MOTV

Menetelmäopetuksen tietovaranto

Verkko-oppimisympäristön tekstiosio 21.1.2003

www.fsd.uta.fi/menetelmaopetus/

Yhteiskuntatieteellinen tietoarkisto
33014 Tampereen yliopisto
www.fsd.uta.fi

Sisällysluettelo

Johdatus menetelmäopetuksen tietovarantoon (versio 1.1)

FSD Yhteiskuntatieteellinen tietoarkisto ryhtyi työstämään vuoden 2000 syksyllä yhteiskuntakuntatieteellistä menetelmäopetusta tukevaa verkkotietovarantoa. Nyt varannon tietosisältö kattaa perustavat kvantitatiivisen tutkimuksen menetelmät. Myöhemmin varantoa laajennetaan ja siihen pyritään liittämään myös kvalitatiivisia menetelmiä ja aineistoja.

Tämä PDF-muotoinen tulostustiedosto kattaa tietovarannon artikkelit lisä- ja viitetietoineen, mutta ei SPSS-harjoituksia tai –ohjeita. Tietovarannon artikkelit on kirjoitettu nimenomaan verkkomediaa ajatellen, ja se näkyy myös tämän julkaisun taittoratkaisuissa. Lisäksi joihinkin lisäesimerkkeihin viitataan internet-linkillä, ja niihin tutustuminen siis vaatii tietovarannon WWW-version käyttöä.

Painettuja tekstejä on muutamissa kohdissa editoitu, mutta vastaavuus verkkoversioon on lähes täydellinen. Tuloste edustaa varannon artikkelikokoelmaa painohetkellä (ks. päivämäärä kansilehdeltä). Uudemmat artikkelit ovat luettavina vain verkossa.

Tietovarannon sisältö ja käytön yleisohjeet

Menetelmäopetuksen tietovaranto (MOTV) on suunnattu ensisijaisesti yliopistojen eri yhteiskuntatieteellisten oppiaineiden johdattaville kvantitatiivisten tutkimusmenetelmien kursseille. Lisäksi se soveltuu alan aihealueiden opetukseen ja oppimiseen muilla oppialoilla ja muissa oppilaitoksissa. Tietovarantoa voi käyttää Internetin kautta tai siitä voi tilata romppuversion. MOTV on hypertekstimuotoinen tutkimusmenetelmäsivusto, johon on integroitu alalla suositun SPSS -tilasto-ohjelmiston suomenkielinen oppimisympäristö ja tilastollisiin tunnuslukuihin ja menetelmiin liittyviä harjoituksia.

Menetelmäsivustoa voi hyödyntää ilman SPSS-osiota, mikä mahdollistaa muidenkin tilasto-ohjelmistojen vaivattoman käytön MOTV:n sisältöjä hyödyntävillä kursseilla. Toisaalta SPSS-oppimisympäristöä voi hyödyntää aihesivustosta riippumatta, joten se sopii kurssimateriaaliksi myös pelkästään ohjelmiston käyttöön keskittyville kursseille. Ohjelmisto-osion käyttö edellyttää, että käyttäjän tietokoneeseen on asennettu uudehko versio SPSS for Windows-ohjelmistosta (versio 9 tai uudempi). Esimerkit on lähes poikkeuksetta kuvattu versiolla SPSS 10 for Windows.

Opettajat ja opiskelijat voivat myös ladata sivuston linkeistä harjoitusaineistoja omille tietokoneilleen, ja tietovaranto tarjoaa kursseille muutakin oppimateriaalia. Tekijät toivovat, että käyttäjät säästävät oppilaitosten tulostimia ja kopiokoneita tilaamalla tietovarannon artikkelit kattavan edullisen painetun julkaisun tietoarkistosta. Yhteydenotot osoitteeseen *fsd@uta.fi*.

Tietovarannon tekijänoikeudet ovat FSD:llä sekä osiot valmistaneilla ja kehitystyössä mukana olevilla henkilöillä. Osioiden vastuuhenkilöt on nimetty sisällysluettelossa. Menetelmäosion teksteistä suurin osa on peräisin *Mikko Mattilalta* ja SPSS-osion sisällöt ovat pääosin *Eija Paason* käsialaa. Sivuston ja muiden julkaisujen teknisestä suunnittelusta ja toteuttamisesta on vastannut *Tuomas J. Alaterä* FSD:stä. *Sami Borg* tietoarkistosta on toiminut hankkeen koordinaattorina sekä joiltakin osin myös sisällöntuottajana. Kehitystyössä on ollut mukana myös tietoarkiston pääsuunnittelija *Jouni Sivonen*, joka on valmistanut tietovarannon pohja-aineistot sekä laatinut tietovarannon harjoituksia.

