia diagnosoidessa on hyvä tarkistaa, miltä tilanne näyttää jäännösvaihtelun kohdalla.

Kuvassa 5.7 jäännökset on piirretty ”todennäköisyyspaperille”. Historiasta kumpuavalla nimellä tarkoitetaan koordinaatistoa, jossa suuruusjärjestykseen asetetut havaintoarvot, tässä siis mallin jäännökset, piirretään vasten normaalijakauman *kertymäfunktion* arvoja. Sen arvot ovat todennäköisyyksiä, jotka kertyvät nollasta ykköseen, kun jäännösten vaihteluväli käydään läpi.



Kuva 5.7. Regressiomallin jäännökset todennäköisyyspaperilla.

Kuvan 5.7 pystyakseli on skaalattu siten, että jäännösten ja todennäköisyyksien käyräviivainen yhteys kuvautuu normaalisuusoletuksen pätiessä lineaarisena. On helpompi arvioida, osuvatko pisteet suoralle kuin käyrälle viivalle. Mitä paremmin pisteet muodostavat kuvaan suoran, sitä paremmin normaalisuusoletus pätee. Viivan kiemurtelu on merkki oletuksen rikkoutumisesta. Tässä kuvassa ei ole suurta hätää. Aineiston reunoilla on pientä mutkittelua, mikä on aivan tyypillistä, eikä siitä tarvitse olla huolissaan.

Todennäköisyyspaperikuvasta on useita muunnelmia, joiden periaate on sama: testattavan muuttujan jakaumaa verrataan normaalijakaumaan, ja kuvasta arvioidaan, miten hyvin normaalisuusoletus pätee. Kuvalliset keinot ovat hyviä tapoja testata normaalisuutta, sillä kuvasta voi aina havaita jotain sellaista, mikä jää numeerisissa testeissä piiloon.

Tulosteessa 5.10 on esimerkki numeerisesta normaalisuustestistä. Jäännösten normaalisuus ei senkään perusteella jää kyseenalaiseksi ($p = 0.39$).

Tuloste 5.10. Regressiomallin jäännösten normaalisuustestaus.

One-Sample Kolmogorov-Smirnov Test

		Jäännös
N		243
Normal Parameters	Mean	.00000
	Std. Deviation	3.894459
Most Extreme Differences	Absolute	.058
	Positive	.022
	Negative	-.058
Kolmogorov-Smirnov Z		.899
Asymp. Sig. (2-tailed)		.394

s. 207

Viimeisenä regressiodiagnostiikan kohtana tarkastellaan yksittäisten havaintojen vaikutuksia malliin.

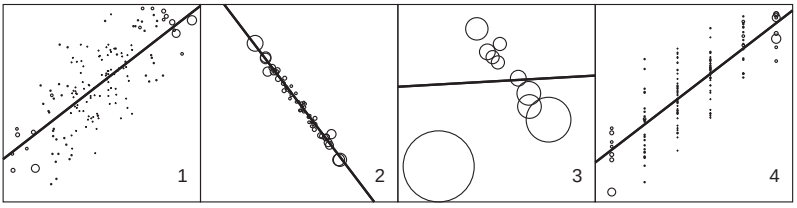
5.3.2 Vaikutusvaltaisuus ja poikkeavuus

Jäännösvaihtelu ei paljastanut mitään epäilyttävää, joten on perusteltua tutkia myös yksittäisten havaintojen vaikutuksia malliin. Keskeiset käsitteet ovat *vaikutusvaltaisuus* ja *poikkeavuus*.

Koska mallien avulla pyritään ilmaisemaan jotakin yleispätevää, ei yksittäisillä havainnoilla saisi olla liikaa vaikutusvaltaa mallin perusteella tehtäviin johtopäätöksiin. Mallit ovat herkkiä myös poikkeaville havainnoille, sellaisille, joiden jäännös on huomattavan suuri. Ongelmia on luvassa ainakin, jos aineistosta paljastuu havaintoja, jotka ovat sekä vaikutusvaltaisia että poikkeavia.

Vaikutusvaltaisuus

Vaikutusvaltaisuus on mallin selittäjien keskinäinen ominaisuus, johon liittyvää tunnuslukua kutsutaan myös *vipuarvoksi*, sillä mitä vaikutusvaltaisempi havainto on, sitä enemmän se voi ”vetää mallia puoleensa”. Viputulkinta avautuu parhaiten kuvasta 5.8. Siihen on koottu luvussa 3 esitellystä, simuloitujen hajontakuvien sarjasta (kuva 3.9, s. 78) ensimmäiset neljä, joissa muuttujien välinen lineaarinen riippuvuus oli selvimmin havaittavissa. Näissä yksinkertaisissa tilanteissa regressiomallia edustaa kuviin piirretty regressiosuora.



Kuva 5.8. Simuloituja regressiokuvia vipuarvoineen.

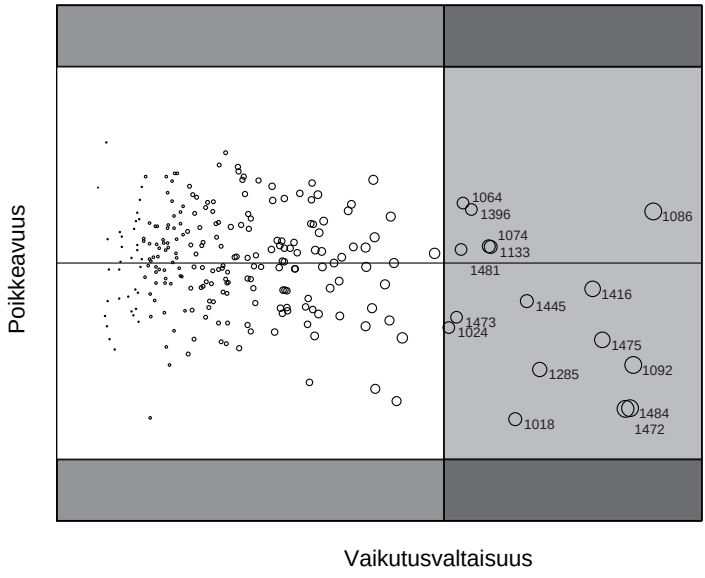
Kuvassa 5.8 ainoa ihmetystä herättävä asetelma on ruutu 3, jossa yksi havainto rikkoo muutoin selvän lineaarisen riippuvuuden. Havaintoja kuvaavien pisteiden koko on verrannollinen havainnon vipuarvoon, joten kyseisellä havainnolla on valtava ”vipuvoima”: se todella vetää

regressiosuoraa puoleensa. Koska havainto on kaukana muiden pisteiden muodostamasta, jyrkästi laskevasta suorasta, se muuttaa koko mallin tulkinnan. Suora kääntyy jopa aavistuksen nousevaan suuntaan.

Vaikutusvaltaisuus ei ole sellaisenaan vaarallista. Se on vain uhka, joka toteutuu, jos havainto ei sovi malliin, toisin sanoen, jos havainnon jäännös on poikkeuksellisen suuri. Silloin seuraa jotain samantyyppistä kuin kuvan 5.8 ruudussa 3. Muissakin ruuduissa on vaikutusvaltaisia havaintoja – viputulkinnan mukaisesti aineiston äärilaidoilla – mutta koska havaintojen jäännökset ovat kohtuullisia, ne eivät aiheuta ongelmia.

Poikkeavuus

Poikkeavuus on hankalampi ongelma, koska suuri jäännös voi johtua yllättävästä, mutta silti aidosta vaihtelusta. Kyseessä voi myös olla aineistoon piiloutunut virhe, joka paljastuu vasta tässä vaiheessa.



Kuva 5.9. Havaintojen vaikutusvaltaisuus ja poikkeavuus.

Vaikutusvaltaisuutta ja poikkeavuutta kannattaa arvioida yhtaikaa. Kuvassa 5.9 niitä kuvaavat arvot on piirretty vastakkain. Numeeriset tiedot on kokonaan häivytetty ja korostettu vain niitä alueita, joihin tulee kiinnittää huomiota. Pisteen koko on verrannollinen havainnon vaikutusvaltaan. Valkoisella alueella ei ole minkäänlaista huolta: siellä olevat havainnot eivät ole vaikutusvaltaisia eivätkä poikkeavia. Mitä harmaammalle alueelle mennään, sitä enemmän harmaita hiuksia alueelle sijoittuvat havainnot aiheuttavat. Tässä kuvassa sellaisia ovat ainoastaan havaintotunnuksin merkityt, vaikutusvaltaisimmat havainnot. Niistä ei tarvitse olla huolissaan, koska ne eivät ole poikkeavia. Havaintotunnusten avulla kyseisiä havaintoja voidaan hieman pitää silmällä.

Jos kuvan 5.9 ylä- ja alareunojen harmaille alueille osuisi havaintoja, ne olisi syytä tutkia tarkemmin ja koettaa selvittää, onko aineistossa virhe. Tilanne on mallin kannalta sitä vakavampi, mitä enemmän oikealle havainnot näillä alueilla sijoittuvat. Kaikkein tummimmissa nurkissa lymyävät havainnot olisivat vaarallisimpia, sillä ne olisivat sekä poikkeavia että vaikutusvaltaisia (vrt. kuvan 5.8 ruutu 3).

Kuvan 5.9 tummimmilla alueilla on harkittava jopa havainnon poistamista analyysista, mutta sitä ei saa liian kevyin perustein mennä tekemään. Olisi liian helppoa, jos kaikki ”kiusallisen poikkeavat” havainnot vain napsittaisiin pois; mallintaminen muuttuisi mielivaltaiseksi aineiston manipuloinniksi, eikä malli enää kuvaisi edes aineistoa, todellisuudesta puhumattakaan. Tällä kertaa nurkat ovat onneksi tyhjillään, joten voidaan todeta, että havaintodiagnostiikka ei paljasta mitään huolestuttavaa, ja laadittu regressiomalli toimii oikein.

6 Aineiston ryhmittely

Edellisen luvun tarkastelut osoittivat, kuinka tiivistetyllä aineistolla päästään kiinni havaintojen vertailuun ja mallintamiseen regressioanalyysilla. Tutkimuskysymyksistä riippuu, mitä muuta kannattaa tehdä. Tässä luvussa perehdytään aineiston ryhmittelyyn, joka on yksi monista mahdollisuuksista jatkaa ja syventää analyyseja. Ryhmittelyä kutsutaan usein myös *klusteroinniksi*.

Tarkasteltavilla menetelmillä voidaan ryhmitellä niin muuttujia kuin havaintoja. Havaintojen ryhmittely on tyypillisempää. Muuttujien ryhmittelyyn soveltuu paremmin luvussa 4 läpikäyty faktorianalyysi, jos sen oletukset vain täyttyvät. Toisin kuin faktori- ja regressioanalyysi, ryhmittelymenetelmät eivät perustu tilastollisiin malleihin, vaan niissä edetään laskennallisten perusteiden ja kokeilujen avulla. Menetelmien oletukset ovat myös jossain määrin väljempiä.

Tässä luvussa käsitellään menetelmiä, jotka soveltuvat sekä muuttujien että havaintojen ryhmittelyyn. Menetelmistä kertoo tarkemmin muun muassa kirja *Finding Groups in Data: An Introduction to Cluster Analysis* (Kaufman & Rousseeuw, 1990).

6.1 Hierarkkinen ja visuaalinen ryhmittely

Aineiston ryhmittely on aikamoista seikkailua, jossa ei tiedä, mihin päätyy. Sitä voisi jopa luonnehtia ”sokeaksi hapuiluksi” aineiston seassa. Parhaimmillaan voidaan saada aikaan onnistuneita ryhmityksiä, huonoimmillaan pelkkää sekasotkua.

Ryhmittelyn haasteellisuudesta saa alustavan käsityksen kuvasta 6.1, jossa on *hierarkkisen ryhmittelyn* lopputulos vaakasuuntaisena puukuviona. Jotta kuvion yksityiskohdat olisivat luettavissa, on ulkonäkötutkimuksen aineistoa tässä rajattu niin, että mukana on vain 50 satunnaisesti valittua havaintoa vuodelta 1997.

Kuvan 6.1 oikeassa reunassa ovat havaintojen tunnisteet. Ryhmittelyn voi ajatella alkavan sieltä, tilanteesta, jossa kaikki havainnot muodostavat oman ryhmänsä. Vasemmalle päin siirryttäessä havaintoja yhdistellään uusiksi ryhmiksi tai niitä liitetään mukaan jo luotuihin ryhmiin. Vaakasuuntaisten viivojen pituudet ovat suhteessa käytettyyn etäisyysmittaan, josta puhutaan tarkemmin tuonnempana.

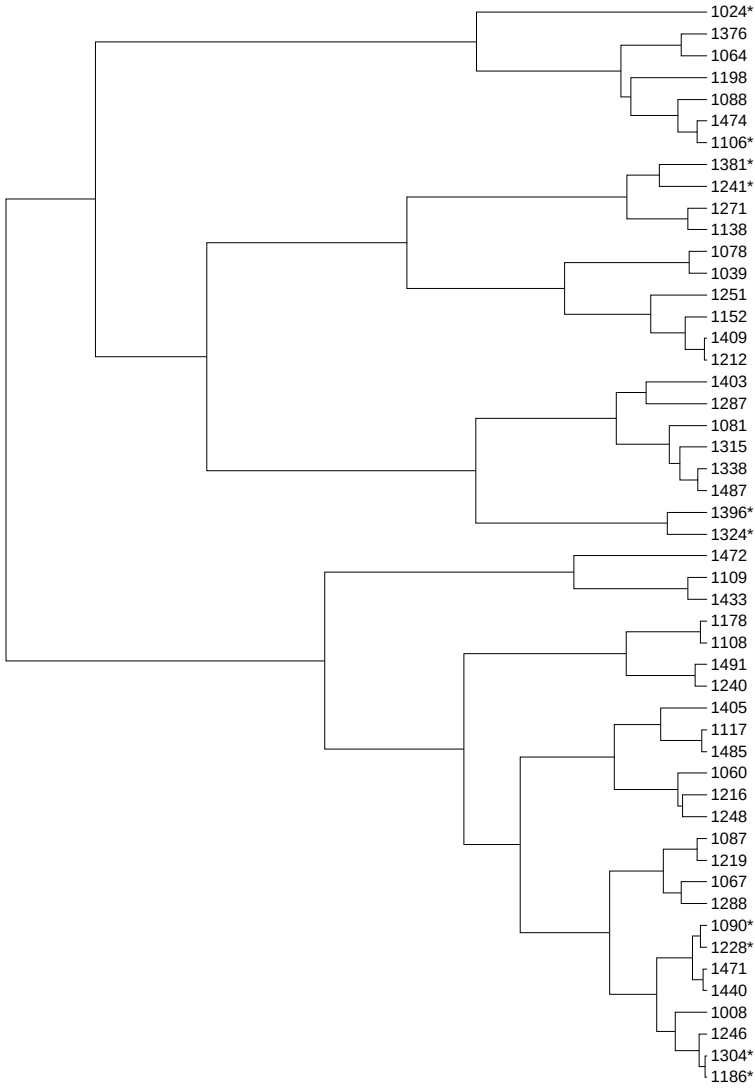
Kun tullaan kuvan 6.1 vasempaan reunaan, kaikki havainnot ovat samassa ryhmässä. Yhtä hyvin voitaisiin toimia käänteisessä järjestyksessä: lähteä vasemmalta, alkaa erottaa havaintoja ryhmistä ja päätyä oikeaan reunaan. Kiinnostavaa ei välttämättä ole kumpikaan lopputuloksista, vaan niiden välinen alue, puun ”oksisto”. Voidaan kysyä esimerkiksi: ”*Montako ryhmää pitäisi muodostaa?*”, ”*Miten suuria ryhmät ovat?*” ja ”*Mitä havaintoja ryhmiin kuuluu?*”.

Yksikäsitteisiä vastauksia ei saada, sillä ryhmittelyn tulos riippuu siitä, minkä muuttujien perusteella havaintoja ryhmitellään. Tässä on käytetty faktoripistemuuttujia (ks. taulukko 4.1, s. 110), joten ryhmittelyn tulkintaan on ainakin jonkinlainen mahdollisuus. Alkuperäisillä muuttujilla ei kannata ryhmittelyyn ryhtyä; syyt ovat samat kuin regressioanalyysin yhteydessä mainitut (ks. kohta 5.2.2, s. 128).

Havaintojen profiilit

Ennen kuin katsotaan, mistä muista asioista ryhmittelytulos riippuu, tarkastellaan taulukkoa 6.1 (s. 154). Siinä on kymmenen, kuvassa 6.1 tähdellä (*) merkityn, havainnon faktoripisteiden arvot. Tiedot on taulukossa lueteltu samassa järjestyksessä kuin kuvassa. Havainnot 1304 ja 1186, jotka kuvassa näkyvät alimmaisina, on ryhmittelyssä yhdistetty ensimmäiseksi. Ne ovat *profiileiltaan* samantyyppisiä, minkä voi päätellä myös taulukon luvuista.

Vastaavasti keskenään samantyyppisiä ovat havainnot 1090 ja 1228. Loput taulukon 6.1 havainnoista ovat kuvan 6.1 yläosasta, josta voi hahmottaa yhdestä kolmeen ryhmää. Yläosan suurimmasta ryhmästä on valittu sen kaksi alinta havaintoa, 1324 ja 1396, jotka taas ovat keskenään melko samanlaisia, tosin eivät yhtä samanlaisia kuin



Kuva 6.1. Osa-aineiston hierarkkinen ryhmittely.

aiemmat, joten ne on yhdistetty vasta hieman myöhemmässä vaiheessa. Ylempää on poimittu vielä neljä havaintoa, joista numero 1024 vaikuttaa profiililtaan erilaiselta kuin muut.

	Faktoripistemuuttujat			
Tunnus	TUNTO	PANOS	PAINE	NEGAT
1024	–1.55	1.80	0.90	–1.89
1106	–2.06	0.15	1.05	–1.01
1381	0.37	–0.27	0.51	2.37
1241	–0.41	–0.19	–0.76	2.24
1396	0.34	0.40	–2.82	1.03
1324	0.42	0.17	–1.94	0.00
1090	0.91	–0.99	0.23	–0.39
1228	0.86	–1.00	0.77	–0.28
1304	0.28	0.18	0.32	–0.39
1186	0.35	0.41	0.40	–0.33

Taulukko 6.1. Kuvan 6.1 tähdellisten havaintojen arvoja.

Taulukoissa on tapana esittää lukuarvoja parin desimaalin tarkkuudella, mutta vertailuissa on hyvä muistaa mittauksen keskivirhe (ks. taulukko 4.2, s. 119). Yksi desimaali riittäisi useimmiten mainiosti, niin myös taulukossa 6.1. Olennaiset erot näkyvät itse asiassa jo, kun luvut pyöristetään lähimpään kokonaislukuun. Taulukko 6.2 esittää samat tiedot pelkkinä symboleina. Näin tarkastellut erot ylittävät ainakin mittausvirheistä johtuvan vaihtelun, joten voidaan luottavaisemmin puhua todellisista eroista.

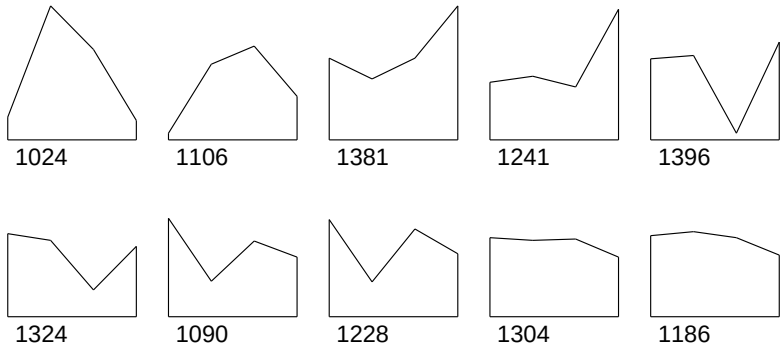
Profiilikuva

Piirtäminen on paras keino tarkastella profiileja, kuten havaittiin jo keskiarvoprofiilien yhteydessä (kuva 5.3, s. 141). Taulukossa 6.1 esitetyt luvut on piirretty kuvaan 6.2 yhdistämällä faktoripisteiden arvot murtoviivoiksi. Näin jokaiselle havainnolle muodostuu sitä ilmentävä profiili. Kuvasta voidaan helposti todeta samat yhtäläisyydet, joita edellä sanallisesti selitettiin.

Varsinkin pienemmillä aineistoilla voi olla kätevää ryhmitellä havaintoja myös *visuaalisesti*, erilaisten profiilikuvien avulla. Kuvat voi tulostaa paperille, leikata irti ja järjestellä pinoihin; lopputulos on usein aivan yhtä hyvä kuin ryhmittelymenetelmillä. Tässä yhteydes-

	Faktoripistemuuttujat			
Tunnus	TUNTO	PANOS	PAINE	NEGAT
1024	− −	+ +	+	− −
1106	− −	.	+	−
1381	.	.	+	+ +
1241	.	.	−	+ +
1396	.	.	− − −	+
1324	.	.	− −	.
1090	+	−	.	.
1228	+	−	+	.
1304	.	.	.	.
1186	.	.	.	.

Taulukko 6.2. Taulukon 6.1 tiedot symboleina.

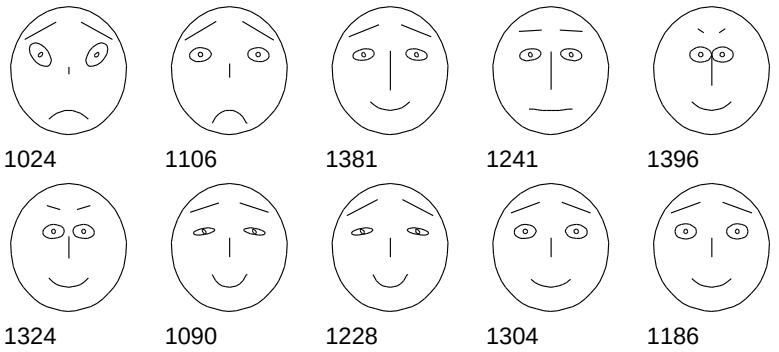


Kuva 6.2. Kuvan 6.1 tähdellisten havaintojen profiilikuva.

sä on tilan säästämiseksi tyydytty vain 50 havainnon ryhmittelyyn ja kymmenen havainnon tarkempaan kuvailuun, mutta menettelytavat sopivat käytännössä myös suurempien aineistojen työstämiseen. Tuhansien havaintojen kanssa lienee jo parempi tiivistää ensin aineistoa myös havaintojen suhteen.