MOTV:a saa käyttää vapaasti erilaisiin oppilaitosten opetustarkoituksiin sillä edellytyksellä, että tietoarkistoa informoidaan käytöstä (*käyttöilmoituslomake verkossa*). Kurssien opetustilanteiden ulkopuolisesta yksittäiskäytöstä informointi ei ole pakollista, mutta on silti toivottavaa.

Tietovarannon tavoitteet

Tietoarkiston tehtävänä on edistää yhteiskuntatieteellisissä tutkimushankkeissa kertyvien aineistojen uudiskäyttöä tutkimuksessa ja opetuksessa. FSD:n ja MOTV:n yhtenä tavoitteena on madaltaa kynnyksiä olemassa olevien aineistojen käyttöön helpottamalla harjoitusaineistomateriaalin saamista opetukseen ja oppimiseen. Verkkoympäristö tukee myös aineistojen käsittelyyn välttämättömien välineiden, kuten tilasto-ohjelmistojen, käytön oppimista sekä lisäinformaation saantia käyttäjille tärkeistä menetelmistä ja aihealueista. Lisäksi on suotavaa, että MOTV:n käytön myötä FSD:n maksuttomat peruspalvelut tulisivat laajemminkin tutuiksi erityisesti yliopistollisten menetelmäkurssien opiskelijoille.

Käyttö Internetin kautta ja mahdollisuus rompun käyttöön takaavat tietovarannon valtakunnallisen käytettävyyden. Peruskäyttäjiksi voidaan määritellä yliopistojen yhteiskuntatieteelliset, kvantitatiiviseen tutkimukseen liittyvien menetelmäkurssien opettajat ja opiskelijat. Käytännössä käytettävyys ei kuitenkaan tunne, eikä sen tarvitse noudattaa, tiukkoja oppiaine- tai oppilaitosrajoja: muiden muassa kasvatustieteilijät tai monet ammattikorkeakoulujen opettajat ja opiskelijat pystyvät varmaankin hyödyntämään MOTV:n tietosisältöjä työssään.

Korostettakoon, että MOTV:n tavoitteena ei ole korvata alan painettuja oppikirjoja. Se ei myöskään vähennä tilastotieteen perusteiden opettamistarvetta yhteiskunta- ja käyttäytymistieteilijöille. MOTV:n näkökulma ja kehittämisen tavoitteet ovat aineisto- ja verkkopainotteisia. Pyrkimyksenä on niin ikään ollut yhteiskuntatieteilijöiden usein vaikeaksi kokeman matemaattisen esitystavan välttäminen. Ensisijaisesti on tavoiteltu kohderyhmälle tarkoituksenmukaista verkko-oppimisympäristöä.

Sivuston käyttötapa

Tietovarannon ensimmäinen versio soveltuu sekä kontaktiopetukseen että itsenäiseen opiskeluun. Kontaktiopetus toteutuu tarkoituksenmukaisimmin tietokoneluokassa joko Internet-avusteisesti tai romppuja hyödyntäen. ATK-luokka ei kuitenkaan ole ainut vaihtoehto. Tietovarantoon oheistetut kalvot ja diat ovat käyttökelpoisia dataprojektori- tai piirtoheitinavusteisissa opetustilanteissa. Luentomaisen opetuksen tukena on mahdollista käyttää myös menetelmä- eli aiheosioista painettua julkaisua.

Itseopiskelu onnistunee niinikään parhaiten ATK-luokassa, koska tietovarannon täysipainoinen hyödyntäminen edellyttää SPSS-ohjelmiston käyttöä ja ilman romppua myöskin jatkuvasti auki olevaa Internet-yhteyttä. Useimmat yliopistot tarjoavat opiskelijoilleen myös melko edullisia lisenssejä SPSS-ohjelmiston asentamiseksi omaan tietokoneeseen, joten kotiopiskelukin on realistinen mahdollisuus. Tietovarannon SPSS-osio sisältää omat yksityiskohtaiset opiskeluohjeet sekä tarjoaa muita vinkkejä opiskeluun ja harjoitusten tekoon.