Naamakuva

Luultavasti eksoottisin tilastollisen aineiston kuvaustapa tunnetaan nimellä *Chernoffin naamat* (Chernoff, 1973). Siinä havaintojen profiilit kuvataan sarjakuvamaisina naamoina, joiden piirteet, kuten pään koko, suun kaarevuus, nenän pituus ja silmien asento, vaihtelevat muuttujien arvojen perusteella. Kuvassa 6.3 on profiilikuvaa 6.2 vastaava naamakuva, jossa faktoripisteet ”muovaavat” naamoille ilmeitä. Samanaikaisesti voi vaikuttaa 18:aan piirteeseen, mutta tässä on tyydytty seitsemään ja jätetty suurin osa ”peruslukemille”.



s. 208

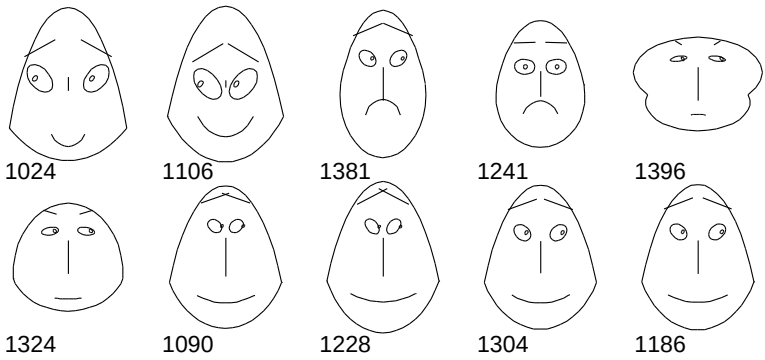
Kuva 6.3. Kuvan 6.1 tähdellisten havaintojen naamakuva.

Myös naamakuvan avulla aineistoa voi ryhmitellä ja ehkä vielä helpommin tunnistaa mahdollisia poikkeavuuksia. Osa piirteistä on puhuttelevampia kuin toiset. Esimerkiksi suun kaarevuus on kuvassa 6.3 kytketty itsetuntoa kuvaavaan faktoripistemuuttujaan, joka taulukon 6.1 (s. 154) kahdella ensimmäisellä havainnolla on pienin. Vastaavasti ulkonäköön panostamisessa, joka visuaalisesti ilmenee silmien asentona ja kokona, on suurin pistemäärä havainnolla 1024, ja pienin havainnoilla 1090 ja 1228.

Nenän pituus ja kulmakarvojen asento eivät liene ensimmäiseksi tunnistettavia piirteitä, mutta tarkemmalla tutkiskelulla niissä voi havaita erilaisuuksia tai samankaltaisuuksia. Tässä mielessä naamakuva täyttää hyvän tilastollisen kuvan kriteerin: yhdellä vilkaisulla näkee jotain olennaista, ja lähempi tarkastelu paljastaa lisää yksityis-

kohtia. Naamakuvista on hyötyä myös aikasarjojen tarkastelussa, sillä ilmeiden muuttumisesta on helppo havaita ajassa tapahtuvia, monimutkaisiakin muutoksia.

Koska muuttujia voi yhdistää piirteisiin mielivaltaisesti, samasta aineistosta voi piirtää useita, aivan erinäköisiä naamakuvia. Tässä suhteessa naamakuva ei eroa varsinaisista ryhmittelymenetelmistä. Yksi variaatio näkyy kuvassa 6.4.



Kuva 6.4. Kuvan 6.3 naamakuvan muunnelma.

Piirteiden valinnaisuuden vuoksi naamakuvaa saatetaan pitää liian subjektiivisena, mutta valintoja tutkija joutuu kuitenkin tekemään. Tärkeintä on, että kuvien ja muiden tulosten perusteella tehdyt johtopäätökset ovat totuudenmukaisia. Tämä on keskeinen *tutkimusetiikan* periaate. Toinen tutkimuseettinen periaate koskee aineiston *anonymisointia*, jota käsittelee Kuula (2006, 200–213). Naamakuvissa ja muissa edellä piirretyissä kuvissa havaintoja on merkitty niiden nelinumeroisilla tunnisteilla. Koodit ovat yksikäsitteisiä, mutta anonyymeja lukuja, joista tutkimukseen osallistuneita naisia ei voi tunnistaa.

Naamakuva on kätevä työkalu aineiston visualisointiin. Yksittäisten vastausprofiilien ohella naamoilla voidaan kuvata minkä tahansa ryhmien profiileja esimerkiksi niiden keskiarvojen perusteella. Naamakuvista sekä muista moniulotteisen aineiston erikoiskuvista kertoo lisää *Survo ja minä* -kirja (Mustonen, 1996, 139–159).

Läheisyys ja etäisyys

Ryhmittelyn lopputulos riippuu olennaisesti muuttujista. Lisäksi siihen vaikuttaa ryhmittelymenetelmä. Edellä tehdyssä hierarkkisessa ryhmittelyssä on useita vaihtoehtoja riippuen muun muassa siitä, millä tavoin havaintoja yhdistetään ryhmiin.

Ryhmittelyn perustapa on ”lähimmän naapurin” menetelmä, jossa etäisyyttä ryhmään mitataan vain sen lähimmän jäsenen perusteella. Menetelmä ei yleensä anna tyydyttävää tulosta, koska ryhmät ketjuuntuvat helposti ja niitä voi olla vielä vaikeampi erottaa toisistaan kuin luvun alussa (s. 153) esitetystä kuvasta 6.1. Kuva perustuu päinvastaiseen, ”etäisimmän naapurin” menetelmään, jossa ryhmien mahdolliset erot tulevat selkeämmin näkyviin. On paljon muitakin menetelmiä, eikä yksikään ole tilastollisesti toista parempi. Tyypillisesti ryhmittelyjä kokeillaan useilla menetelmillä pyrkien mahdollisimman helposti tulkittavaan lopputulokseen.

Tulokset riippuvat myös *etäisyysmitasta*, joka määrää, millä periaatteella havaintoja vertaillaan ryhmittelyn eri vaiheissa. Tavallisen, euklidisen etäisyyden lisäksi on käytettävissä muitakin, esimerkiksi sen neliö, joka korostaa eroja selvemmin. Sitä on sovellettu myös kuvan 6.1 ryhmittelyssä.

Kaikki tämän luvun menetelmät pohjautuvat havaintojen tai muuttujien välisiin etäisyyksiin. Analyysien lähtökohta on *etäisyysmatriisi*, jossa on taulukkomuotoon koottuna kaikki parittaiset etäisyydet. Tulosteessa 6.1 on edellä käsiteltyjen kymmenen havainnon etäisyysmatriisi, jossa etäisyysmittana on käytetty niin sanottua *City Block* -etäisyyttä. Nimensä mitta on saanut siitä, että ”kulkeminen” havainnosta tai ”paikasta” toiseen tapahtuu amerikkalaisen kaupungin ruutukaavan mukaisesti; talojen läpi ei siis voi oikaista. Euklidinen etäisyys vastaisi kulkemista linnuntietä. Tulosteessa on hahmottamisen helpottamiseksi korostettu suurimpia etäisyyksiä.

Etäisyysmatriisi muistuttaa luvussa 3 käsiteltyä korrelaatiomatriisia (tuloste 3.11, s. 79), jossa luvut kuvasivat muuttujien välisiä lineaarisia yhteyksiä. Nyt tarkastelun kohteena ovat havaintojen väliset etäisyydet. Korrelaatiomatriisin lävistäjä koostui ykkösistä, mutta etäisyysmatriisin lävistäjä koostuu nollista – ”havainnon etäisyydestä itseensä”. Myös korrelaatiota voidaan soveltaa etäisyysmittana vähentämällä sen arvot ykkösestä.

Tuloste 6.1. Kuvan 6.1 tähdellisten havaintojen etäisyysmatriisi.

```
     1024 1106 1381 1241 1396 1324 1090 1228 1304 1186
1024 0.00 3.19 8.64 8.92 9.93 8.33 7.42 6.95 5.53 5.35
1106 3.19 0.00 6.77 7.05 8.56 6.50 5.55 5.08 3.72 4.00
1381 8.64 6.77 0.00 2.26 5.37 5.31 4.30 4.13 3.49 3.51
1241 8.92 7.05 2.26 0.00 4.61 4.61 5.74 6.13 4.77 5.09
1396 9.93 8.56 5.37 4.61 0.00 2.22 6.43 6.82 4.84 4.60
1324 8.33 6.50 5.31 4.61 2.22 0.00 4.21 4.60 2.80 2.98
1090 7.42 5.55 4.30 5.74 6.43 4.21 0.00 0.71 1.89 2.19
1228 6.95 5.08 4.13 6.13 6.82 4.60 0.71 0.00 2.32 2.34
1304 5.53 3.72 3.49 4.77 4.84 2.80 1.89 2.32 0.00 0.44
1186 5.35 4.00 3.51 5.09 4.60 2.98 2.19 2.34 0.44 0.00
```

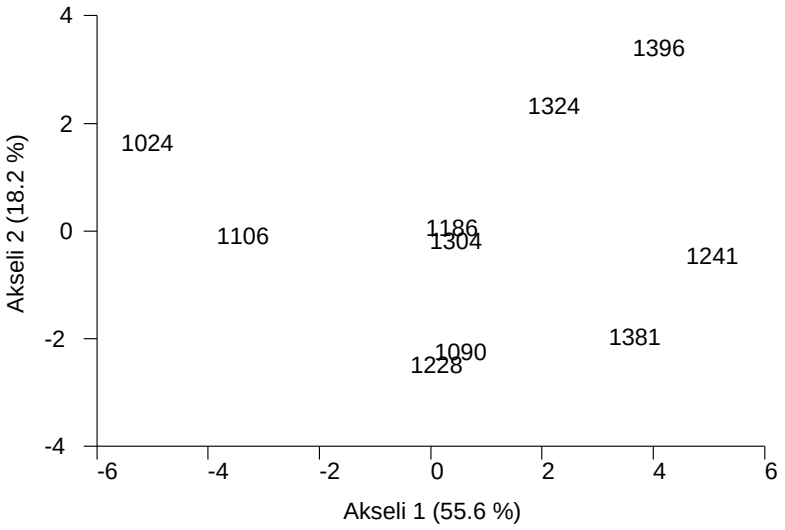
Usein etäisyysmatriisit ovat symmetrisiä, mutta mikään ei estä ajattelemasta etäisyyksiä epäsymmetrisinä – pitää vain vähän venyttää käsitystä sanasta ”etäisyys”. Maantieteellinen on vain yksi vaihtoehto; etäisyyksiä voi määritellä lukemattomilla tavoilla. Esimerkiksi kahden paikan välinen etäisyys voidaan määritellä matkaan kuluvana aikana tai lentolipun hintana, jolloin etäisyydet ovat heti epäsymmetrisiä.

6.2 Moniulotteinen skaalaus

Toisenlaisen tavan aineiston kuvailuun ja havaintojen ryhmittelyyn tarjoaa menetelmä nimeltään *moniulotteinen skaalaus*. Siinä etäisyysmatriisin asetelma pyritään puristamaan kaksiulotteiseksi hajontakuvaksi, jossa havainnot sijoittuvat toisiinsa nähden siten, että kaikki niiden väliset etäisyydet pätevät. Tavoite ei ole aivan helppo, ja ristiriitaisuuksilta ei voida välttyä. Aineiston ei edes tarvitse olla kovin suuri tai monimutkainen.

Ratkaisua voi tarkentaa lisäämällä ulottuvuuksia, mutta se kostautuu visualisoinnin vaikeutena. Joskus saattaa olla perusteltua tarkastella skaalausta kolmessa ulottuvuudessa, mutta useimmiten tyydytään kahteen. Olennaista ei ole etäisyysmatriisin mahdollisimman täydellinen kuvaaminen, vaan yleiskäsitysten saaminen havaintojen välisistä suhteista.

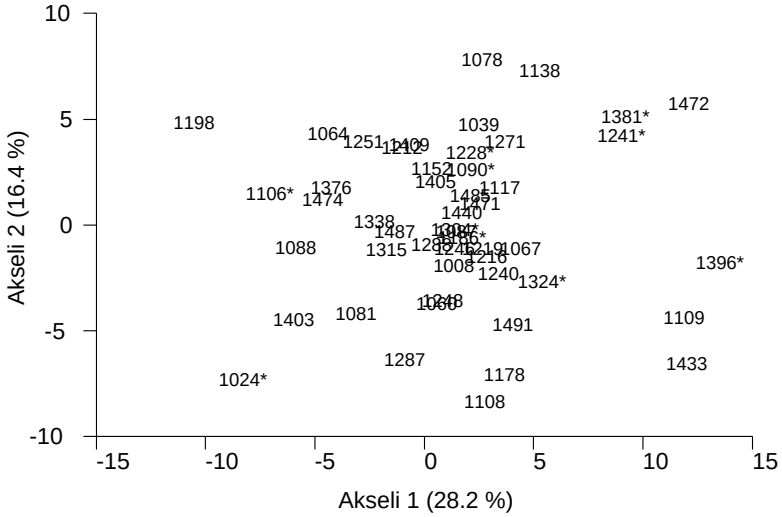
Kuvassa 6.5 on tehty moniulotteinen skaalaus edellä käsitellystä kymmenen havainnon etäisyysmatriisista (tuloste 6.1). Kaksi ensimmäistä ulottuvuutta, jotka tässä on piirretty vastakkain, selittävät etäisyyksien vaihtelusta noin 74 %. Ensimmäinen selittää jo yli puolet. ”Täydellinen” kuvaus vaatisi tässä pienoistapauksessa ainakin kuusi ulottuvuutta, joiden kuvaamiseen tarvittaisiin enemmän kuvia kuin olisi järkevää piirtää. Monimuuttujamenetelmille on tyypillistä, että keskitytään vain tärkeimpiin ulottuvuuksiin.



Kuva 6.5. Kuvan 6.1 tähdellisten havaintojen skaalaus.

Edellä samankaltaisiksi todetut havainnot 1186 ja 1304 asettuvat aika tarkalleen kuvan 6.5 keskelle, ja poikkeavin havainto 1024 vasempaan reunaan. Etäisyysmatriisissa (tuloste 6.1) suurimmat etäisyydet ovat sen ja oikean reunan kolmen havainnon välillä. Kuva välittää etäisyysmatriisin olennaisimmat tiedot.

Kuvassa 6.6 on vastaavanlainen skaalaus, jossa ovat mukana kaikki hierarkkisessa ryhmittelyssä (kuva 6.1, s. 153) käytetyt 50 havaintoa. Myös etäisyysmitta on sama kuin siinä käytetty, siis neliöllinen euklidinen etäisyys. Tähdelliset havainnot erottuvat joukosta, ainakin profiileiltaan hieman tavallisesta poikkeavammat. Niiden lisäksi nousee esiin muitakin, kuten vasemman reunan numero 1198.



Kuva 6.6. Kuvan 6.1 havaintojen skaalaus.

Vertailemalla kuvia 6.6 ja 6.1 (s. 153) huomaa ryhmittelyn haastavuuden ja monikäsitteisyyden. Puukuviosta on hahmottuvinaan useitakin ryhmiä, hajontakuvasta ei. Kuvien välillä on sekä yhteneväisyyksiä että eroavaisuuksia. Tietojen puristaminen kahteen ulottuvuuteen ei ole helppoa, mikä näkyy myös kuvan 6.6 akselien alemmissa selitysosuuksissa.

Yhtä hyvin muuttujia kuin havaintoja

Vastapainoksi sille, että ryhmittelymenetelmät ovat kuin hapuilua aineiston syövereissä, ne ovat myös varsin joustavia menetelmiä. Ne eivät välitä siitä, ryhmitelläänkö havaintoja vai muuttujia; aineistoa voi ryhmitellä miten päin haluaa. Periaatteessa luvussa 4 kuvattu aineiston tiivistäminen voitaisiin tehdä ryhmittelemällä muuttujia jollain tässä luvussa kuvatulla menetelmällä, mutta näin ei kannata menetellä. Mittausmallin perusteella tehty faktorianalyysi on luotettavampi tapa huolehtia aineiston tärkeimmästä tiivistysvaiheesta.

Faktorianalyysi ei silti sovellu kaikkien mittausten tiivistämiseen, koska se edellyttää numeerista mittaustasoa. Esimerkki 2.4 (s. 29) on mittaustasoltaan kaukana faktorianalyysissa vaaditusta. Koska vastaaja on saanut valita annetuista vaihtoehdoista niin monta kuin on halunnut, tiedot on pitänyt koodata aineistossa peräti 74 muuttujaan. Kunkin muuttujan arvo on ykkönen, jos kyseinen luonnehdinta on valittu, muuten nolla. Dikotomiset mittaukset täyttävät periaatteessa korrelaatioiden laskemisen ehdot, mutta eivät ne siihen kovin hyvin sovellu. Sen sijaan ryhmittelymenetelmät toimivat hyvin, koska etäisyyksiä voidaan mitata myös luokittelutasoisista tiedoista.

	1	0
1	a	b
0	c	d

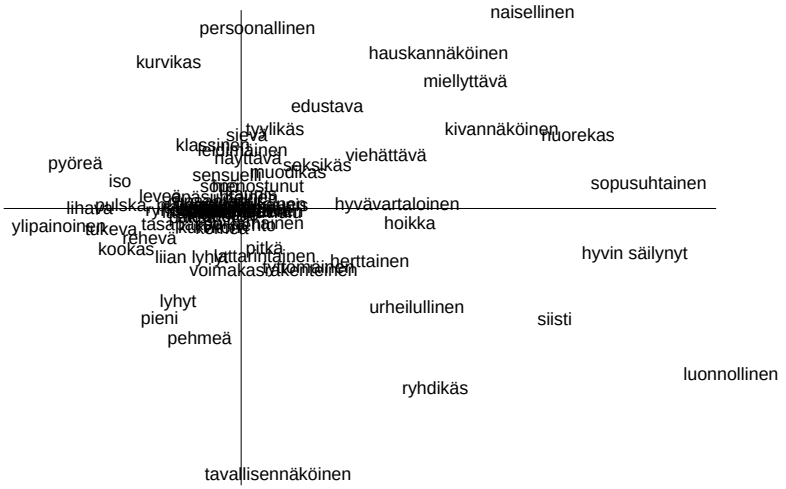
Taulukko 6.3. Kahden luokittelutason tiedon vertailukaavio.

Jatkuville muuttujille sopivia etäisyysmittoja on paljon, mutta luokittelutasolle soveltuvia on vielä enemmän. Ne kaikki perustuvat taulukossa 6.3 esitettyyn kaavioon. Siinä vertaillaan kahta havaintoa tai muuttujaa, jotka saavat vain arvoja 0 tai 1. Vertailutuloksia on merkitty kirjaimin a, b, c ja d. Tuloksena on a, jos molemmat saavat arvon 1. Päinvastaisessa tapauksessa tuloksena on d. Vertailun mennessä jomminkummin päin ristiin saadaan b tai c.

Kun vertailu on suoritettu läpi aineiston, saadaan taulukko tulosten a, b, c ja d lukumääristä. Etäisyysmitta muodostetaan sen jälkeen

jollakin kymmenistä mahdollisista lausekkeista, joissa käytetään näitä kirjaimia. Tyypillinen mitta on sellainen, jossa lasketaan tulosten a ja d suhteellinen osuus kaikista vertailuista.

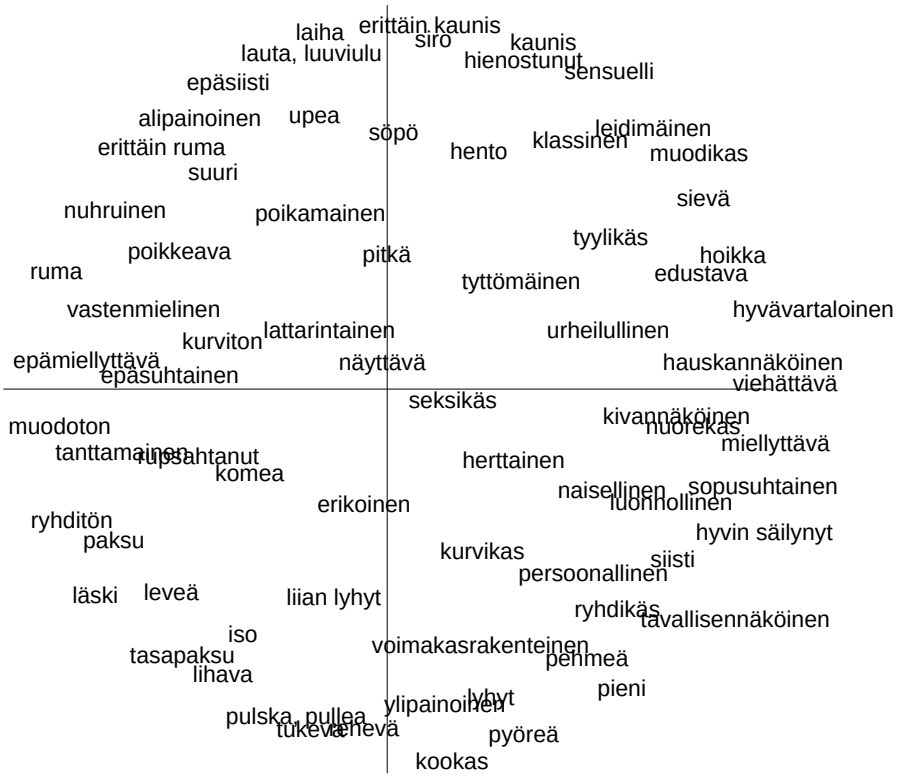
Esimerkin 2.4 (s. 29) tilanteessa voidaan vertailla luonnehdintoja kuvaavia muuttujia toisiinsa havaintojen, siis vastausten, perusteella. Nyt pyrkimyksenä on tiivistää aineistoa muuttujien suunnassa ja päästä eroon tilaa vievistä 74 dikotomiasta. Etäisyysmittana on käytetty edellä mainittua a- ja d-tapausten suhteellista osuutta. Etäisyysmatriisille on tehty *klassinen moniulotteinen skaalaus*, jota on vielä tarkennettu *pienimmän neliösumman skaalauksella*. Näitä menetelmiä käytettiin myös aiemmissa skaalausesimerkeissä (kuvat 6.5 ja 6.6). Menetelmistä kertoo tarkemmin Mustonen (1995, 148–170).



Kuva 6.7. Ulkonäön sanallisten luonnehdintojen skaalaus.

Tuloksena on kuva 6.7, josta on jätetty numeeriset tiedot kokonaan pois. Jäljellä ovat vain kaksi koordinaattiakselia ja sanalliset luonnehdinnat, jotka ryhmittyvät kahteen ulottuvuuteen. Vaakasuuntaisen ulottuvuuden voisi kenties nimetä ”sopusuhtaisuuden – ylipainoisuuden” akseliksi, pystysuuntaisen kenties ”tavallisuuden – persoonallisuuden” akseliksi. Alun perin karkeasti mitatuista käsityksistä on siis saatu ainakin alustavasti kaksi jatkuvaa ulottuvuutta.

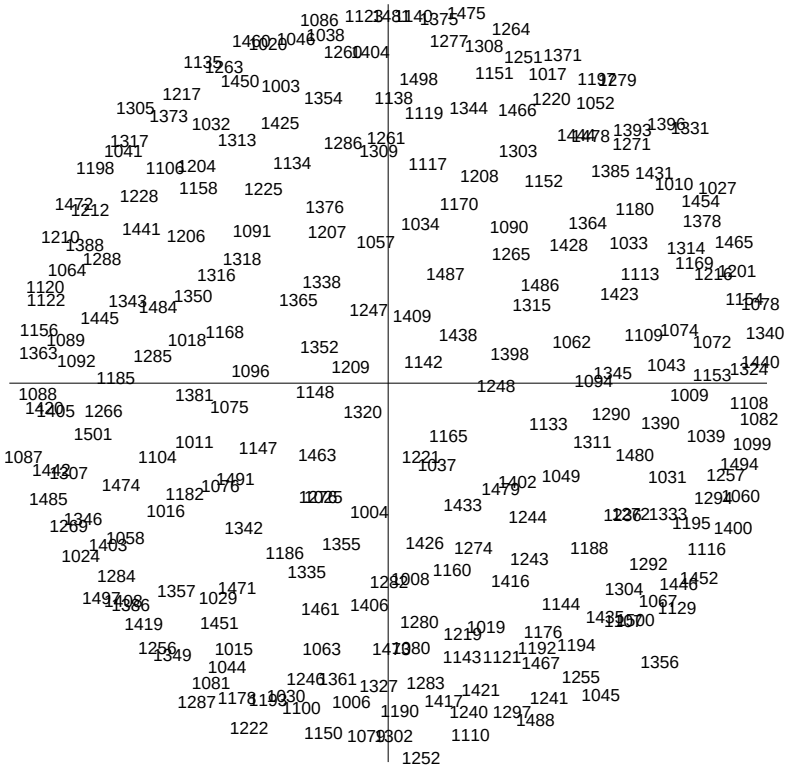
Muunnelma kuvasta 6.7 on kuva 6.8. Ainoa ero on etäisyysmitta. Luvut a, b, c ja d on laskettu samalla tavalla, mutta mitta on muodostettu ilman d-tapauksia, siis niitä, joissa vastaaja ei ole rastittanut kumpaakaan vertailtavista sanoista. Painoa saavat enemmän ne, joissa on paljon yhteisiä rastituksia. Sanat ryhmittyvät koordinaatistoon ympyrän muotoon. Useammin yhdessä esiintyvät luonnehdinnat ovat lähempänä toisiaan. Osin kuvat muistuttavat toisiaan, mutta jälkimmäinen on ehkä vielä selkeämmin tulkittavissa johtuen siitä, että sanat ovat paremmin näkyvillä.



Kuva 6.8. Sanallisten luonnehdintojen toinen skaalaus.

Koska sanalliset luonnehdinnat ryhmittyvät aika siististi, voisi niitä esittäneet *luonnehtijat* yrittää ryhmitellä vastaavasti. Käännetään siis

tarkastelu havaintojen ryhmittelyyn, mutta vertaillaan samoja nollia ja ykkösiä kuin edellä. Painottamalla jälleen pelkkiä a-tapauksia saadaan kuva 6.9, jossa havainnot ryhmittyvät ympyrän muotoon samaan tapaan kuin kuvassa 6.8.



Kuva 6.9. Luonnehdintoja esittäneiden skaalaus.

Ulottuvuuksia vastaavia muuttujia voidaan hyödyntää jatkoanalyyseissa. Esimerkiksi voidaan tehdä regressioanalyysi, jossa selitetään itsetuntoa näillä sanallisista arvioista muodostetuilla pisteillä. Yksityiskohdat sivuutetaan, mutta arviot selittäisivät itsetuntoa kuvaavasta faktoripistemuuttujasta jopa 34 %, joten asiaa voisi tutkia tarkemmin. Ainakin sanallisilla luonnehdinnoilla näyttää olevan selviä yhteyksiä varsinaisilla mittareilla mitattuihin ulottuvuuksiin.

6.3 Medoidiryhmittely

Edellä käsitellyissä menetelmissä ryhmien muodostamista enemmän painottuvat aineiston kuvailu ja mahdollisten poikkeavien havaintojen tunnistaminen. Ryhmien tarkempi hahmottaminen jää tutkijalle, eikä tehtävä ole välttämättä helppo, etenkään suuremmilla aineistoilla.

Kun halutaan vain muodostaa tietty määrä ryhmiä, toisin sanoen *osittaa* aineisto, sovelletaan yleensä keskiarvoryhmittelyä (*k-means*). Seuraavassa tutustutaan periaatteiltaan samantapaiseen *medoidiryhmittelyyn* (*k-medoids*), jonka erikoinen nimi viittaa mediaanin yleistyksiin. Toteutukseltaan medoidiryhmittely on keskiarvoryhmittelyä *robustimpi*, toisin sanoen se ”sietää” poikkeavia havaintoja paremmin.

Medoidiryhmittelyä tehdään paljolti samaan tapaan kuin muitakin ryhmittelyjä. Aluksi havainnoista muodostetaan etäisyysmatriisi jonkin etäisyysmitan suhteen. Tässä yhteydessä on käytetty samaa mittaa kuin kohdan 6.1 hierarkkisessa ryhmittelyssä, siis neliöllistä euklidista etäisyyttä. Muuttujia ei edelleenkään kannata olla liikaa, jotta pysyy selvillä siitä, mihin ryhmittely perustuu. Medoidiryhmittely on seuraavassa tehty kahdella tavalla: ensimmäinen perustuu vain faktoripisteisiin, ja toisessa otetaan mukaan samat taustamuuttujat, paino ja ikä, joita käytettiin luvun 5 regressioanalyysissa.

Molemmissa ryhmittelyissä muodostetaan vuoden 1997 aineistosta *neljä ryhmää*. Ryhmien lukumäärä ei yleensä ole itsestään selvää kuten havaittiin jo hierarkkisen ryhmittelyn yhteydessä. Jollain tavoin ”oikean” määrän löytäminen voi vaatia paljon kokeiluja ja ryhmien perusteellisempaa tutkimista, esimerkiksi taulukointia aineiston muita muuttujia vastaan, mikä tässä yhteydessä sivuutetaan. Ryhmiä tarkastellaan enemmän seuraavassa luvussa, mutta todellisuudessa nämä vaiheet vuorottelevat, kunnes tyydyttävä ratkaisu löydetään – jos löydetään.

Faktoripisteet ryhmittelymuuttujina

Lähtökohtana on siis samalla tavalla muodostettu etäisyysmatriisi kuin hierarkkisessa ryhmittelyssä, mutta nyt se on kooltaan suurempi, sillä aineistoa ei ole tarvetta rajata keinotekoisesti kuten tämän luvun alussa tehtiin.

Medoidiryhmittely neljään ryhmään tiivistyy tulosteeseen 6.2, joka kertoo ryhmien keskipisteiksi eli *medoideiksi* valittujen havaintojen järjestysnumerot, ryhmien koot sekä eräitä muita tietoja. Ryhmissä 1 ja 3 on molemmissa vähän alle 80 havaintoa, ryhmissä 2 ja 4 puolestaan 57 kummassakin. Tulosteen oikeaan reunaan on lisätty medoidihavaintojen profiilit ryhmittelymuuttujien suhteen (vrt. taulukko 6.1, s. 154). Profiilit ovat varsin erilaisia, sillä menetelmä valitsee medoidit mahdollisimman erilaisiksi, jotta niiden ympärille rakennettavat ryhmät erottuisivat toisistaan.

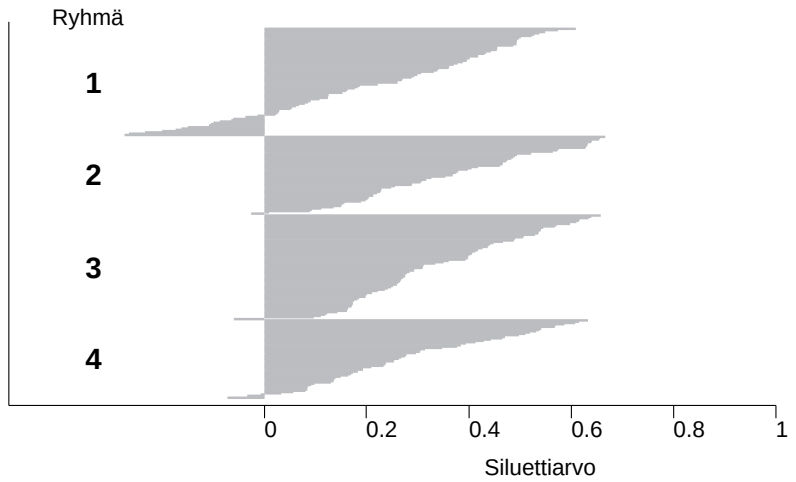
Tuloste 6.2. Medoidiryhmittely faktoripistemuuttujilla.

```
Cluster analysis by medoids of Kaufman and Rousseeuw (1987)
Data UNRYHM  N=268
Group Medoid       n Mean             TUNTO PANOS PAINE NEGAT
 1    112         78 0.218            -0.94 -0.03  0.40  0.40
 2    89          57 0.363             0.75  1.08  0.39 -0.69
 3    71          76 0.344             0.81 -0.76  0.24  0.10
 4    124         57 0.289             0.08  0.04 -1.41 -0.03
Mean of all silhouette values is 0.3
```

Ryhmittely ei ole kuitenkaan onnistunut kovin hyvin, minkä voi päätellä jo tulosteen 6.2 perusteella niin kutsuttujen *siluettiarvojen* ryhmäkeskiarvoista. Havainnon siluettiarvo kertoo, miten hyvin ryhmittely on sen osalta onnistunut. Mitä selvemmin havainto kuuluu ryhmään, sitä lähempänä sen siluettiarvo on ykköstä. Jos arvo on lähellä nollaa, se on merkki siitä, että havainto on jossain kahden ryhmän ”välissä”. Siluettiarvo voi myös olla negatiivinen, jolloin havainto on menetelmän mielestä todennäköisesti väärässä seurassa.

Tulosteessa 6.2 siluettiarvojen ryhmäkohtaiset keskiarvot vaihtelevat 0.2:n ja 0.4:n välillä, ja niiden keskiarvo koko aineistossa on 0.3. Tunnusluvut tiivistävät tiedon tehokkaasti, mutta piilottavat mielenkiintoiset yksityiskohdat, niin kuin jo aineiston esikäsittelyn yhteydessä luvussa 3 havaittiin. Kuva on jälleen ylivoimaisesti paras tapa tarkastella tilannetta.

Kuvassa 6.10 siluettiarvot on piirretty ryhmittäin ja siluettiarvoittain järjestetystä aineistosta havaintonumeroita vasten, jolloin näkyy, mistä nämä arvot ovat saaneet nimensä: aineistosta piirtyy siluetti, josta voidaan arvioida ryhmittelyn onnistumista. Joka ryhmässä on havaintoja, joiden siluettiarvo on negatiivinen, etenkin ryhmässä 1. Nämä havainnot ovat siis menetelmän mielestä väärässä ryhmässä. Suurimmatkin arvot ovat vain vähän yli 0.6:n suuruisia.



Kuva 6.10. Ensimmäisen medoidiryhmittelyn siluetti.

Saatu tulos on vain yksi vaihtoehto lukemattomien muiden joukossa. Eri määrällä ryhmiä saataisiin erilaisia tuloksia, samoin jos etäisyysmittaa muutettaisiin. Onneksi sentään medoidihavainnot löytyvät yhden tällaisen kokeilun puitteissa yksikäsitteisesti. Se on menetelmän yksi kiistaton etu verrattuna perinteisempiin keskiarvoryhmittelyihin, joissa alkuryhmitystä saatetaan joutua hakemaan satunnaisesti ja kokeilemaan eri vaihtoehtoja.

Faktoripisteet ja taustamuuttujat

Kun pidetään kiinni ryhmien lukumäärästä ja etäisyysmitasta, mutta otetaan mukaan taustamuuttujat paino ja ikä, ryhmittely selkiytyy huomattavasti, mikä ei ole yllätys. Tulokset näkyvät tulosteesta 6.3.

Tuloste 6.3. Medoidiryhmittely faktoripisteillä ja taustoilla.

```
Cluster analysis by medoids of Kaufman and Rousseeuw (1987)
Data UNRYHM  N=264
Group Medoid       n Mean            TUNTO PANOS PAINE NEGAT paino ika
 1    68          51 0.629           -0.19 -0.18 -0.09  1.02  66.5  61
 2    109         82 0.537           -0.84 -0.25 -0.57 -0.08  63.0  25
 3    34          31 0.292           -0.31 -0.53  0.93 -0.66  90.0  51
 4    88         100 0.514            0.11 -0.68 -0.24 -0.79  60.0  44
Mean of all silhouette values is 0.517
```

s. 209

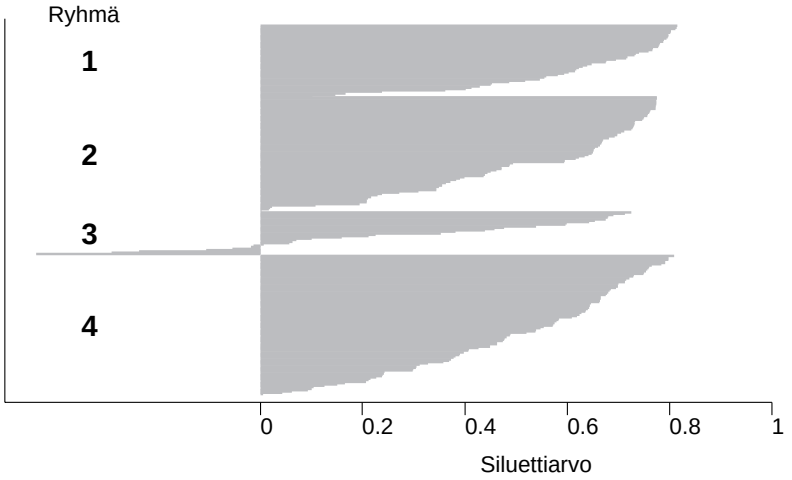
Muuttujien valinta on ryhmittelyyn vaikuttavista tekijöistä tärkein, sillä se on lähimpänä varsinaista asiasisältöä. Etäisyysmitta on aika tekninen seikka. Ryhmien lukumäärälle ei yleensä ole sisällöllisiä perusteita samaan tapaan kuin faktoreiden määrälle mittausmallissa. Voi tietenkin olla, että halutaan tehdä sama määrä ryhmiä kuin aiemmissa tutkimuksissa, mutta luultavasti päädytään kokeilemaan muitakin vaihtoehtoja.

Ryhmät näyttävät koostuvan koko lailla eri havainnoista kuin edellä, mikä selviää tutkimalla tulosteen 6.4 vertailutaulukkoa. Siinä `G1` viittaa ryhmittelyyn pelkillä faktoripisteillä ja `G4` niillä ja taustamuuttujilla. Puuttuvien havaintojen vuoksi `G1`:n frekvenssien summat eroavat tulosteen 6.2 (s. 167) lukumääristä.

Tuloste 6.4. Kahden medoidiryhmittelyn vertailutaulukko.

```
   G4   1   2   3   4 sum
G1 **
 1     13  26  18  19  76
 2      8  23   2  24  57
 3     21  19   4  32  76
 4      9  14   7  25  55
  sum  51  82  31 100 264
```

Jälkimmäisen ryhmittelyn (`G4`) siluettikuva 6.11 näyttää myös selvästi paremmalta. Nyt tosin esiintyy vielä suurempia negatiivisia arvoja, mutta vain yhdessä ryhmässä. Kyseessä on ryhmä 3, joka on myös havaintomäärältään pienin. Ryhmien koot vaihtelevat enemmän kuin ensimmäisessä ryhmittelyssä (`G1`). Koot näkyvät tulosteen 6.4 summariveistä sekä tulosteista 6.2 (s. 167) ja 6.3 (s. 169). Siluettiarvojen maksimit ovat 0.8:n luokkaa, mutta joka ryhmässä on yhä kaiken kokoisia arvoja. Näin ollen ryhmät sijoittuvat ainakin osittain päällekkäin tai hyvin lähelle toisiaan.



Kuva 6.11. Toisen medoidiryhmittelyn siluetti.

Lopulta ryhmittelyjen tuloksena saatiin aikaan melko vähän: neljä erikokoista ryhmää. Sisällölliset tulkinnat jäivät vähemmälle, vaikka se on kaikkein olennaisinta. Koska ryhmittelyt eivät perustu tilastollisiin malleihin, ei tulosten arviointiin ole samanlaista pohjaa kuin faktorianalyysissa. Ryhmiä on tutkittava tarkemmin ja yritettävä hahmottaa, mikä yhdistää niihin kuuluvia havaintoja. Tähän perehdytään seuraavassa luvussa.

7 Ryhmien visualisointi

Edellisessä luvussa nähtiin, kuinka aineistoa voitiin tiivistää edelleen ryhmittelemällä joko havaintoja tai muuttujia tai molempia. Ryhmittelyyn käytettiin aiemmin muodostettuja faktoripisteitä, mutta myös aineiston taustamuuttujia.

Koska ryhmittelyä voidaan tehdä lukemattomilla tavoilla, eivät muodostetut ryhmät ole yksikäsitteisiä eivätkä aina edes kovin selkeitä tulkita. Missään tapauksessa ei kannata tyytyä löydettyihin ryhmiin sellaisenaan. Tässä luvussa perehdytään menetelmiin, joilla ryhmiä päästään tarkemmin tutkimaan ja visualisoimaan.

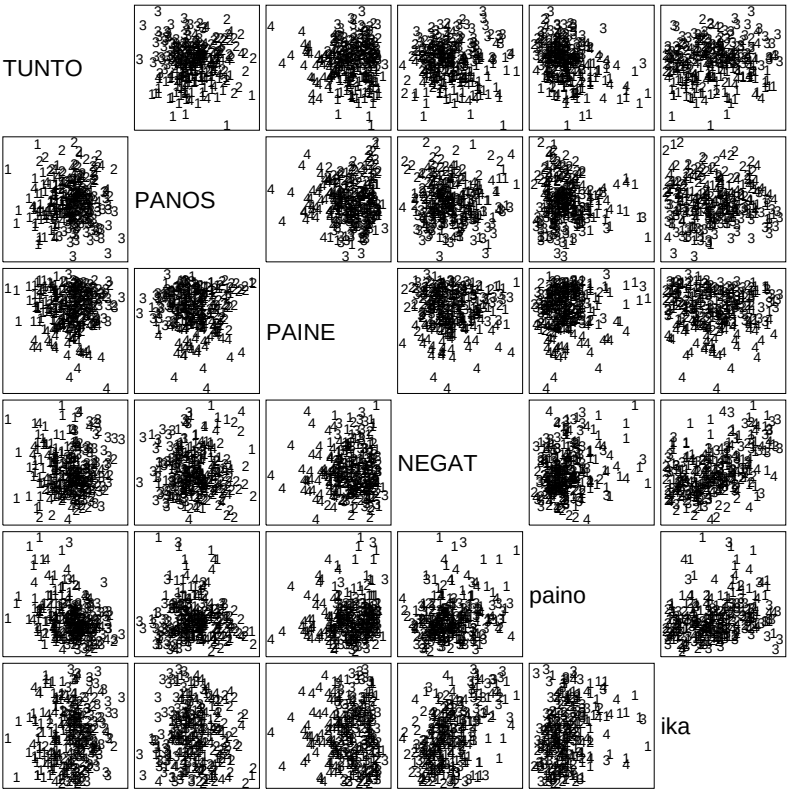
7.1 Hajontakuvan yleistyksiä

Luvussa 3 todettiin hajontakuvan olevan tilastollisista kuvista tärkeimpiä. Sillä on etenkin menetelmien yhteydessä monia käyttötarkoituksia, ja niinpä erilaisia hajontakuvia on ollut sittemmin esillä joka luvussa. Perustilanteiden ohella sen avulla on kuvattu muuttujia faktoriavaruudessa (kuva 4.2, s. 105), regressiomallin havaintokohtaista diagnostiikkaa (kuva 5.9, s. 149), ulkonäön sanallisia luonnehdintoja (kuva 6.7, s. 163) ja medoidiryhmittelyn siluetteja (kuva 6.10, s. 168).

Kirjan viimeisen luvun kuluessa laajennetaan vielä hajontakuvan toimialaa. Edellä tiivistettyä ja ryhmiteltyä aineistoa visualisoidaan piirtämällä erilaisia kuvia sekä suoraan että tiivistämällä tuloksia edelleen kahdella tilastollisella menetelmällä, erotteluanalyysilla ja korrespondenssianalyysilla.

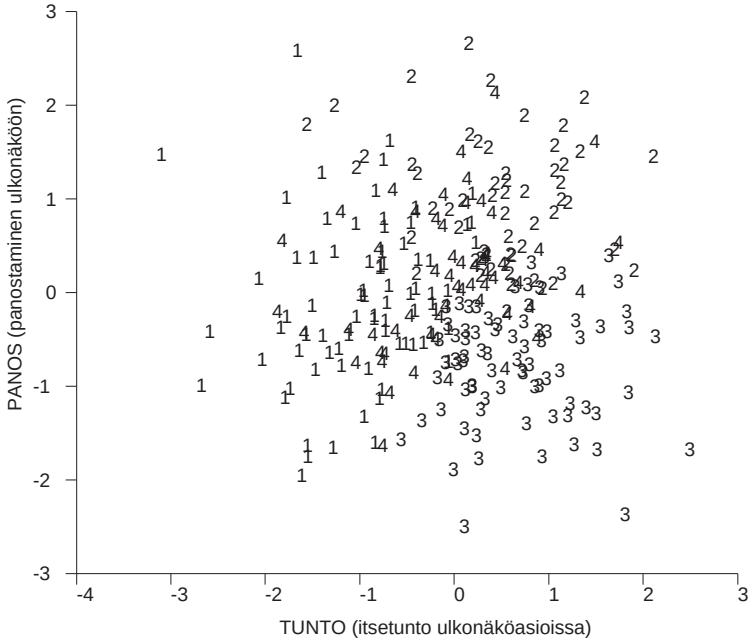
Hajontakuvamatriisi

Aineiston tiivistämisenkin jälkeen muuttujia on yleensä niin paljon, että yleiskuvan saaminen pelkistä parittaisista hajontakuvista on vaikeaa. Ennen kuin katsotaan, miten voidaan puristaa tietoja edelleen tiiviimpään muotoon, tutustutaan kuvan 7.1 *hajontakuvamatriisiin*. Siinä on koottu samaan kuvaan useamman muuttujan väliset hajontakuvat. Muuttujina ovat neljä faktoripistemuuttujaa sekä taustamuuttujat paino ja ikä. Havaintopisteiden tilalla on ryhmien numerot luvussa 6 tehdystä medoidiryhmittelystä, jossa käytettiin vain faktoripisteitä (ks. tuloste 6.2, s. 167).



Kuva 7.1. Hajontakuvamatriisi; faktoripisteet, paino ja ikä.

Kuvaan 7.1 sisältyy 30 hajontakuvaa, mutta lävistäjän eri puolilla ovat samat kuvat, vain toisinpäin piirrettyinä. Erilaisia kuvia on siis 15. Hajontakuvamatriisi muistuttaa korrelaatiomatriisia (ks. tuloste 3.11, s. 79), jossa yksittäisen korrelaatiokertoimen tilalla on kokonainen hajontakuva. Kussakin osakuvassa muuttujan arvot on piirretty vastaavan suuntaisesti kuin niiden nimet lukevat lävistäjällä, siis esimerkiksi ensimmäisen pystyrivin kuvissa muuttuja `TUNTO` vaihtelee vaakasuunnassa ja ensimmäisen vaakarivin kuvissa pystysuunnassa. Kuvista on häivytetty kaikki akselien merkinnät, sillä vain yleisvaikutelma kiinnostaa.

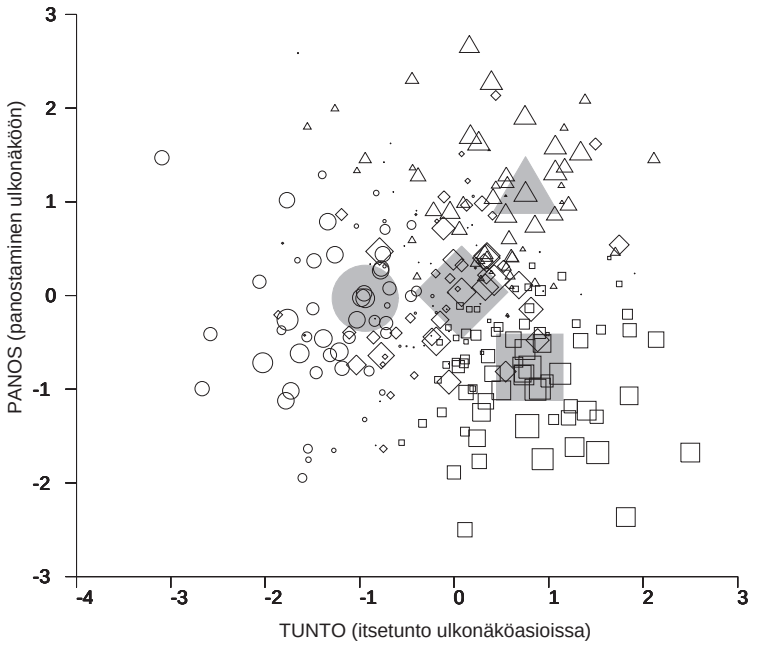


Kuva 7.2. Faktoripisteiden hajontakuva medoidiryhmittäin.

Yksityiskohtia voi tutkia piirtämällä haluamiaan osakuvia erikseen. Näin on tehty kuvassa 7.2, joka on ”suurennos” hajontakuvamatriisin 7.1 vasemman reunan ylimmästä kuvasta. Siinä ovat vastakkain kaksi ensimmäistä faktoripistemuuttujaa, `TUNTO` ja `PANOS`.

Havainnollisia symboleita vai symbolisia havaintoja?

Kuvassa 7.3 hajontakuvaa 7.2 on muunneltu merkitsemällä havainnot numeroiden sijaan eri symboleilla. Ympyrät, kolmiot, neliöt ja vinoneliöt erottelevat medoidiryhmittelyllä (ks. tuloste 6.2, s. 167) löydetyt neljä ryhmää. Symbolin koko on verrannollinen siluettiarvoon (ks. kuva 6.10, s. 168). Ryhmien keskukset eli medoidit on lisäksi piirretty taustalle harmaalla värillä, vastaavin symbolein mutta suurempina.



Kuva 7.3. Faktoripisteiden symbolikuva medoidiryhmittäin.

Lisäämällä tavanomaiseen kaksiulotteiseen hajontakuvaan tietoja muista muuttujista kuvasta tulee aidosti moniulotteinen. Kuvaa 7.3 voidaan pitää *neliulotteisena*, koska sen piirtämiseen on käytetty kahden faktoripistemuuttujan ohella ryhmäkoodia ja siluettiarvoa. Koordinaatistossa faktoripisteet määräävät havainnon sijainnin, ryhmäkoodi

symbolin ja siluettiarvo symbolin koon. Muita hajontakuvan yleistämismahdollisuuksia esittelevät muun muassa Mustonen (1995, 1–5) ja Robbins (2005).

Kuvia piirtämällä ryhmien olemus alkaa paremmin hahmottua. Medoidihavaintojen avulla kuvasta 7.3 nähdään, että itsetunnon suhteen ympyräryhmä on keskimääräistä heikompi, kolmio- ja neliöryhmät parempia ja vinoneliöryhmä niiden välissä. Neliöryhmä panostaa ulkonäköön vähemmän ja kolmioryhmä enemmän. Ympyrä- ja vinoneliöryhmät asettuvat keskiarvon kohdalle.

Ryhmien välillä on siis nähtävissä useammanlaisia eroja riippuen siitä, mistä suunnasta katsotaan. Kahden muuttujan kuvasta tilanne on vielä käsitettävissä, mutta kun mukaan otetaan lisää muuttujia, mahdollisia katselusuuntia tulee yhä enemmän ja tulkinta vaikeutuu. Aavistuksen tästä sai kuvasta 7.1 (s. 172). Kuvien hyödyllisyydestä huolimatta kokonaiskuva ei välttämättä hahmotu ilman pidemmälle vietyä aineiston tiivistämistä. Tähän tarkoitukseen soveltuu seuraavaksi tarkasteltava menetelmä.

7.2 Erotteluanalyysi

Myös useampiulotteisessa tilanteessa on mahdollista löytää suuntia, joista katsottuna ryhmien erot näyttäytyvät mahdollisimman selkeinä. Suuntien löytämiseen ja visualisointiin erikoistunut tilastollinen monimuuttujamenetelmä on *erotteluanalyysi*. Sen avulla päästään myös syvällisemmin kiinni ryhmien olemukseen, sillä muotoa ja sijaintia tärkeämpää on tietää, millaisia ryhmät ovat sisällöiltään.

Erotteluanalyysissa kysytään: *”Mikä erottaa ryhmät toisistaan?”* Ryhmiä ei yritetä muodostaa niin kuin ryhmittelymenetelmissä, vaan ryhmät oletetaan ”tunnetuiksi”. Tässä vaiheessa tunteminen on vielä pinnallista; ryhmät tunnetaan vain koodinimillä.

Esimerkiksi edellä visualisoituja ryhmiä voidaan työstämisvaiheessa kutsua numeroin 1, 2, 3 ja 4 tai symbolein ympyrät, kolmiot, neliöt ja vinoneliöt, mutta tulosten esittämistä ja ymmärtämistä ajatellen ryhmille on annettava kuvaavammat nimet. Haaste on vastaava kuin faktoreiden nimeämisessä, jota käsiteltiin kohdassa 4.3.2 (s. 96).

Mittauskehikko ja muuttujien valinta

Erotteluanalyysi sijoittuu luontevasti mittauskehikkoon (ks. kuva 5.1, s. 122). Vertailuperuste on aineiston ositus, joka on voitu muodostaa ryhmittelymenetelmillä, taustamuuttujien avulla tai näiden yhdistelyllä. Analyysi synnyttää ryhmien ja muuttujien määrästä riippuen yhden tai useampia tulosasteikkoja, joille havainnot asettuvat siten, että ryhmien väliset erot tulevat selvimmin esiin.

Erotteluanalyysin ainoa luokittelutasoinen muuttuja on ryhmän koodi. Varsinaiset muuttujat oletetaan jatkuviksi, ja niihin pätee sama kuin ryhmittelyssä ja regressioanalyysissa: alkuperäiset muuttujat kannattaa unohtaa ja toimia faktoripisteillä. Mahdolliset taustamuuttujat on myös syytä tuoda analyysiin mahdollisimman jatkuvina. Muuttujien valinta perustuu siihen, minkä asioiden suhteen ryhmäeroja halutaan selvittää.

Ulkonäkötutkimuksen erotteluanalyysi

Seuraavassa tarkastellaan, mikä erottaa kahdella eri medoidiryhmittelyllä luvussa 6 luodut ulkonäköaineiston ryhmät toisistaan. Ensimmäinen ryhmittely perustui vain faktoripisteisiin (tuloste 6.2, s. 167). Toisessa olivat mukana myös taustamuuttujat paino ja ikä (tuloste 6.3, s. 169). Molemmissa luotiin neljä ryhmää.

Tuloste 7.1. Ensimmäisen medoidiryhmittelyn erotteluanalyysi.

```
   Eig.val.  %      Can.corr  Chi^2     df P
1  1.515222  44.08  0.776158  595.0488  12 0.0001
2  1.123631  32.69  0.727398  352.4679   6 0.0001
3  0.798683  23.23  0.666361  154.3953   2 0

Correlations between variables and discriminators
       Discr1 Discr2 Discr3
TUNTO   -0.24  -0.89   0.26  itsetunto ulkonäköasioissa
PANOS    0.48  -0.26  -0.76  panostaminen ulkonäköön
PAINE   -0.81   0.07  -0.58  sosiaaliset ulkonäköpaineet
NEGAT   -0.15   0.40   0.21  negatiivinen suhtautuminen
```

Havainnollisuuden vuoksi molempien ryhmittelyjen erotteluanalyyseissa käytetään samoja muuttujia kuin ryhmittelyissä. Todellisuudessa erotteluanalyysissa ei tarvita tietoa ryhmien syntytavasta.

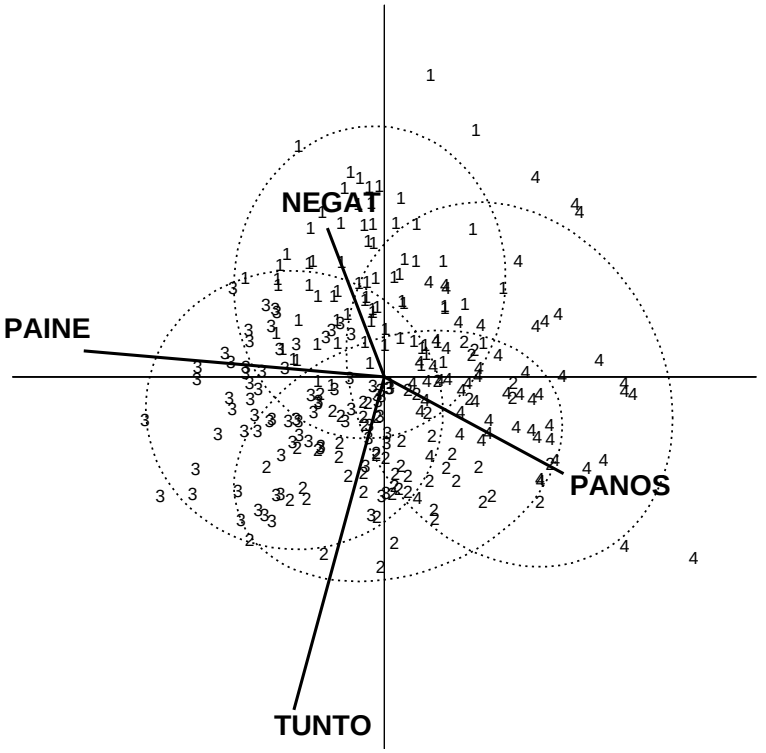
Erotteluanalyysi perustuu ryhmien keskiarvoihin, hajontoihin ja korrelaatioihin. Niiden informaation analyysi tiivistää *erottelumuuttujiksi* tai *erottelijoiksi*, joita on aina yksi vähemmän kuin ryhmiä. Tulosteen 7.1 yläosan yhteenvedossa erottelumuuttujiin viitataan numeroilla 1, 2 ja 3; alaosassa vastaavasti `Discr`-alkuisilla nimillä. Ensimmäinen selittää ryhmien välisistä eroista 44 %, toinen 33 % ja kolmas loput, 23 %. Tilastollinen testi osoittaa erot merkitseviksi, mikä ei taaskaan ole kovin vahva todiste suuntaan eikä toiseen.

Tärkeämpää on tutkia, mihin erottelut perustuvat ja riittääkö kaksi ulottuvuutta erojen kuvaamiseen. Visualisointi on jälleen yksinkertaisempaa, jos tarkastelut tiivistetään kahteen ulottuvuuteen, vaikka osa informaatiosta hukataan (vrt. kohta 6.2, s. 159).

Erotteluanalyysin tulkinta ja visualisointi

Tulosteen 7.1 alaosan *rakennematriisia* tulkitaan samaan tapaan kuin faktorimatriisia, sillä luvut ovat erottelumuuttujien ja analyysin muuttujien välisiä korrelaatioita. Tulosteen yläosan perusteella voidaan katsoa kahden erottelumuuttujan riittävän, sillä ne selittävät yhteensä lähes 80 % ryhmäeroista. Rakennematriisista nähdään, että ensimmäisen erottelijan ulottuvuudella korostuvat toisilleen vastakkaisina panostaminen ulkonäköön ja sosiaaliset ulkonäköpaineet. Toisella korostuvat itsetunto ulkonäköasioissa ja negatiivinen suhtautuminen ulkonäköön. Kolmatta erottelumuuttujaa ei siis tässä tulkita.

Kun muodostetaan kahta ensimmäistä erottelijaa vastaavat tulosasteikot eli *erottelupisteet* aineistoon, voidaan tulkintoja syventää visuaalisesti piirtämällä erottelupisteet vastakkain. Näin saadaan *erotteluavaruus*, jota havainnollistaa kuva 7.4. Siinä havainnot on merkitty niiden ryhmäkoodeilla 1, 2, 3 ja 4. Muuttujat on piirretty origosta lähtevinä vektoreina. Koordinaattiakseleina ovat siis kaksi ensimmäistä erottelumuuttujaa, jotka tulosteessa 7.1 esiintyivät myös nimillä `Discr1` ja `Discr2`. Kuva 7.4 on niin sanottu *kaksoiskuva (biplot)*, sillä samaan kuvaan on yhdistetty tietoja sekä havainnoista että muuttujista. Kaksoiskuvia hyödynnetään myös kohdassa 7.3 (s. 183) käsiteltävässä korrespondenssianalyysissa.



Kuva 7.4. Ensimmäisen medoidiryhmittelyn erotteluavaruus.

Kuvassa 7.4 ryhmien ympärille piirretyt *hajontaellipsit* auttavat hahmottamaan ryhmän sijaintia ja muotoa sekä ryhmien välisiä eroja. Ellipsit on piirretty 95 %:n tasolle, mikä tarkoittaa, että noin 5 % kunkin ryhmän havainnoista sijaitsee sen ellipsin ulkopuolella. Kyseiset havainnot ovat parhaiten toisistaan erottuvia, jopa jonkinlaisia ääritapauksia. Kuvan keskiosan havaintoja on vaikeampi erottaa toisistaan. Kaikkiaan ryhmäerot eivät näytä kovin selviltä. Vaakasuunnassa katsottuna ryhmä 4 panostaa enemmän ulkonäköön ja ryhmä 3 kokee enemmän sosiaalisia ulkonäköpaineita. Pystysuunnassa katsottuna ryhmässä 1 esiintyy enemmän negatiivista suhtautumista ulkonäköön, mutta muuten ryhmät menevät aika lailla päällekkäin.

Taustamuuttujien vaikutus ryhmien erotteluun

Kun erotteluun otetaan mukaan paino ja ikä, tulos kirkastuu ainakin tulosteen 7.2 perusteella selvästi. Nyt ensimmäinen erottelija selittää jo yksin lähes 80 %, toiselle jää reilut 20 % ja kolmas voidaan unohtaa saman tien.

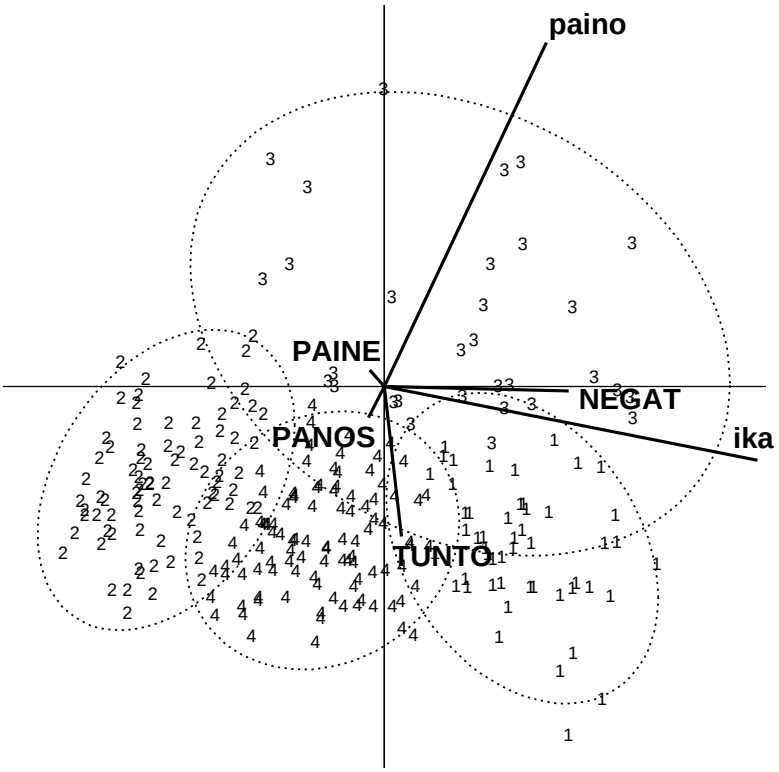
Tuloste 7.2. Toisen medoidiryhmittelyn erotteluanalyysi.

```
   Eig.val.  %      Can.corr  Chi^2     df P
1  4.393933  77.80  0.902556  649.7074  18 0
2  1.215083  21.52  0.740641  214.9065  10 0
3  0.038400   0.68  0.192302    9.7218   4 0.0454

Correlations between variables and discriminators
        Discr1 Discr2 Discr3
TUNTO     0.05  -0.39   0.25  itsetunto ulkonäköasioissa
PANOS    -0.04  -0.08   0.09  panostaminen ulkonäköön
PAINE    -0.04   0.04  -0.72  sosiaaliset ulkonäköpaineet
NEGAT     0.48  -0.01  -0.57  negatiivinen suhtautuminen
paino     0.43   0.90  -0.05  paino (kg)
ika       0.98  -0.19   0.02  ikä (vuosina)
```

Tulosteen 7.2 rakennematriisi paljastaa, että ensimmäinen erottelija on käytännössä ikä ja toinen paino. Faktoripisteiden osuus ryhmäerojen selittämisessä jää vähäisemmäksi muilla paitsi kolmannella ulottuvuudella, jolla ei kuitenkaan ole mitään selitysvoimaa.

Erotteluavaruus kertoo tulokset visuaalisemmin kuvassa 7.5. Ryhmä 3 eroaa muista ryhmistä ennen kaikkea painon, käänteisesti myös itsetunnon suhteen. Siinä on myös eniten hajontaa sekä painon että iän suhteen. Muut ryhmät järjestyvät paljolti iän mukaan siten, että ryhmässä 1 on keskimäärin vanhempia, ryhmässä 4 nuorempia ja ryhmässä 2 vielä nuorempia vastaajia. Negatiivista suhtautumista ulkonäköön esiintyy eniten vanhemmissa ikäryhmissä.



s. 209

Kuva 7.5. Toisen medoidiryhmittelyn erotteluavaruus.

Havaintojen luokittelu

Erotteluanalyysin tuloksia on tapana arvioida luokittelemalla havainnot uudelleen erottelun perusteella ja vertaamalla näin saatua luokitusta alkuperäiseen luokitukseen. Tätä analyysin periaatteessa erillistä vaihetta kutsutaan toisinaan *luokitteluanalyysiksi*. Sen avulla on myös mahdollista luokitella uusia havaintoja olemassa oleviin ryhmiin. Esimerkiksi jos kyselyyn saadaan uusia vastauksia, mutta ryhmittelyjä ei haluta tehdä uudelleen, voidaan tutkia, mihin ryhmään kyseiset vastaajat mahtaisivat vastausprofiilinsa perusteella kuulua.

Hyvä tapa soveltaa luokittelua on jakaa aineisto satunnaisesti kahteen osaan, tehdä erotteluanalyysi toisella osalla ja luokitella sen perusteella toisen osan havainnot. Näin saadaan todellisempi käsitys luokittelun onnistumisesta. Ulkonäkötutkimuksessa voisi myös luokitella vuoden 2005 vastaukset vuoden 1997 erottelun perusteella, koska luvun 4 lopussa todettiin, ettei rakenne-eroja vuosien välillä ilmennyt.

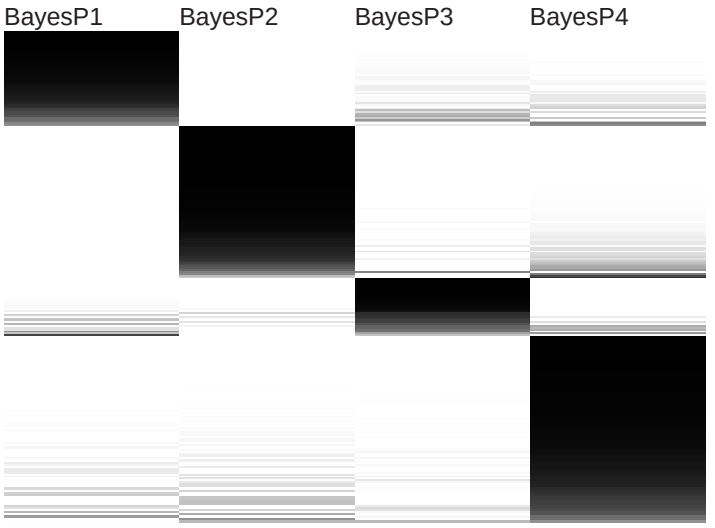
Nyt tyydytään tarkastelemaan vain vuoden 1997 aineistoa ja luokitellaan sen havainnot erottelun perusteella uudelleen neljään ryhmään. Havaintojen luokittelussa käytetään tilastollisia perusteita, muun muassa erilaisia etäisyysmittoja. Tässä sovelletaan niin sanottua *bayesiläistä* periaatetta. Siinä ryhmiin kuulumiselle asetetaan ennakkoarviot, jotka analyysi normaalijakaumaoletuksen pätiessä tarkentaa niin, että joka havainnolle saadaan todennäköisyyspohjainen arvio ryhmiin kuulumisesta.

Tuloste 7.3. Havaintojen luokittelu erotteluanalyysin perusteella.

```
   Bayes0   1   2   3   4 sum      Bayes1   1   2   3   4 sum
G4 ******                      G4 ******
 1          48   0   0   3  51   1          49   0   0   2  51
 2           0  79   0   3  82   2           0  79   0   3  82
 3           3   0  25   3  31   3           1   0  29   1  31
 4           1   2   0  97 100   4           0   0   0 100 100
      sum   52  81  25 106 264        sum   50  79  29 106 264
```

Tulosteessa 7.3 alkuperäinen ryhmäkoodi `G4` on taulukoitu ristiin kahden erilaisen Bayes-luokittelun kanssa: `Bayes0` on tehty olettaen ryhmien kovarianssirakenteet samoiksi, `Bayes1` huomioi ryhmien omat rakenteet. Tuloksissa ei ole suuria eroja, mutta jälkimmäinen on hieman parempi. Enin osa havainnoista on molemmissa taulukoissa lävistäjällä, mikä kertoo luokittelun onnistuneen hyvin. Jälkimmäisessä (`Bayes1`) ryhmän 4 havainnot luokittuvat täysin oikein, muissa ryhmissä on pieniä eroavaisuuksia. Kuten edellä todettiin, tulos on liian optimistinen, sillä luokittelu tehtiin samoilla havainnoilla kuin erottelu.

Bayes-luokittelu antaa siis havaintokohtaiset, ryhmiin kuulumisen todennäköisyydet, joita tutkimalla saa vielä tarkemman käsityksen tilanteesta. Kuva 7.6 esittää havaintomatriisia, jonka pystysuunnassa ovat ryhmien Bayes-todennäköisyyksiä vastaavat muuttujat ja vaakasuunnassa aineiston havainnot. Joka havainnosta muodostuu ohut rivi, jonka värisävy vaihtelee mustasta valkoiseen muuttujien arvojen mukaisesti. Mitä lähempänä havainnon arvo on ykköstä, sitä mustempi on sen sävy ja sitä varmemmin havainto kuuluu kyseiseen ryhmään. Vastaavasti mitä lähempänä havaintoarvo on nollaa, sitä vaaleampi on sävy ja sitä epätodennäköisempää on sen kuuluminen ryhmään.



Kuva 7.6. Havaintojen Bayes-todennäköisyydet ryhmittäin.

Aineisto on kuvassa 7.6 järjestetty niin, että ryhmiin kuuluminen erottuu selvästi. Joitakin hajanaisia havaintoja lukuun ottamatta luokittelu on selkeä. Poikkeushavaintojen profiileja voisi olla hyödyllistä tutkia tarkemmin, mutta siihen ei tässä syvennytä. Lisää erotteluanalyysista, muun muassa sen yhteyksistä logistiseen regressioanalyysiin, kertoo sovelluspainotteinen teos *Multivariate Data Analysis* (Hair, Jr. ym., 1998, 239–325).

7.3 Korrespondenssianalyysi

Viimeiseksi perehdytään *korrespondenssianalyysiin*, jossa visualisoidaan ryhmien välisiä suhteita. Kuvat ovat pääosassa ja numeeriset tulokset vähäisemmässä asemassa kuin muissa monimuuttujamenetelmissä. Kuvat ovat haastavia, mutta parhaimmillaan antoisia.

Korrespondenssianalyysia sovelletaan myös laadullisissa tutkimuksissa, sillä muuttujat saavat olla mittaustasoiltaan millaisia tahansa, myös luokittelutasoisia. Esimerkiksi tekstuaalisista aineistoista voidaan koodauksilla rakentaa luokkia, joiden välisten suhteiden tutkimisessa visualisoinnista on paljon apua.

7.3.1 Kahden muuttujan taulukko

Korrespondenssianalyysin lähtökohta on kahden muuttujan ristiintaulukko, jota käsiteltiin jo luvussa 3. Jatkuvat muuttujat on siis luokiteltava: analyysissa muuttujat esiintyvät vain luokittelu- tai järjestystasoisina. Ristiintaulukoiden tapaan korrespondenssianalyysilla päästään käsiksi millaisiin yhteyksiin tahansa, ei vain lineaarisiin, kuten useissa muissa monimuuttujamenetelmissä. Käytännönläheinen korrespondenssianalyysin perusteos on *Correspondence Analysis in Practice* (Greenacre, 2007).

Menetelmän yleistyksissä tavallista taulukkoa laajennetaan eri tavoin, jotta analyysiin saadaan sisällytettyä useampia muuttujia. Tämän luvun lopussa tutustutaan yhteen, kyselytutkimuksessa hyödylliseen korrespondenssianalyysin yleistykseen, mutta aluksi katsotaan perustilannetta, jossa muuttujia on vain kaksi.

Riippumattomuushypoteesi

Johdatukseksi korrespondenssianalyysiin tarkastellaan vastaajan peruskoulutusta ja yhdessäoloaikaa nykyisen kumppanin kanssa. Peruskoulutus oli esillä jo luvussa 2, esimerkissä 2.5 (s. 31). Tässä yhteydessä sen kaksi ensimmäistä luokkaa on yhdistetty ja muodostettu kolmiluokkainen muuttuja `PK`, jonka luokat ja koodit ovat kansakoulu (`kk`), peruskoulu (`pk`) ja ylioppilas (`yo`). Yhdessäoloaikaa tutkittiin puolestaan luvussa 3, esimerkiksi sen frekvenssijakauma oli tulosteessa 3.3 (s. 58). Tässä luokitusta on tiivistetty jättämällä pois ne, jotka eivät olleet vakituisessa parisuhteessa.

Tulosteessa 7.4 peruskoulutus ja yhdessäolo vuosina on taulukoitu vastakkain. Tässä yhteydessä on käytetty myös vuoden 2005 vastauksia, sillä muuten luokkafrekvenssit olisivat jääneet verrattain pieniksi. Koko aineiston käyttö on perusteltavissa, koska rakenneeroja ei ilmennyt.

Riippumattomuuden tilastollinen testaus

Ristiintaulukon luokittelijoiden välistä riippuvuutta voidaan arvioida tilastollisella testillä, jota kutsutaan *khiin neliö* -testiksi tai *khii toiseen* -testiksi. Testin nollahypoteesi väittää, että luokittelijat, tässä tapauksessa peruskoulutus ja yhdessäoloaika, ovat toisistaan riippumattomia. Toisin sanottuna oletetaan, että yhdessäoloaika ei riipu peruskoulutuksesta tai päinvastoin; ensin mainittu lienee tulkinnallisesti järkevämpi oletus.

Sekä taulukoilla että testillä on yhteys korrespondenssianalyysiin, joka toisaalta voidaan nähdä laajennuksena kohdassa 6.2 (s. 159) esitetystä moniulotteisesta skaalauksesta. Korrespondenssianalyysissa skaalaus tapahtuu taulukon molempien luokittelijoiden suhteen samanaikaisesti, ja etäisyysmittana (vrt. kohta 6.1, s. 151) käytetään khiin neliötä. Katsotaan aluksi, mitä khiin neliö -testistä voidaan päätellä.

Tulosteen 7.4 ylimmässä taulukossa on *havaitut frekvenssit*, siis varsinainen ristiintaulukko. Seuraavassa on *odotetut frekvenssit*, jotka vastaavat nollahypoteesin riippumattomuusoletusta. Ne saadaan taulukon *reunajakaumien* eli luokittelijoiden omien frekvenssijakaumien avulla, kertomalla taulukon kutakin solua vastaavat reunafrekvenssit keskenään ja jakamalla tulo kokonaismäärällä. Esimerkiksi kymmenen vuotta tai alle yhdessä olleiden ja kansakoulun suorittaneiden odotettu frekvenssi 34 saadaan jakamalla lukujen 91 ja 144 tulo 384:llä. Odotetut frekvenssit on tässä pyöristetty kokonaisluvuiksi.

Riippumattomuushypoteesia testaava khiin neliö -testi vertaa toisiinsa havaittuja ja odotettuja frekvenssejä. Erojen suuntia ei huomioida, vaan vertailu tapahtuu neliöllisesti ja suhteessa odotettuihin frekvensseihin. Näin saadaan tulosteen 7.4 alimman taulukon solujen luvut, joita kutsutaan khiin neliön *kontribuutioiksi*. Niiden summa, 74, on khiin neliö -testisuure.

Tuloste 7.4. Peruskoulutuksen ja yhdessäoloajan ristiintaulukot.

```
Observed frequencies with marginals (yhdessa/PK)
                     kk      pk      yo     Sum
         10/alle     16      43      85     144
         11-20       11      35      52      98
         21-30       20      18      28      66
         31-40       30      10      14      54
         yli 40      14       4       4      22
         Sum         91     110     183     384

Expected frequencies with marginals (yhdessa/PK)
                     kk      pk      yo     Sum
         10/alle     34      41      69     144
         11-20       23      28      47      98
         21-30       16      19      31      66
         31-40       13      15      26      54
         yli 40       5       6      10      22
         Sum         91     110     183     384

Contributions to X^2 with marginals (yhdessa/PK)
                     kk      pk      yo     Sum
         10/alle    9.63    0.07    3.91   13.61
         11-20      6.43    1.71    0.60    8.74
         21-30      1.22    0.04    0.38    1.64
         31-40     23.13    1.93    5.35   30.41
         yli 40    14.81    0.84    4.01   19.66
         Sum       55.21    4.60   14.25   74.06
```

Testin p-arvoon tarvittavat *vapausasteet* määräytyvät taulukon varsinaisten rivien ja sarakkeiden lukumäärien tulona, kun molemmista vähennetään ensin ykkönen. Tarkasteltavassa taulukossa on viisi riviä ja kolme saraketta, joten soluja on 15. Vapaasti voidaan valita kahdeksan; loput saadaan reunasummien avulla.

Riittävät todisteet nollahypoteesia vastaan khiin neliö -testillä saadaan, kun testisuure on selvästi vapausasteitaan suurempi. Tässä tapauksessa jo testisuureen arvo 16 riittäisi nollahypoteesin hylkäämiseen tavanomaisella viiden prosentin riskillä. Nyt havaittu 74 vastaa kahdeksalla vapausasteella p-arvoa, jossa on peräti 13 nollaa, joten tulos ei jää epäselväksi: peruskoulutus ja yhdessäoloaika riippuvat toisistaan. Se, mitä tämä tarkemmin tarkoittaa, voi sen sijaan jäädä epäselväksi, jos tyydytään pelkkään tilastolliseen testiin. Katsotaan seuraavaksi, miten tulkintaa voidaan syventää visuaalisesti.

7.3.2 Kahden muuttujan kuva

Luokiteltujen muuttujien kuvaajia ovat esimerkiksi luvussa 3 esitetyt pylväskuvat, joita voidaan piirtää myös ristiintaulukoista. Tyypillisesti yksi luokittelija määrää pylväät ja toinen jakaa ne luokkiensa mukaisiin osiin. Tällöin huomio keskittyy lähinnä lukumääriin tai prosenttiosuuksiin. Pylväskuvista on vaikea saada selkoa luokittelijoiden mahdollisista riippuvuuksista.

Korrespondenssianalyysissa riippuvuuksiin päästää kiinni skaalaamalla taulukon luokittelijat numeerisille asteikoille. Kun ne piirretään vastakkain, saadaan luokat sijoitettua kuvaan, ikään kuin kartalle. Kuten edellä mainittiin, analyysilla on yhteyksiä moniulotteiseen skaalaukseen (kohta 6.2, s. 159). Kaksisuuntaisen skaalauksen mahdollistaa juuri ristiintaulukko, joka on lähtökohtaisesti rikkaampi kuin moniulotteiseen skaalaukseen yleensä käytetty, symmetrinen etäisyysmatriisi.

Tulosteessa 7.5 on yhteenveto peruskoulutuksen ja yhdessäolon ristiintaulukon korrespondenssianalyysista. Siitä nähdään edeltä tutut khiin neliö -testin tiedot, p-arvoa myöten, joka on esitetty niin sanotussa ”tieteellisessä muodossa”. Tällaisia lukuja esiintyy toisinaan ohjelmien tulosteissa. Ruman näköinen luku tarkoittaa käytännössä nollaa. Yli kymmenen merkitsevän numeron esitystarkkuudella on harvoin, jos koskaan, käyttöä tilastollisissa yhteyksissä.

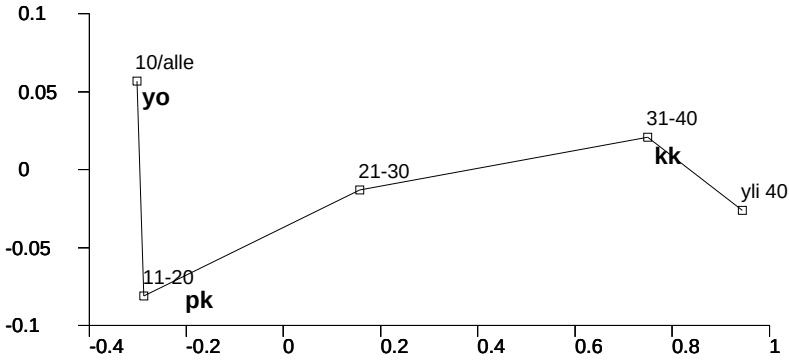
Tuloste 7.5. Korrespondenssianalyysin numeerinen yhteenveto.

```
Correspondence analysis on data yhdPK: Rows=5 Columns=3

    Canonical      Eigen-          Chi^2    Cumulative
    correlation    value                    percentage
 1  0.4357         0.1898      72.8890720       98.42
 2  0.0552         0.0030      1.17097315      100.00
                   0.1929       74.06 (df=8 P=7.66831e-013)
```

Tulosteen 7.5 muut luvut ovat samantyyppisiä ulottuvuuksien voimakkuuksien mittoja, joita on esitetty aiemmin, muun muassa erotteluanalyysin yhteydessä (vrt. tuloste 7.1, s. 176). Tästä näkyy jo, että ensimmäinen ulottuvuus kertoo lähes kaiken kerrottavissa olevan, sillä toiselle jää vain rippeet.

Korrespondenssianalyysin varsinainen tulos on kaksoiskuva, joka saadaan piirtämällä ulottuvuuksia vastaavat numeeriset asteikot vastakkain, sekä taulukon rivi- että sarakeluokittelijan suhteen, ja merkitsemällä pisteiden tilalle luokkien sanalliset kuvaukset. Kuvassa 7.7 on lisäksi kytketty viivalla toisiinsa yhdessäoloaikaa kuvaavat luokat. Tämä keino auttaa visualisoinnissa, mutta sitä on syytä käyttää harkiten, ja ainoastaan järjestystasoisiin luokittelijoihin. Peruskoulutuskin olisi tällainen, mutta luokittelutasoisten asioiden, kuten työllisyystilanteen, luokkien yhteen kytkeminen ei olisi perusteltua.

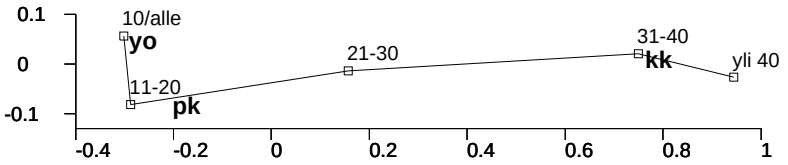


Kuva 7.7. Harhaanjohtava korrespondenssianalyysikuva.

Kuvasta 7.7 nähdään luokittelijoiden välillä selvä vastaavuus: korkeamman peruskoulutuksen suorittaneiden yhdessäoloaika on lyhyempi ja päinvastoin. Kuva antaa hyvän käsityksen taulukon sisällöstä; se on sen ainoa tavoite. Näin pienen taulukon kohdalla samat asiat voidaan nähdä lukuja tarkastelemalla, mutta mitä suurempi on taulukko, sitä vaikeammaksi käy yhteyksien hahmottaminen.

Vakava ongelma kuvassa 7.7 on sen pielessä oleva *kuvasuhde:* koordinaattiakselit eivät ole vertailukelpoisia keskenään, sillä pystyakseli on vaaka-akseliin nähden aivan liian pitkä. Virhe vääristää kuvan tulkinnan. Todellisuudessa pystysuuntaisen vaihtelun määrä on vaakasuuntaista huomattavasti vähäisempää, mikä nähtiin jo tulosteesta 7.5. Toisen ulottuvuuden osuudeksi taulukon kokonaisvaihtelusta jäi alle kaksi prosenttia. Kuva 7.7 valehtelee ja johtaa harhaan.

Kuva 7.7 on piirretty *tarkoituksellisesti väärin*, jotta nähtäisiin, miten helposti kuvilla saatetaan vääristellä tietoja. Kun kuva piirretään oikein (ks. kuva 7.8), näkymä on totuudenmukainen. Vaihtelua on käytännössä vain vaakasuunnassa, joten taulukon visualisointiin riittäisi yksi ulottuvuus. Pystysuuntaiset erot ovat niin pieniä, että ylitulkinnan riski vaikuttaa ilmeiseltä.



Kuva 7.8. Kahden muuttujan korrespondenssianalyysi.

Visuaalisen valehtelun vaara

Oikeasta kuvasuhteesta on pidettävä kiinni tilastollisia kuvia piirrettäessä, sillä väärin laadittu kuva valehtelee. Kuvasuhteen merkitys korostuu, jos koordinaattiakseleiden lukuarvot jätetään pois. Lukujen poistamisella voidaan keventää kuvan visuaalista ilmettä, mikäli luvut eivät välitä kiinnostavaa tietoa. Olennaista on tällöin vain kuvan origon sijainti. Sen voi kertoa koordinaattiakseleiden avulla ilman merkintöjäkin, mutta akseleiden pituuksien on oltava oikeassa suhteessa toisiinsa.

Kuvasuhteen kanssa on oltava erittäin tarkkana, sillä kuvanpiirto-ohjelmistot eivät siihen välttämättä automaattisesti kykene. Tarkemmin kuvasuhdetta ruotii Kuusela (2000, 90–94). Edelleen ajankohtainen kirja tilastollisten kuvien piirtämisen ja tulkinnan sudenkuopista on *How to Lie with Statistics* (Huff, 1954). Teos on suomennettu nimellä *Kuinka tilastoilla valehdellaan*.

Räikeimmillään tilastoilla valehdellaan esittämällä ”kolmiulotteisia” pylväs- tai piirakkakuvia tilanteissa, joilla ei ole mitään yhteyksiä aitoon moniulotteisuuteen. Tutkijan vastuulla on esittää tiedot totuudenmukaisesti, olivat ne lukuja tai kuvia.

7.3.3 Burtin matriisi

Kahden muuttujan ristiintaulukolla ei päästä pidemmälle, vaikka sitä visualisoitaisiin näyttävästi. Moniulotteiset, sisäkkäiset taulukot eivät auta asiaa, sillä suuri osa luokittelijoiden yhdistelmistä käy harvinaisiksi, vaikka aineistossa olisi miten paljon havaintoja tahansa. Esimerkki kolmiulotteisesta taulukosta nähtiin luvun 3 lopussa, kun tarkasteltiin tietojen puuttuvuutta mittareittain: sisäkkäin taulukoiduista tiedoista suurin osa oli nollia (ks. tuloste 3.13, s. 85).

Korrespondenssianalyysin moniulotteinen laajennus perustuu luokiteltujen muuttujien parittain muodostettuihin, yhteen koottuihin ristiintaulukoihin. Taulukkokokoelmaa kutsutaan *Burtin matriisiksi* (Greenacre, 2007, 140–141). Esimerkki kahdeksan muuttujan Burtin matriisista näkyy tulosteessa 7.6 (s. 190). Kokonaisuudessaan esimerkkinä tarkasteltava taulukko jatkuu viiden luokittelijan verran oikealle, mutta tilan säästämiseksi katsellaan vain osaa siitä. Muuttujien luokat nähdään tulosteen vasemmasta reunasta. Muuttujat ovat

1. *medoidiryhmittelyn ryhmäkoodi* (tuloste 6.3, s. 169), jossa viidennen luokan muodostavat vuoden 2005 havainnot,
2. *ikä* (tuloste 3.6, s. 65) luokiteltuna seitsemään luokkaan,
3. *työllisyystilanne* (esimerkki 2.3, s. 28) luokiteltuna tiiviimmin (tuloste 3.8, s. 67),
4. *liikunnan harrastaminen* (esimerkki 2.6, s. 31) luokiteltuna tiiviimmin,
5. *painon luonnehdinta* (esimerkki 2.7, s. 32) luokiteltuna tiiviimmin,
6. *huolissaan olo painosta* (esimerkki 2.8, s. 33),
7. *painoindeksiluokitus* (muodostettu vastaajalta kysyttyjen pituuden ja painon avulla ja soveltaen yleisesti käytettyä luokitusta, jossa rajat ovat 18.5, 25 ja 30), sekä
8. monessa kohtaa aiemmin käsitelty faktoripistemuuttuja ”*itsetunto ulkonäköasioissa*”, luokiteltuna neljään luokkaan mittauksen keskivirheen perusteella (ks. taulukko 4.2, s. 119).

Tulosteesta 7.6 nähdään, että Burtin matriisi on symmetrinen. Matriisi koostuu useista lohkoista, joita on tulosteessa korostettu hahmottamisen helpottamiseksi. Mustapohjainen lävistäjä muodostuu muuttujien omista frekvenssijakaumista. Näissä lohkoissa on lävistäjän ulkopuolella vain nollia, sillä luokat ovat toisensa poissulkevia. Harmaa- ja valkopohjaiset lohkot ovat muuttujien välisiä, tavallisia ristiintaulukoita.

Tuloste 7.6. Osa kahdeksan muuttujan Burtin matriisista.

		1					2							3					4
1	1	51	0	0	0	0	0	0	0	0	8	29	14	10	2	4	35	0	...
	2	0	82	0	0	0	12	26	44	0	0	0	0	39	4	20	0	19	...
	3	0	0	31	0	0	0	0	4	6	13	6	2	12	7	7	5	0	...
	4	0	0	0	100	0	0	0	8	63	29	0	0	81	3	15	0	1	...
	5	0	0	0	0	232	10	30	39	49	42	43	19	118	12	39	39	24	...
2	18-20	0	12	0	0	10	22	0	0	0	0	0	0	1	0	7	0	14	...
	21-25	0	26	0	0	30	0	56	0	0	0	0	0	22	2	11	0	21	...
	26-35	0	44	4	8	39	0	0	95	0	0	0	0	62	4	23	0	6	...
	36-45	0	0	6	63	49	0	0	0	118	0	0	0	88	9	16	2	3	...
	46-55	8	0	13	29	42	0	0	0	0	92	0	0	65	9	18	0	0	...
	56-65	29	0	6	0	43	0	0	0	0	0	78	0	22	4	10	42	0	...
	66-74	14	0	2	0	19	0	0	0	0	0	0	35	0	0	0	35	0	...
3	työssä	10	39	12	81	118	1	22	62	88	65	22	0	260	0	0	0	0	...
	yrittäjä	2	4	7	3	12	0	2	4	9	9	4	0	0	28	0	0	0	...
	työtön/muu	4	20	7	15	39	7	11	23	16	18	10	0	0	0	85	0	0	...
	eläkeläinen	35	0	5	0	39	0	0	0	2	0	42	35	0	0	0	79	0	...
	opiskelija	0	19	0	1	24	14	21	6	3	0	0	0	0	0	0	0	44	...
4	LIIKKUU	10	38	21	37	88	11	24	46	51	37	19	6	97	19	43	13	22	...
	liikkuu	18	30	6	32	85	8	22	30	39	34	28	10	101	6	24	27	13	...
	ei_liiku	23	14	4	30	56	3	10	17	28	20	31	18	62	3	15	38	9	...
5	alipaino	2	6	0	3	8	1	4	6	2	1	3	2	7	1	3	4	4	...
	normaali	13	41	0	48	78	10	27	50	43	27	17	6	98	8	41	14	19	...
	ylipaino	29	28	8	44	112	11	20	27	62	46	38	17	124	12	28	38	19	...
	YLIPAINO	6	6	23	5	32	0	5	11	11	16	19	10	29	7	12	22	2	...
6	aina_huoli	7	7	0	14	15	2	2	11	11	6	8	3	21	1	11	8	2	...
	lähes_aina	3	18	2	32	46	5	9	26	31	13	13	4	53	11	21	9	7	...
	usein	21	19	6	24	68	7	12	22	30	33	21	13	78	5	16	26	13	...
	joskus	10	20	6	18	57	3	13	19	32	19	19	6	64	4	19	17	7	...
	harvoin	3	11	9	7	36	4	13	12	9	15	9	4	32	5	12	8	9	...
	ei_koskaan	7	7	8	5	10	1	7	5	5	6	8	5	12	2	6	11	6	...
7	alipaino	0	3	0	1	8	4	3	3	1	1	0	0	4	0	2	0	6	...
	normaali	25	63	0	80	123	13	38	71	82	45	30	12	168	14	53	28	28	...
	lihavuus	24	16	8	18	67	4	11	14	22	33	33	16	65	5	19	37	7	...
	LIHAVUUS	2	0	23	1	26	0	3	6	12	11	15	5	21	9	9	12	1	...
8	--Tunto	5	15	10	9	32	1	14	16	17	12	4	7	37	4	12	10	8	...
	-Tunto	12	28	15	25	78	8	16	32	33	27	32	10	80	5	31	27	15	...
	Tunto+	25	32	5	46	79	12	21	28	48	40	25	13	99	14	30	27	17	...
	Tunto++	9	7	1	20	34	1	4	17	18	11	16	4	41	4	9	13	4	...

Burtin matriisi muistuttaa korrelaatiomatriisia (ks. tuloste 3.11, s. 79). Lävistäjän ykkösiä vastaavat frekvenssijakaumat ja korrelaatioita ristiintaulukot. Toisaalta se muistuttaa hajontakuvamatriisia (ks. kuva 7.1, s. 172), jossa hajontakuvien tilalla ovat ristiintaulukot.

Burtin matriisin haittapuolena voidaan pitää sen laajuutta: tulosteen 7.6 matriisissa on kaikkiaan 1 444 lukua, mikä on yli 20-kertainen määrä verrattuna kahdeksan muuttujan korrelaatiomatriisin 64 lukuun. Edut menevät kuitenkin haittojen edelle, koska Burtin matriisin avulla voidaan analysoida muitakin kuin lineaarisia riippuvuuksia. Ryhmäkoodi ja työllisyystilanne ovat luokittelutasoisia muuttujia, joiden analysointiin hajontakuvat ja korrelaatiot soveltuvat huonosti, jos lainkaan.

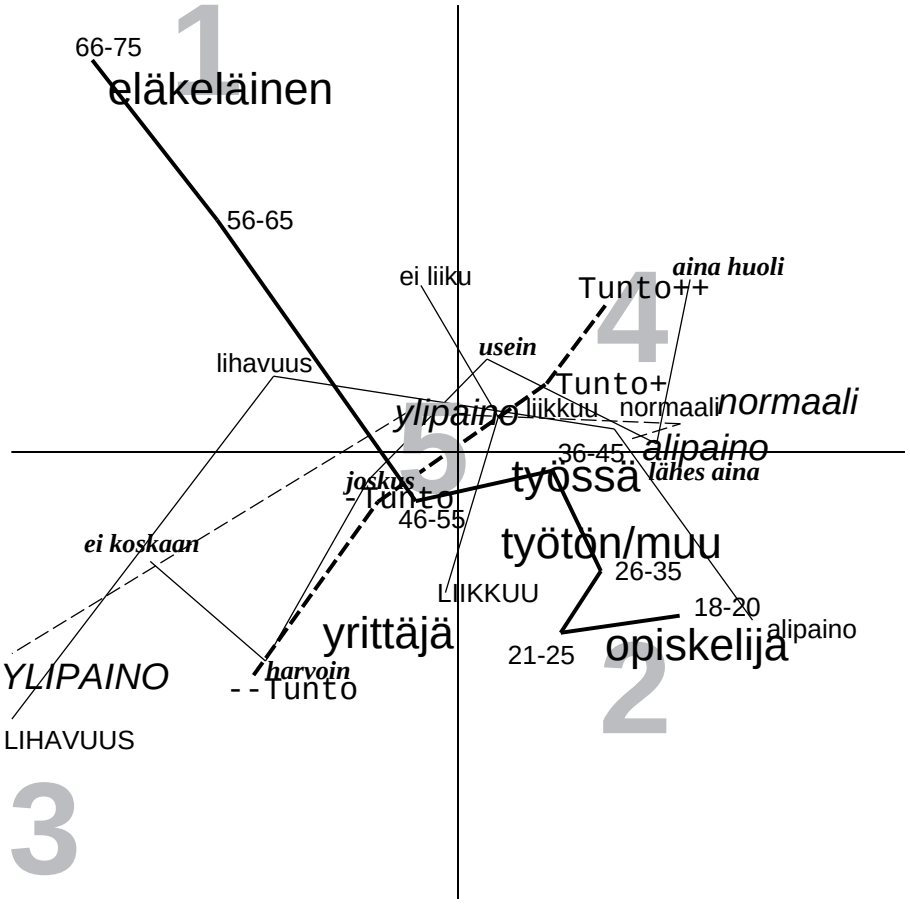
7.3.4 Usean muuttujan kuva

Burtin matriisiin sisältyviä yhteyksiä voidaan kuvata *usean muuttujan korrespondenssianalyysilla*. Tämä korrespondenssianalyysin yleistys soveltuu etenkin laajojen, luokitteluihin perustuvien kyselyaineistojen kuvaamiseen. Menetelmää on kehitetty ja sovellettu erityisesti Ranskassa, jossa sitä toisinaan kutsutaan ”luokiteltujen muuttujien faktorianalyysiksi”.

Yhteiskuntatieteissä analyysia on tehnyt tunnetuksi ranskalainen sosiologi Pierre Bourdieu, esimerkiksi teoksessaan *Distinction* (Bourdieu, 1984). Menetelmän perusteisiin johdattaa muun muassa Greenacre (2007). Pidemmälle meneviin mahdollisuuksiin ja muunnelmiin perehdyttää laaja teos *Multiple Correspondence Analysis and Related Methods* (Greenacre & Blasius, 2006).

Usean muuttujan korrespondenssianalyysissa numeeriset tulokset kiinnostavat yleensä vielä vähemmän kuin tavallisessa korrespondenssianalyysissa. Tavoitteena on jälleen visualisoida olennaiset tiedot kaksiulotteisena kuvana. Akseleiden lukuarvot on tapana jättää pois, sillä useiden luokiteltujen muuttujien samanaikaisessa visualisoinnissa kuvaan tulee helposti ”ruuhkaa”. Kiinnostavia ovat vain luokkien ja luokittelijoiden väliset suhteet ja niiden sijoittuminen ”korrespondenssiavaruuteen”.

Kuva 7.9 tiivistää usean muuttujan korrespondenssianalyysin keskeiset tulokset visualisoimalla Burtin matriisiin (tuloste 7.6) sisältyvät yhteydet. Kuvaa tulkitaan samalla tavalla kuin kahden muuttujan kuvaa (ks. kuva 7.8, s. 188), nyt vain muuttujien määrä on nelinkertainen ja luokkien määrä noin viisinkertainen.



Kuva 7.9. Kahdeksan muuttujan korrespondenssianalyysi.

Ryhmäkoodia ja työllisyystilannetta lukuun ottamatta muuttujien luokat on kuvassa 7.9 yhdistetty erilaisilla viivoilla. Ne visualisoivat yhteyksiä, jotka ovat vain osittain lineaarisia. Esimerkiksi ikä vaihtelee kuvan oikeasta alakulmasta vasempaan yläkulmaan melko lineaarisesti; vain nuoremmissa ikäluokissa on vähän mutkittelua.

Ryhmässä 2 on enemmän nuorempia, ryhmässä 1 vanhempia vastaajia. Tämän voi todeta myös helposti Burtin matriisin (tuloste 7.6) vasemmasta yläkulmasta, iän ja ryhmäkoodin välisistä frekvenssitaulukoista. Yhteys ilmeni myös aiemmin piirretystä erotteluavaruudesta (kuva 7.5, s. 180), samoin kuin ryhmän 3 vastaajille ominainen, suurempi paino. Vuoden 2005 havainnoista muodostuva ryhmä 5 sijoittuu kuvan keskelle, koska sen havaintoja ei ole ryhmitelty. Jos ryhmittely tehtäisiin, voitaisiin nähdä mahdolliset vastaavuudet vuoden 1997 ryhmiin 1, 2, 3 ja 4. Edellä havaittujen yhteyksien perusteella vastaavuuksia varmasti löytyisi.

Kuvasta 7.9 voi nähdä myös muita mielenkiintoisia seikkoja. Lihavuuteen taipuvaisten itsetunto ulkonäköasioissa on huono, mutta he eivät ole painostaan huolissaan; enemmän huolissaan ovat ne, joilla on hyvä itsetunto. Yrittäjät ja opiskelijat harrastavat eniten liikuntaa; yrittäjät ovat oman käsityksensä mukaan ylipainoisempia muihin verrattuna ja opiskelijat vastaavasti normaali- tai jopa alipainoisia. Painon arviointi itse luonnehdittuna näyttää menevän hyvin yksiin painoindeksiluokituksen kanssa.

Osa kuvasta havaituista yhteyksistä perustuu pieniin frekvensseihin. Burtin matriisia olisi syytä tiivistää paremmilla luokituksilla. Kaikkiaan tässä esitetyt tulkinnat ovat jälleen pinnallisia, mutta tutkittavan ilmiön tunteville kuvan 7.9 tapaiset esitykset tarjoavat hyvän pohjan sisällöllisiin tulkintoihin.

Korrespondenssianalyysin kuvien perusteella hahmotettujen riippuvuuksien tarkastelua voidaan jatkaa *frekvenssiaineistojen* analyysimenetelmillä. Niitä ovat muun muassa kohdassa 5.2.1 (s. 126) mainittuihin, yleistettyihin lineaarisiin malleihin kuuluvat *log-lineaariset mallit*. Johdatuksen näihin malleihin esittää Alkula ym. (1994, 220–232). Mallien yhteyksistä korrespondenssianalyysiin kertoo esimerkiksi Greenacre & Blasius (2006).

Kirjan menetelmällinen sisältö päättyy tähän. Liitteessä A perehdytään ohjelmistoihin ja dokumentointiin sekä tutustutaan kirjassa esitettyjen kuvien ja tulosteiden työkaavioihin.

A Ohjelmistot ja dokumentointi

Kirjan varsinaiset luvut keskittyivät mittaukseen ja tiedonkeruuseen, aineiston esikäsittelyyn sekä tilastollisiin malleihin ja menetelmiin. *Ohjelmistot* olivat esillä vain tulosteiden ja kuvien välityksellä. Käytännössä tulosteet ja kuvat eivät riitä; on myös tiedettävä, miten ne on saatu aikaan. *Dokumentointi* on avain toistettavuuteen ja laadukkaaseen työskentelyyn.

Tässä liitteessä perehdytään dokumentoivaan työskentelytapaan kirjassa käytettyjen ohjelmistojen avulla. Esimerkkeinä tarkastellaan aineiston perustamista sekä tulosteiden ja kuvien työkaavioita.

A.1 Ohjelmistot

Tässä kirjassa käytetyt ohjelmistot ovat Survo (SURVO MM, ver. 3.00) ja SPSS (SPSS 15.0 for Windows). Survo on suomalainen, professori Seppo Mustosen kehittämä tietojenkäsittely-ympäristö (Survo Systems, 2007) ja SPSS on amerikkalainen, alun alkaen Norman H. Nien, C. Hadlai Hullin ja Dale H. Bentin kehittämä tilastollisten ohjelmien paketti (SPSS, 2007).

Survo on kirjan keskeinen työväline, jolla on kirjoitettu tekstit, piirretty kuvat ja tehty suurin osa analyyseista. SPSS:llä on tehty osa luvun 3 taulukoista sekä luvun 5 regressio- ja varianssianalyyseista.

Kirja on taitettu Survon ja LaTeX-ladontaohjelmiston (Lamport, 1994; Mittelbach & Goossens, 2004) yhteistyönä, tekniikalla, josta kertoo yksityiskohtaisemmin ja värikkäämmin Vehkalahti (2007).

A.1.1 Survo ja SPSS

Survon alaa ovat tilastolliset analyysimenetelmät ja kuvat, aineistojen hallinta sekä julkaisujen ja sovellusten laatiminen (Mustonen, 2001). Ohjelmiston nimi johtuu sanasta *survey* tai tiedon tiivistämiseen viittaavasta verbistä *survoa* (Mustonen, 1992, 2).

Survon varhaisimmat 1960-luvun versiot ovat osa suomalaisen tietojenkäsittelyn kiehtovaa esihistoriaa, jota valottaa muun muassa Mustonen (2007). Nyky-Survo sai alkunsa vuonna 1979 *nuotinpainatuksesta*. Idea tekstinkäsittelyyn perustuvasta tavasta käyttää tietokonetta myös tilastollisiin analyyseihin syrjäytti valikot silloisessa SURVO 76:ssa, joka oli maailman ensimmäisiä interaktiivisia tilastollisia ohjelmistoja (Mustonen, 1980, 1996, 2001).

Myös SPSS:n historia ulottuu 1960-luvulle, jolloin se tuli tunnetuksi erityisesti yhteiskuntatieteisiin sopivana, tilastollisten sovellusohjelmien pakettina, *"Statistical Package for Social Sciences"*. Nykyään ohjelmistosta käytetään vain lyhennenimeä.

SPSS:n perinteistä alaa ovat edelleen yhteiskuntatieteissä keskeiset ristiintaulukointi sekä regressio- ja varianssianalyysit. Menetelmävalikoiman suhteen SPSS on yksi laajimmista tilastollisista ohjelmistoista, mikäli käytössä ovat peruspaketin ohella tyypillisimmät laajennusosat (mm. *Advanced Models*, *Categories* ja *Missing Values*).

Käyttöliittymä

Pintapuolisesti Survo ja SPSS saattavat vaikuttaa aivan erilaisilta, mutta pohjimmiltaan niissä on samoja yhtäläisyyksiä kuin muissakin tilastollisten aineistojen analysointiin soveltuvissa ohjelmistoissa:

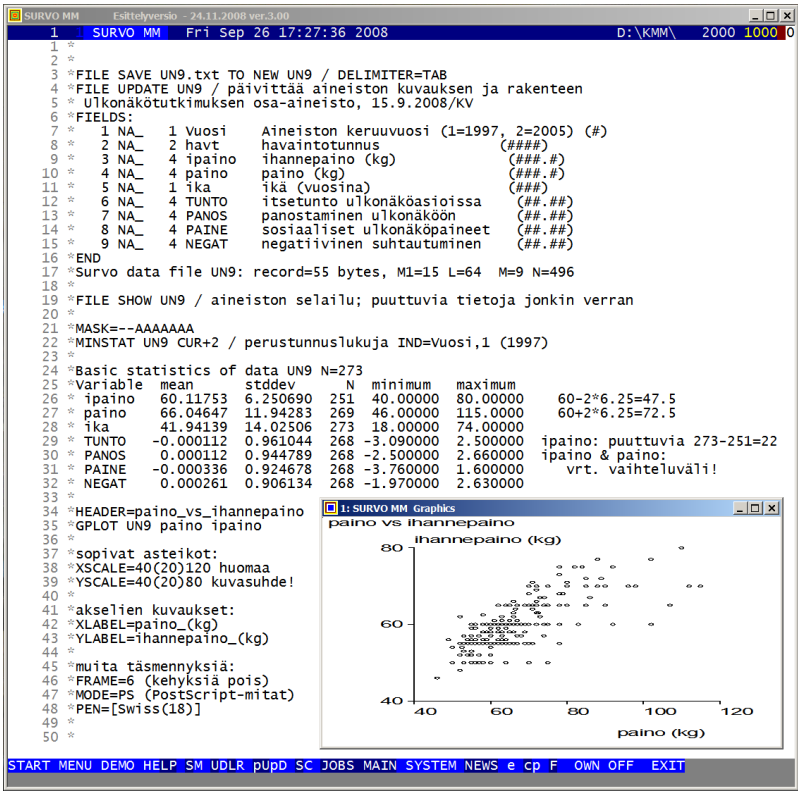
- Asiat ilmaistaan täsmällisesti komennoin ja avainsanoin.
- Työskentelyn apuna hyödynnetään lisäksi valikoita.
- Toimintoja ohjataan näppäimistöllä tai hiirellä.

Esimerkkejä puhtaasti hiirellä ohjattavista toiminnoista ovat Survon graafinen faktorirotaatio ja SPSS:n kuvien muokkaus. Työvaiheiden toistamisen ja dokumentoinnin kannalta keskeisimpiä ovat kuitenkin komennot ja avainsanat.

Survo

Survo on tekstinkäsittelyyn perustuvan käyttöliittymänsä ansiosta näyttänyt varsin samanlaiselta aikakausista ja versioista riippumatta.

Näkymä A.1 kuvaa Survon pääikkunaa ja siihen sijoittuvaa *toimituskenttää*, josta ohjataan kaikkia Survon toimintoja sekä käsitellään niin tekstiä, taulukoita, lukuja, lausekkeita ja aineistoja kuin matriiseja, komentoja, kuvanpiirto- ja muita kaavioita sekä näiden tulosteita.



Näkymä A.1. Survo-ohjelmiston käyttöliittymä.

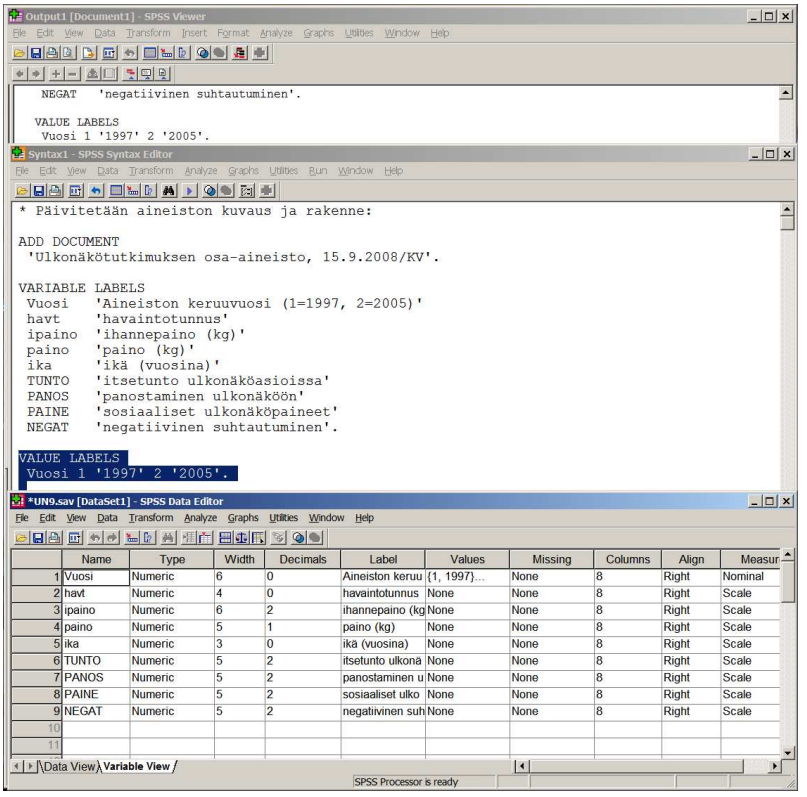
Komento *aktivoidaan* joko `Esc`-näppäimellä tai hiiren kaksoispainalluksella. Mahdollinen tuloste ilmestyy samaan toimituskenttään komennossa osoitetulta riviltä alkaen.

Survo muodostaa oman käyttöympäristön, jonka pelisääntöihin voi tutustua asennus- ja käyttöönotto-oppaasta (Mustonen, 2003) tai Survon ilmaisversion ja opetusohjelmien välityksellä.

SPSS

SPSS on mukauttanut käyttöliittymänsä kulloiseenkin ympäristöön. Windowsissa SPSS näyttää tyypilliseltä Windows-ohjelmalta.

Näkymä A.2 kuvaa SPSS:n pääikkunaa *(Data Editor)* ja aineiston muuttujia *(Variable View)*. Vaihto havaintoihin *(Data View)* tapahtuu hiirellä tai näppäimellä `Ctrl-T`. Näkymään sisältyvät myös komentoikkuna *(Syntax Editor)* ja tulosteikkuna *(Viewer)*.



Näkymä A.2. SPSS-ohjelmiston käyttöliittymä.

SPSS-komento suoritetaan maalaamalla se komentoikkunassa ja painamalla näppäintä `Ctrl-R` *(Run Current)*. Vaikutus ilmenee joko tulosteikkunassa tai pääikkunassa.

SPSS-oppaita on runsaasti, mutta useimmissa korostetaan vain valikkoja. Dokumentointiin tarvittavat komennot saa komentoikkunaan, kun valintojen lopuksi painaa OK-näppäimen sijasta *Paste*.

A.1.2 Aineiston perustaminen

Luvun 2 lopussa mainittiin lyhyesti, millä tavoin kyselytutkimusaineisto siirretään tiedonkeruussa käytetyiltä lomakkeilta ohjelmistoilla käsiteltävään muotoon. Seuraavassa perehdytään aiheeseen yksityiskohtaisemmin Survon ja SPSS:n avulla.

Esimerkkinä tarkasteltava tekstitiedosto `UN9.txt` muodostuu ulkonäkötutkimuksen 496 havainnosta ja yhdeksästä, erityisesti luvuissa 5–7 analysoidusta muuttujasta. Listauksessa A.1 tästä osaaineistosta on näkyvissä 25 havaintoa: 13 alusta, viisi keskeltä ja seitsemän lopusta. Jokainen havainto muodostaa tekstitiedostossa yhden rivin. Ensimmäisellä rivillä ovat muuttujien nimet.

Listaus A.1. Ulkonäkötutkimuksen aineistoa tekstimuodossa.

```
Vuosi|havt|ipaino|paino|ika|TUNTO|PANOS|PAINE|NEGAT
    1|1003| 65.00| 75.0| 53|-0.33|-1.36| 0.58| 0.30
    1|1004| 67.00| 86.0| 55| 0.23| 0.35| 0.77|-0.32
    1|1006| 60.00| 64.0| 36| 0.29| 0.37|-0.46|-1.15
    1|1008| 58.00| 63.0| 34| 1.06| 0.10| 0.98|-1.07
    1|1009| 46.00| 46.0| 26| 0.86| 0.75| 0.22|-0.85
    1|1010| 53.00| 53.0| 74| 0.02|-0.71| 0.69| 0.91
    1|1011| 60.00| 70.0| 71| 0.33| 0.09|-1.23|-0.21
    1|1015| 60.00| 72.0| 21|-1.34| 0.79| 0.88| 1.06
    1|1016| 71.00| 80.0| 67|-0.98|-0.02| 0.96| 0.84
    1|1017| 60.00| 55.0| 24|-1.48| 0.38| 0.66|-0.48
    1|1018| 60.00|102.0| 27|-1.19| 0.87|-1.05|-0.29
    1|1019|     -| 60.0| 22| 1.82|-2.36| 0.49| 1.38
    1|1020| 58.00| 65.0| 37|-1.86|-0.20|-1.84| 0.68
    .         .      .        .     .     .     .
    .         .      .        .     .     .     .
    .         .      .        .     .     .     .
    2|1672| 58.00| 58.0| 29| 1.62|-0.15|-0.62| 0.22
    2|1673|     -|    -| 53|    -|    -|    -|    -
    2|1680| 65.00| 90.0| 67|-0.20| 0.18|-2.43|-0.70
    2|1681| 64.00| 64.0| 40| 1.54|-1.23|-3.88| 1.04
    2|1685| 60.00| 63.0| 19| 0.74|-0.37| 0.43|-0.58
    .         .      .        .     .     .     .
    .         .      .        .     .     .     .
    .         .      .        .     .     .     .
    2|1965| 60.00| 61.0| 65| 1.11|-0.64|-1.12| 0.70
    2|1969| 70.00| 78.0| 24|-1.00|-0.47| 0.77|-1.25
    2|1970| 58.00| 68.0| 31| 1.04|-0.62| 0.79|-1.31
    2|1977| 56.00| 47.0| 66|-0.46|-0.39| 0.47| 0.13
    2|1988| 55.00| 63.0| 58|-0.40|-0.73| 0.46|-0.71
    2|1991| 65.00| 65.0| 25|-0.74| 0.13| 0.29|-1.08
    2|1998| 58.00| 65.0| 38| 0.14| 1.03|-1.18| 0.42
```

Tiedot ovat järjestyksessä keruuajankohdan (`Vuosi`) ja havaintotunnuksen (`havt`) mukaan. Puuttuvia tietoja on merkitty viivalla (`-`). Todellisuudessa muuttujien arvoja erottavat *tab*- eli sarkainmerkit, mutta havainnollisuuden vuoksi ne on listauksessa A.1 korvattu pystyviivoilla (`|`).

Seuraavassa tiedot siirretään tekstitiedostosta ohjelmien omiin tiedostoihin. Esimerkki on keinotekoinen, koska siinä rajoitutaan vain yhdeksään muuttujaan, joista osa on luotu vasta aineiston esikäsittelyn ja analyysien myötä. Käytännössä siirto tapahtuisi vastaavasti.

Survo

Kaaviossa A.1 tiedot siirretään ja talletetaan tekstitiedostosta Survon havaintotiedostoksi `FILE SAVE` -komennolla.

Kaavio A.1. Aineiston perustaminen (Survo).

```
FILE SAVE UN9.txt TO NEW UN9 / DELIMITER=TAB

FILE STATUS UN9 / näyttää aineiston kuvauksen ja rakenteen
 Copied from text file UN9.txt
FIELDS: (active)
   1 NA_   1 Vuosi    (######)
   2 NA_   2 havt     (####)
   3 NA_   4 ipaino   (###.##)
   4 NA_   4 paino    (###.#)
   5 NA_   1 ika      (###)
   6 NA_   4 TUNTO    (##.##)
   7 NA_   4 PANOS    (##.##)
   8 NA_   4 PAINE    (##.##)
   9 NA_   4 NEGAT    (##.##)
END
Survo data file UN9: record=55 bytes, M1=15 L=64  M=9 N=496
```

Perustetun aineiston rakenne ilmenee kaavion alaosasta, jonka on tuottanut `FILE STATUS` -komento. `FILE SAVE` on automaattisesti päätellyt, että kaikki muuttujat ovat numeerisia, osa kokonais- ja osa desimaalilukuja. Kaavioon sisältyvät yksityiskohdat selviävät muun muassa Survon käyttöoppaasta (Mustonen, 1992, 86–88).

Kaaviossa A.2 aineistoa on dokumentoitu lisäämällä muuttujien sanalliset kuvaukset sekä päivämäärällä varustettu, aineiston sisältöä kuvaava kommentti. Rakenne päivittyy aktivoimalla kaavion yläpuolelle vaihdettu `FILE UPDATE` -komento.

Kaavio A.2. Aineiston dokumentointi (Survo).

```
FILE UPDATE UN9 / päivittää aineiston kuvauksen ja rakenteen
 Ulkonäkötutkimuksen osa-aineisto, 15.9.2008/KV
FIELDS:
   1 NA_   1 Vuosi    Aineiston keruuvuosi (1=1997, 2=2005) (#)
   2 NA_   2 havt     havaintotunnus              (####)
   3 NA_   4 ipaino   ihannepaino (kg)             (###.#)
   4 NA_   4 paino    paino (kg)                   (###.#)
   5 NA_   1 ika      ikä (vuosina)                (###)
   6 NA_   4 TUNTO    itsetunto ulkonäköasioissa    (##.##)
   7 NA_   4 PANOS    panostaminen ulkonäköön       (##.##)
   8 NA_   4 PAINE    sosiaaliset ulkonäköpaineet   (##.##)
   9 NA_   4 NEGAT    negatiivinen suhtautuminen    (##.##)
END
```

SPSS

Kaaviossa A.3 tiedot siirretään tekstitiedostosta SPSS-muotoon komennolla `GET DATA`. Komento on saatu *File*-valikon *Read Text Data* -toiminnolla, joka vaiheittain, osittain automaattisesti, selvittää aineiston rakenteen kuusisivuisella lomakkeella.

Kaavio A.3. Aineiston perustaminen (SPSS).

```
GET DATA  /TYPE = TXT
 /FILE = 'D:\KMM\UN9.txt'  /DELCASE = LINE /DELIMITERS = "\t"
 /ARRANGEMENT = DELIMITED  /FIRSTCASE = 2  /IMPORTCASE = ALL
 /VARIABLES =
 Vuosi F6.0 havt F4.0 ipaino F6.2 paino F5.1 ika F3.0
 TUNTO F5.2 PANOS F5.2 PAINE F5.2 NEGAT F5.2.
EXECUTE.
DATASET NAME DataSet1 WINDOW=FRONT.

SAVE OUTFILE='D:\KMM\UN9.sav'
 /COMPRESSED.
```

Suorittamalla kaavion komennot aineisto haetaan SPSS:n pääikkunaan ja talletetaan SPSS:n havaintotiedostoksi.

Kaaviossa A.4 aineistoa on dokumentoitu lisäämällä päivämäärällä varustettu, aineiston sisältöä kuvaava kommentti sekä muuttujien sanalliset kuvaukset. `Vuosi`-muuttujalle on asetettu myös sen yksittäisten arvojen kuvaukset.

Kaavio A.4. Aineiston dokumentointi (SPSS).

```
* Päivitetään aineiston kuvaus ja rakenne:

ADD DOCUMENT
 'Ulkonäkötutkimuksen osa-aineisto, 15.9.2008/KV'.

VARIABLE LABELS
 Vuosi   'Aineiston keruuvuosi (1=1997, 2=2005)'
 havt    'havaintotunnus'
 ipaino  'ihannepaino (kg)'
 paino   'paino (kg)'
 ika     'ikä (vuosina)'
 TUNTO   'itsetunto ulkonäköasioissa'
 PANOS   'panostaminen ulkonäköön'
 PAINE   'sosiaaliset ulkonäköpaineet'
 NEGAT   'negatiivinen suhtautuminen'.

VALUE LABELS
 Vuosi 1 '1997' 2 '2005'.
```

A.1.3 Dokumentoiva työskentelytapa

Edellä esitetyt kaaviot ovat esimerkkejä *dokumentoivasta työskentelytavasta*, joka helpottaa olennaisesti työvaiheiden toistamista. Toistamista on mahdoton välttää, sillä se kuuluu tutkimustyön luonteeseen. Tuloksiin ei yleensä päästä suorinta tietä, vaan siihen vaaditaan lukuisia yrityksiä ja erehdyksiä. Yksinkertaisimmillaan samoja työvaiheita toistetaan, kun jotain analyysia tehdään eri muuttujilla tai eri havainnoilla.

Lisää toistoa on tiedossa, kun aineistosta löytyy virheitä. Tällöin joudutaan yleensä uusimaan monia, virheen korjausta edeltäneitä työvaiheita, kuten muunnoksia, muokkauksia ja analyyseja. Laajimmillaan toistaminen on, kun aineisto vaihtuu. Monessa tutkimuksessa halutaan toistaa olennaisesti samat, aiemmin tehdyt vaiheet. Mitä paremmin työvaiheet on dokumentoitu, sitä helpompaa se on.

Valikot ovat käteviä kertaluonteisissa töissä, mutta tutkimustyössä vain harvat tehtävät ovat kertaluonteisia. Valikkojen varassa työskentelemällä toistaminen on hidasta ja turhauttavaa, eikä työskentelystä jää yleensä mitään jälkiä.

Dokumentoiva työskentelytapa jättää jäljet, joita seuraamalla voi myöhemmin jäljittää oman ajatusprosessinsa. Tuloksista näkee, *mitä* kaikkea on tehty, mutta jäljistä näkee, *miten* kaikki on tehty.

A.2 Kuvien ja tulosteiden työkaavioita

Esimerkkeinä dokumentoivasta työskentelytavasta tarkastellaan kirjan kuvien ja tulosteiden tekemisessä käytettyjä Survo- ja SPSS-työkaavioita. Kaikkiaan kirjassa on yli 80 kuvaa tai tulostetta; vain muutaman työkaavio on voitu mahduttaa tähän. Kaikkia koskee kuitenkin sama periaate: kun kuva tai tuloste pitää tuottaa uudelleen, työ ei ala alusta, vaan se jatkuu siitä, mihin on viimeksi jääty.

Käytännössä periaate tarkoittaa, että työkaavion toiminnot ainoastaan aktivoidaan tai suoritetaan uudelleen, jolloin tuloksena saadaan uusi versio kuvasta tai tulosteesta. Kaavioissa on melko vähän kommentteja, sillä ne dokumentoivat pitkälti itse itsensä.

Kaaviossa A.5 tutkitaan jakaumia ja tunnuslukuja Survon `STAT`-komennolla. Tulokset tulevat komentorivin alapuolelle (`CUR+1`). Täsmennys `VARS` luettelee muuttujat (tässä vain `sv`), ja `CLASSMAX` määrää luokkien maksimimäärän automaattisessa luokittelussa. Oletusarvo on 30, joten täsmennys ei olisi tässä välttämätön.

Kaavio A.5. Tuloste 3.1, sivu 53 (Survo).

```
STAT MV2007A CUR+1 / VARS=sv CLASSMAX=30
```

Kaavion A.6 SPSS-komennot tarkenteineen saadaan komentoikkunaan `Paste`-toiminnolla valitsemalla ensin valikoista kohdat *Analyze – Descriptive Statistics – Descriptives* ja muuttujaksi `sv`. Tunnuslukujen (rivi 3) osalta valinnat vastaavat *Options*-alakohdassa tarjolla olevia oletuksia. Piste rivin 3 lopussa päättää SPSS-komennon.

Kaavio A.6. Tuloste 3.2, sivu 55 (SPSS).

```
1 DESCRIPTIVES
2   VARIABLES=sv
3   /STATISTICS=MEAN STDDEV MIN MAX .
```

Kaaviossa A.7 tehdään Survon `TAB`-komennolla ristiintaulukko muuttujista `yhdessa` ja `pari`, jotka annetaan täsmennyksellä `VARIABLES`.

Halutunlaiset luokittelut muuttujille ilmaistaan omina täsmennyksinään riveillä 2 ja 3. Ensimmäinen luku on luokan alaraja, loput ylärajoja. Suluissa annetaan luokille nimet.

Avainsana `MISSING` sisällyttää mukaan puuttuvien tietojen lukumäärät. Täsmennykset `CHI2` ja `LABELS` jättävät tulostuksesta pois turhia osia kuten riippumattomuustestin ja taulukon otsikoita.

Kaavio A.7. Tuloste 3.5, sivu 60 (Survo).

```
1 TAB MV2007A END+2 / VARIABLES=yhdessa,pari   CHI2=- LABELS=0
2   yhdessa=0,0(ei),53(on),MISSING(?)
3      pari=1,1(ei),02(on),MISSING(?)
```

Kaavio A.8 piirtää Survossa histogrammin. Tällaiset kaaviot syntyvät vaiheittain. Alkutilanteessa kaavio on paljon yksinkertaisempi.

Varsinainen komento on rivin 2 `HISTO`. Täsmennyksiä ja kommentteja saa esiintyä vapaasti komennon ympärillä (kuten riveillä 1 ja 4). Olennaisinta on, että kuva syntyy täsmälleen sellaisena kuin halutaan eikä sitä tarvitse muokata jälkeenpäin.

Kaavio A.8. Kuva 3.2, sivu 61 (Survo).

```
1 IND=yhdessa,0.01,53 (rajataan nollat pois)
2 HISTO MV2007A yhdessa / yhdessa=0(1)55 FILL=NO DEVICE=PS,K.PS
3 XSCALE=-3:?,0(5)55 YSCALE=0(5)20 HEADER=   FRAME=6
4 Kuvan koko ja mittasuhteet: SIZE=1200,600 XDIV=2,9,1 YDIV=0.5,5.5,0
5   *pen=[Swiss(10)][move(0,0)][rot(0)] PEN=*pen
6 XLABEL=yhdessäoloaika_(vuosina)  LINETYPE=*pen
7 YLABEL=[move(-140,-450)][rot(90)],vastaajien_lukumäärä
```

Kaaviota A.8 hyödynnetään myös kuvan 3.3 (s. 62) piirtämisessä. Ainoat muutettavat kohdat ovat `yhdessa`-muuttujan luokittelu rivillä 2 ja `YSCALE`-täsmennys rivillä 3.

Myös kaavio A.9 pohjautuu kaavioon A.8. Luokkafrekvenssit on taulukoitu `TAB`-komennolla, nimetty taulukko dataksi `YHD5` ja piirretty se `PLOT`-komennolla. Loput täsmennykset ovat samat kuin kaaviossa A.8 (rivit 3–7), vain x-akselin nimeä on siirretty vähän vasemmalle samaan tapaan kuin y-akselin nimeä kaavion A.8 rivillä 7.

Kaavio A.9. Kuva 3.4, sivu 63 (Survo).

```
TAB MV2007A CUR+3 / VARIABLES=yhdessa IND=yhdessa,0.01,53 (nollat pois)
yhdessa=0.1,4.9(alle_5),15(5_-_15),25(16_-_25),35(26_-_35),53(yli_35)

DATA YHD5
yhdessa     f
alle_5     71
5_-_15    120
16_-_25    91
26_-_35    65
yli_35     37

PLOT YHD5 / TYPE=VBAR DEVICE=PS,K.PS
LEGEND=- SHADING=7 VALUES=[Swiss(9)],##%,1
```

Kaavio A.10 sisältää vastaavia SPSS-komentoja; periaate on vain hieman toisenlainen. Frekvenssitaulukon sijaan tehdään riveillä 1–6 uusi luokiteltu muuttuja (YHD5). Komento RECODE saadaan valikosta *Transform – Recode into Different Variables*. Riveillä 7–14 nimetään muuttuja ja sen luokat, minkä jälkeen piirretään pylväskuva.

Kaavio A.10. Kaaviota A.9 vastaavia komentoja (SPSS).

```
DO IF (yhdessa ~= 0) .
RECODE
  yhdessa
  (0.1 thru 4.9=1)  (5 thru 15=2)  (16 thru 25=3)
  (26 thru 35=4)  (35.1 thru Highest=5)  INTO  YHD5 .
END IF .
VARIABLE LABELS YHD5 'yhdessäolo luokiteltuna'.
VALUE LABELS YHD5 /* annetaan luokille nimet */
  1 'alle 5'
  2 '5 - 15'
  3 '16 - 25'
  4 '26 - 35'
  5 'yli 35'.
EXECUTE .
          /* pylväskuva, kaksi erilaista versiota: */
GRAPH
  /BAR(SIMPLE)=COUNT BY YHD5 .

IGRAPH /VIEWNAME='Bar Chart'
 /X1 = VAR(YHD5) TYPE = SCALE
 /Y = $count /COORDINATE = VERTICAL
 /X1LENGTH=3.0 /YLENGTH=3.0 /X2LENGTH=3.0 /CHARTLOOK='NONE'
 /BAR KEY=ON SHAPE = RECTANGLE BASELINE = AUTO. EXE.
```

Kuva piirretään kahdella tavalla. Rivin 16 GRAPH-komennon tuottaa valinta *Graphs – Legacy Dialogs – Bar* ja rivin 19 IGRAPH-komennon valinta *Graphs – Interactive – Bar*. Vaihtoehdot edustavat SPSS:n eri kehitysvaiheita. Vanhempi GRAPH osaa hyödyntää luokkien nimiä, mutta sitä ei voi ohjata komennoilla paljoa pidemmälle. Uudempi IGRAPH tarjoaa enemmän mahdollisuuksia, muttei oletuksena hyödynnä nimiä, vaan sijoittaa pylväiden alle luvut 1–5.

Molemmissa tapauksissa kuvan viimeistely edellyttäisi käsin tapahtuvaa muokkausta. Koska sellaisten työvaiheiden dokumentointimahdollisuudet ja toistettavuus ovat olemattomia, ei tässä kirjassa ole yhtään SPSS:llä piirrettyä kuvaa.

Kaaviossa A.11 asetetaan mittarin k26 muuttujille sanalliset selitteet ja taulukoidaan kahden muuttujan frekvenssijakaumat niin, että taulukkoon tulevat sekä luvut että selitteet. Komento FREQUENCIES saadaan valikosta *Analyze – Descriptive Statistics – Frequencies*.

Kaavio A.11. Tuloste 3.7, sivu 66 (SPSS).

```
1  VALUE LABELS k26.1 TO k26.22
2    1 'Täysin eri mieltä'
3    2 'Osin eri mieltä'
4    3 'Ei samaa eikä eri'
5    4 'Osin samaa mieltä'
6    5 'Täysin samaa mieltä'.
7  SET TNumbers Both. /* taulukkoon sekä luvut että selitteet */
8  FREQUENCIES
9    VARIABLES=k26.3 k26.18
10   /ORDER=  ANALYSIS.
```

SET-komennolla voi säätää monia asetuksia. Tarkempia tietoja on SPSS:n käsikirjassa (*Help – Command Syntax Reference*). Kaaviossa A.12 muuttujat taulukoidaan vastakkain CROSSTABS-komennolla (*Analyze – Descriptive Statistics – Crosstabs*).

Kaavio A.12. Tuloste 3.9, sivu 68 (SPSS).

```
1  SET TNumbers Values. /* tiiviimpi taulukko ilman selitteitä */
2  CROSSTABS
3    /TABLES=k26.18  BY k26.3
4    /FORMAT= AVALUE TABLES
5    /CELLS= COUNT
6    /COUNT ROUND CELL.
```

Kaaviossa A.13 lasketaan ja taulukoidaan puuttuvia tietoja. Kun rivin 14 kaksoispiste korvataan pilkulla, saadaan vastaava sisäkkäinen taulukko (tuloste 3.13).

Kaavio A.13. Tuloste 3.12, sivu 84 (Survo).

```
*mask1=-AAAAAAAAAAAAAAAAAAAA------------------------------
*mask2=----------------------AAAAAAAAAAAAAAAAAAA-----------
*mask3=-------------------------------------------AAAAAAAAAAA

VARSTAT MV2007I,Mittari1:1,#VAL,1,5 / MASK=*mask1
VARSTAT MV2007I,Mittari2:1,#VAL,1,5 / MASK=*mask2
VARSTAT MV2007I,Mittari3:1,#VAL,1,5 / MASK=*mask3
Luokittelut: Mittari1=0,11(poistetaan),21(paikataan),22(jätetään)
             Mittari2=0,10(poistetaan),19(paikataan),20(jätetään)
             Mittari3=0,05(poistetaan),10(paikataan),11(jätetään)

Katsotaan samalla vuosittain: Vuosi=1,1(1997),2(2005) LABELS=0
                                        RESULTS=RSUMS,CSUMS CHI2=-
TAB MV2007I CUR+3 / VARIABLES=Vuosi:Mittari1,Mittari2,Mittari3
```

Kaaviossa A.14 tehdään faktorianalyysi ja graafinen rotaatio, talletetaan työvaiheita ja otetaan tulos esiin tulkintaa varten.

Kaavio A.14. Tuloste 4.2, sivu 101 (Survo).

```
Valitaan muuttujat: MASK=#7(CGQ) ja havainnot: IND=Vuosi,1 (1997)
CORR UN            / korrelaatiot (+ keskiarvot ja hajonnat)
MAT R1997=CORR.M   / korrelaatiomatriisi talteen
FACTA R1997,4      / faktorointi (maximum likelihood, 4 faktoria)
MAT F1997=FACT.M   / rotatoimaton faktorimatriisi talteen
ROTATE F1997,4     / graafinen rotaatio: ROTATION=GRAPHICAL
MAT T1997=TFACT.M  / rotaation muunnosmatriisi talteen
MAT G1997=AFACT.M  / rotatoitu faktorimatriisi talteen
/LOADFACT UN G1997 / tulos järjestettynä ja korostettuna + kuvaukset
```

Kaaviossa A.15 testataan regressiomallin jäännösten normaalisuutta Kolmogorovin ja Smirnovin ei-parametrisella testillä.

Kaavio A.15. Tuloste 5.10, sivu 147 (SPSS).

```
NPAR TESTS /* Analyze - Nonparametric Tests - 1-Sample K-S */
  /K-S(NORMAL)= Jäännös
  /STATISTICS DESCRIPTIVES QUARTILES
  /MISSING ANALYSIS.
```

Kaaviossa A.16 piirretään naamakuva kymmenen havainnon aineistosta TÄHDET, jonka tiedot on esitetty taulukossa 6.1 (s. 154).

Kaavio A.16. Kuva 6.3, sivu 156 (Survo).

```
1  SIZE=1200,800 XDIV=0,1,0 YDIV=0,1,0 HEADER= LINETYPE=[Swiss(10)]
2  PLOT TÄHDET / TYPE=FACES   MASK=-AAAA LABEL=havt DEVICE=PS,K.PS
3
4  VARIABLES: xmin      xmax     Features                        fmin fmax
5   -        *         **       Radius_to_corner_of_face_OP     0.6  1.0
6   -        *         **       Angle_of_OP_to_horizontal       0.0  0.6
7   -        *         **       Vertical_size_of_face_OU        0.6  1.0
8   -        *         **       Eccentricity_of_upper_face      0.5  1.5
9   -        *         **       Eccentricity_of_lower_face      0.5  1.5
10  NEGAT    -1.89*    2.37**   Length_of_nose                  0.1  0.5
11  -        *         **       Vertical_position_of_mouth      0.2  0.8
12  TUNTO    -2.06*    0.91**   Curvature_of_mouth_1/R         -4.0  4.0
13  -        *         **       Width_of_mouth                  0.2  1.0
14  -        *         **       Vertical_position_of_eyes       0.0  0.4
15  PAINE    -2.82*    1.05**   Separation_of_eyes              0.3  0.8
16  PANOS    -1*       1.8**    Slant_of_eyes                  -0.5  0.5
17  PANOS    -1*       1.8**    Eccentricity_of_eyes            0.3  1.0
18  -        *         **       Size_of_eyes                    0.1  0.2
19  -        *         **       Position_of_pupils             -0.1  0.1
20  -        *         **       Vertical_position_of_eyebrows   0.2  0.4
21  PAINE    -2.82*    1.05**   Slant_of_eyebrows              -0.5  0.5
22  PAINE    -2.82*    1.05**   Size_of_eyebrows                0.1  0.5
23 END
```

Riveillä 4–23 kytketään muuttujat eri kasvonpiirteisiin kirjoittamalla muuttujan nimi VARIABLES-sarakkeeseen halutun piirteen kohdalle. Valmis pohja saadaan aktivoimalla pelkkä rivin 2 PLOT-komento täsmennyksellä TYPE=FACES.

Seuraava aktivointi synnyttää kuvan, jossa kaikki naamat ovat peruslukemilla joka piirteen suhteen. Kun muuttujan nimi kirjoitetaan jollekin riveistä 5–22 ja aktivoidaan PLOT jälleen, alkaa kyseinen piirre varioida kuvassa muuttujan arvojen suhteen. Tästä merkkinä kaavioon ilmestyvät muuttujan minimi ja maksimi sarakkeisiin xmin ja xmax. Ne voi kääntää piirrekohtaisesti toisinpäin vaihtamalla lukujen perässä olevien tähtien paikkaa.

Sarake Features kuvaa piirteet lyhyesti Chernoffin (1973) artikkelissa esitetyn kaavakuvan mukaisesti. Piirteet selviävät ilman artikkeliakin yksinkertaisesti kokeilemalla. Oikean reunan sarakkeissa fmin ja fmax olevat luvut säätelevät piirteiden vaihtelun määrää. Niitä ei ole tässä muutettu oletusarvoistaan.

Kaaviossa A.17 tehdään medoidiryhmittely perustuen etäisyysmatriisiin, joka muodostetaan DIST-komennolla rivillä 2. Euklidisista etäisyyksistä lasketaan neliöt rivin 3 matriisikomennolla, minkä jälkeen riveillä 5–6 alustetaan tuloksia varten kaksi uutta muuttujaa. Ryhmittely tapahtuu rivin 8 DCLUSTER-komennolla, jossa viitataan aineistoon, etäisyysmatriisiin ja uusiin muuttujiin.

Kaavio A.17. Tuloste 6.3, sivu 169 (Survo).

```
1 MASK=-------------------------AAAA-------AA--
2 DIST UNRYHM,UNDIST / MEASURE=EUCLIDEAN IND=Vuosi,1 (1997)
3 MAT TRANSFORM UNDIST BY X#^2 /  neliölliset etäisyydet
4 ............................
5 VAR G4:1=MISSING TO UNRYHM   /  ryhmäkoodeja varten
6 VAR S4:4=MISSING TO UNRYHM   /  siluettiarvoja varten
7
8 DCLUSTER UNRYHM,UNDIST,CUR+2 /  GROUPS=4 VARS=G4(G),S4(S)
```

Kaaviossa A.18 piirretään erotteluavaruutta havainnollistava kaksoiskuva. Ensimmäisen osakentän avainsana *GLOBAL* saa täsmennykset voimaan myös muissa osakentissä, ellei niissä määrätä toisin. Näin kootaan yhteisiä, esimerkiksi kuvan koon ja mittasuhteiden sekä koordinaattiakseleiden ja viivatyypin määrittelyjä samaan paikkaan.

Riveillä 6–9 piirretään ryhmittäiset pisteet ja merkataan ne ryhmäkoodeilla (1, 2, 3, 4). Rivillä 11 piirretään muuttujat vektoreina rakennematriisista muodostetusta aineistosta.

Kaavio A.18. Kuva 7.5, sivu 180 (Survo).

```
1  CONTOUR=[line_width(0.8)][line_type(2)],0.95  FRAME=0    *GLOBAL*
2  XSCALE=4,10,16 YSCALE=2,8,14 GRID=XY
3  PEN=[Swiss(10)][line_type(0)] LINETYPE=[Swiss(10)][line_type(0)]
4  SIZE=1200,1200 XDIV=0,1,0 YDIV=0,1,0 HEADER= YLABEL= XLABEL=
5  ..............................................................
6  PLOT UNRYHM,D1,D2 / IND=G4,1 POINT=[Swiss(8)],G4 DEVICE=PS,K1.PS
7  PLOT UNRYHM,D1,D2 / IND=G4,2                     DEVICE=PS,K2.PS
8  PLOT UNRYHM,D1,D2 / IND=G4,3                     DEVICE=PS,K3.PS
9  PLOT UNRYHM,D1,D2 / IND=G4,4                     DEVICE=PS,K4.PS
10 ..............................................................
11 PLOT DCONTXR4 D1,D2 / CONTOUR= SCALE=-1:?_,1:?_  DEVICE=PS,K5.PS
12      POINT=[SwissB(13)],CASE  LINE=[line_width(1.4)],6 LINE2=0,0
13 EPS JOIN K,K1,K2,K3,K4,K5 / osakuvat yhdeksi PostScript-kuvaksi
```

Lähteet ja kirjallisuus

Alkula, Tapani; Pöntinen, Seppo & Ylöstalo, Pekka (1994). *Sosiaalitutkimuksen kvantitatiiviset menetelmät*. Porvoo: WSOY.

Alwin, Duane F. (2007). *Margins of Error: A Study of Reliability in Survey Measurement*. Hoboken, NJ: Wiley.

APA (2001). *Publication Manual of the American Psychological Association*. Washington, DC: American Psychological Association, 5. painos.

Bourdieu, Pierre (1984). *Distinction: A Social Critique of the Judgement of Taste*. Boston: Harvard University Press.

Chernoff, Herman (1973). The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, 68, 361–368.

Cook, R. Dennis & Weisberg, Sanford (1999). *Applied Regression Including Computing and Graphics*. New York: Wiley.

Cudeck, Robert & MacCallum, Robert C., toim. (2007). *Factor Analysis at 100: Historical Developments and Future Directions*. Mahwah, NJ: Lawrence Erlbaum.

DeMaris, Alfred (2004). *Regression with Social Data: Modeling Continuous and Limited Response Variables*. Hoboken, NJ: Wiley.

Eskola, Antti (1968). *Sosiologian tutkimusmenetelmät 2*. Porvoo: WSOY, 2. painos.

Fowler, Jr., Floyd J. (1995). *Improving Survey Questions: Design and Evaluation*. Lontoo: Sage.

Greenacre, Michael (2007). *Correspondence Analysis in Practice*. Boca Raton, FL: Chapman & Hall/CRC, 2. painos.

Greenacre, Michael & Blasius, Jörg, toim. (2006). *Multiple Correspondence Analysis and Related Methods*. Boca Raton, FL: Chapman & Hall/CRC.

Groves, Robert M.; Fowler, Jr., Floyd J.; Couper, Mick P.; Lepkowski, James M.; Singer, Eleanor & Tourangeau, Roger (2004). *Survey Methodology*. Hoboken, NJ: Wiley.

Hair, Jr., Joseph F.; Anderson, Rolph E.; Tatham, Ronald L. & Black, William C. (1998). *Multivariate Data Analysis*. Upper Saddle River, NJ: Prentice Hall, 5. painos.

Heikkilä, Tarja (2004). *Tilastollinen tutkimus*. Helsinki: Edita, 5. painos.

Hirvelä, Satu & Vehkalahti, Kimmo (1993). Tutkimus työviihtyvyydestä. http://www.helsinki.fi/~kvehkala/posteri93.pdf (20.9.2008).

Huff, Darrell (1954). *How to Lie with Statistics*. New York: Norton. Suom. "Kuinka tilastoilla valehdellaan", Otava (1974).

Jyrinki, Erkki (1977). *Kysely ja haastattelu tutkimuksessa*. Helsinki: Gaudeamus, 3. painos.

Kaufman, Leonard & Rousseeuw, Peter J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

Kuula, Arja (2006). *Tutkimusetiikka: aineistojen hankinta, käyttö ja säilytys*. Tampere: Vastapaino.

Kuusela, Vesa (2000). *Tilastografiikan perusteet*. Helsinki: Edita.

Lamport, Leslie (1994). *LaTeX: A Document Preparation System. User's Guide and Reference Manual*. Boston: Addison-Wesley, 2. painos.

Lehtonen, Risto & Pahkinen, Erkki (2004). *Practical Methods for Design and Analysis of Complex Surveys*. Chichester, UK: Wiley, 2. painos.

Little, Roderick J. A. & Rubin, Donald B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Mittelbach, Frank & Goossens, Michel (2004). *The LaTeX Companion*. Boston: Addison-Wesley, 2. painos.

Mustonen, Seppo (1980). SURVO 76 EDITOR, a new tool for interactive statistical computing, text and data management. Research Report 19, Department of Statistics, University of Helsinki.

Mustonen, Seppo (1992). *Survo, An Integrated Environment for Statistical Computing and Related Areas*. Helsinki: Survo Systems.

Mustonen, Seppo (1995). *Tilastolliset monimuuttujamenetelmät*. Helsinki: Survo Systems. http://www.survo.fi/monim/ (28.10.2007).

Mustonen, Seppo (1996). *Survo ja minä*. Helsinki: Survo Systems.

Mustonen, Seppo (2001). SURVO MM: käyttöympäristö tekstin ja numeerisen tiedon luovaan käsittelyyn. http://www.survo.fi/ (28.10.2007).

Mustonen, Seppo (2003). SURVO MM:n asennus- ja käyttöönotto-opas. http://www.survo.fi/opastus/Asennusopas.pdf (11.9.2008).

Mustonen, Seppo (2007). Survo Crossings. *CSCnews 1/2007*. http://www.csc.fi/english/csc/publications/cscnews/back_issues/cscnews1_2007 (29.9.2008).

Nummenmaa, Lauri (2004). *Käyttäytymistieteiden tilastolliset menetelmät*. Helsinki: Tammi.

Nummenmaa, Tapio; Konttinen, Raimo; Kuusinen, Jorma & Leskinen, Esko (1997). *Tutkimusaineiston analyysi*. Porvoo: WSOY.

Pahkinen, Erkki & Lehtonen, Risto (1989). *Otanta-asetelmat ja tilastollinen analyysi*. Helsinki: Gaudeamus.

Presser, Stanley; Rothgeb, Jennifer M.; Couper, Mick P.; Lessler, Judith T.; Martin, Elizabeth; Martin, Jean & Singer, Eleanor, toim. (2004). *Methods for Testing and Evaluating Survey Questionnaires*. Hoboken, NJ: Wiley.

Robbins, Naomi B. (2005). *Creating More Effective Graphs*. New York: Wiley.

Sariola, Sakari (1956). *Sosiaalitutkimuksen menetelmät*. Porvoo: WSOY.

Saris, Willem E. & Gallhofer, Irmtraud N. (2007). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Hoboken, NJ: Wiley.

Sovio, Ulla & Läärä, Esa (2002). Puuttuvan datan ongelma ja sen ratkaisukeinoja terveystutkimuksissa. *Sosiaalilääketieteellinen aikakauslehti*, 39, 312–325.

SPSS (2007). About SPSS Inc. http://www.spss.com/corpinfo/history.htm (29.10.2007).

Survo Systems (2007). Survo-ohjelmiston historiaa ja nykypäivää. http://www.survo.fi/esittely/ (28.10.2007).

Tarkkonen, Lauri (1987). On Reliability of Composite Scales: An essay on the structure of measurement and the properties of the coefficients of reliability – an unified approach. Statistical Studies 7, Helsinki: Suomen Tilastoseura.

Tarkkonen, Lauri & Vehkalahti, Kimmo (2005). Measurement errors in multivariate measurement scales. *Journal of Multivariate Analysis*, 96, 172–189.

Tilastokeskus (2005). SurveyLaboratorio. http://www.stat.fi/tup/surveylab/ (5.9.2008).

Tilastokeskus (2006). Verkkokoulu. http://www.stat.fi/tup/verkkokoulu/ (5.9.2008).

Vahervuo, Toivo (1956). *Psykometriikan metodeja II*. Porvoo: WSOY.

Vahervuo, Toivo & Ahmavaara, Yrjö (1958). *Johdatus faktorianalyysiin*. Porvoo: WSOY.

Valaste, Maria; Vehkalahti, Kimmo & Tarkkonen, Lauri (2008). Generalizability: Reliability or Validity? Teoksessa K. Shigemasu; A. Okada; T. Imaizumi & T. Hoshino, toim., *New Trends in Psychometrics*. Tokio: Universal Academic Press.

Valkonen, Tapani (1981). *Haastattelu- ja kyselyaineiston analyysi sosiaalitutkimuksessa*. Helsinki: Gaudeamus, 6. painos.

Valtari, Maarit [o.s. Leijola] (2001). Suomalaisten naisten ulkonäkötyytyväisyys. Pro gradu -tutkielma, Sosiaalipsykologian laitos, Helsingin yliopisto.

Vehkalahti, Kimmo (2000). Reliability of Measurement Scales: Tarkkonen's general method supersedes Cronbach's alpha. Statistical Research Reports 17, Helsinki: Suomen Tilastoseura.

Vehkalahti, Kimmo (2007). Survo+LaTeX kuvien käytössä ja näytössä. http://www.survo.fi/latex/kuvittelua.pdf (15.9.2008).

Vehkalahti, Kimmo & Everitt, Brian S. (2019). *Multivariate Analysis for the Behavioral Sciences*. Boca Raton, Florida: Chapman and Hall/CRC, 2. painos. https://crcpress.com (13.5.2019).

Vehkalahti, Kimmo; Puntanen, Simo & Tarkkonen, Lauri (2006). Estimation of reliability: a better alternative for Cronbach's alpha. Reports on Mathematics 430, Department of Mathematics and Statistics, University of Helsinki.

Vehkalahti, Kimmo; Puntanen, Simo & Tarkkonen, Lauri (2007). Effects of measurement errors in predictor selection of linear regression model. *Computational Statistics & Data Analysis*, 52, 1183–1195.

Vehkalahti, Kimmo; Puntanen, Simo & Tarkkonen, Lauri (2008). Implications of dimensionality on measurement reliability. Teoksessa Bernhard Schipp & Walter Krämer, toim., *Statistical Inference, Econometric Analysis and Matrix Algebra. Festschrift in Honour of Götz Trenkler*. Heidelberg: Springer.

Vilkka, Hanna (2007). *Tutki ja mittaa: määrällisen tutkimuksen perusteet*. Helsinki: Tammi.

Weiss, David J. & Davison, Mark L. (1981). Test theory and methods. *Annual Review of Psychology*, 32, 629–658.

Yhteiskuntatieteellinen tietoarkisto (2008). Menetelmäopetuksen tietovaranto. http://www.fsd.uta.fi/menetelmaopetus/ (5.9.2008).

Kuvat, esimerkit, tulosteet ja taulukot

Hakemisto