Tietovarannon hyödyllisyys opettajille ja opiskelijoille riippuu paljolti opeteltavasta asiasta sekä oppimistavoitteiden ja -välineiden laadusta. Tietovarannon ensimmäinen versio soveltuu ensisijaisesti johdattaville kursseille, perusasioiden opetteluun ja kertaamiseen. Tältä pohjalta voidaan nimetä ainakin seuraavia tavanomaisia sivuston käyttötapoja ja -tilanteita:

- koko tietovaranto pohja- tai oheismateriaalina yhteiskunta- ja käyttäytymistieteilijöille suunnatuilla johdattavilla menetelmäkursseilla
- koko tietovaranto pohja- tai oheismateriaalina SPSS-ohjelmiston käyttöön harjaannuttavilla kursseilla
- tietovaranto itseopiskelumateriaalina yksittäisten aihealueiden harjoitteluun ja oppimiseen

Koska eri oppialojen ja eritasoisten kurssien opetustavoitteet poikkeavat toisistaan, tietovarantoon ei toistaiseksi sisälly suoraa mahdollisuutta tietyn tasoisten kurssien suorittamiseen. MOTV tulee kehittymään jatkossa entistä kiinteämmin virtuaaliyliopiston tavoitteita tukevaksi tietovarannoksi, mikä tarkoittaa myös erilaisten tutkintovaatimukset

täyttävien tenttimismahdollisuuksien liittämistä osaksi tietovarantoa. Tällä hetkellä tätä tarkoitusta palvelevat tai ainakin sivuavat eri menetelmiin liittyvät harjoitukset, joita opettajat voivat hyödyntää kurssien suorittamisessa.

Tietovaranto tarjoaa myös hyvän tuen eri alojen seminaari- ja opinnäytetöiden tekoon sekä uusien tutkimusaineistojen keruun suunnitteluun. Linkit Yhteiskuntatieteellisen tietoarkiston aineistoihin muistuttavat käyttäjiä siitä, että FSD tarjoaa maksutta tieteellisen tutkimuksen ja opetuksen käyttöön satoja kansallisesti ja kansainvälisesti merkittäviä data-aineistoja, jotka ovat tutkimuksellisesti vielä suurelta osin hyödyntämättä.

Mainittakoon lisäksi, että tietyt tietovarannon osat, kuten aiheosion monimuuttujamenetelmät, soveltuvat jo vähän pidemmällekin ehtineiden opiskelijoiden menetelmäopintojen oppimateriaaliksi. Yleensä ottaen MOTV:n sisällöllisten linjausten taustaajatuksena on ollut tuottaa helposti omaksuttavaa materiaalia tutkimusmenetelmien perusasioista niille, jotka opettelevat asiaa vailla aiempia tietoja tai kertaavat jo aiemmin opittuja tietoja. Pidemmälle ehtineille Internet ja alan kirjallisuus tarjoavat puolestaan runsaasti ainakin englanninkielistä oppimateriaalia kehittyneemmistä tutkimusmenetelmistä. Jatkossa tätä materiaalia tullaan valikoivasti linkittämään MOTV:n sivuille nykyistä enemmän.

Tietovarannon harjoitusaineistot

MOTV:n ensimmäisen version menetelmäosan ja SPSS-osion esimerkit ja harjoitukset perustuvat lähinnä kolmeen harjoitusaineistoon. Ensimmäinen esimerkkiaineisto pohjautuu vuonna 1995 kerätyn kansainvälisen World Values Survey 1996 -aineiston Suomea koskevaan kyselyaineistoon. Käsittelyn helpottamiseksi tähän esimerkkiaineistoon on valittu vain osa data-aineiston muuttujista. Toinen pohjautuu Tilastokeskuksen 'Maailma numeroina' -tietopalvelun tietoihin vuodelta 2000. Uusimpana joukkoon on lisätty International Social Survey Programme 2000 (ISSP) osa-aineisto. Näiden lisäksi varannossa on kymmenkunta muuta harjoitusaineistoa.

Kaikki aineistot ovat tallennettavissa suoraan Internetin kautta omalle tietokoneelle seuraamalla ohjeita tietovarannon etusivulta.

Tietoarkistosta tilattavat tietovarantoromput sisältävät automaattisesti mainitut harjoitusaineistot, ja tietoarkisto voi tarvittaessa toimittaa normaalein aineistotilausmenettelyin menetelmäkursseille myös muita aineistoja. MOTV:n kehittämisen alkuvaiheista saakka on lisäksi suunniteltu erillisen tutkimusaineiston kokoamista varta vasten opetustarkoitukseen. Hankkeen toteutuessa tästä aineistosta tulee muodostumaan tietovarannon keskeinen esimerkki- ja harjoitusaineisto.

Lähteet, palaute ja linkit muihin vastaaviin kokoelmiin

Tietovarannon menetelmäjaksojen loppuun on yleensä liitetty joitakin kirjallisuusvinkkejä. Mainitut teokset ovat myös tietovarannon sisältöjen pohjamateriaalia. Tietovarannon kehittäjät toivovat, että suositeltavan kirjallisuuden listat voisivat pidentyä käytöstä saadun palautteen myötä. Opettajien ja opiskelijoiden toivotaan ilmoittavan tietovarantoon täydennyksiä, muutosehdotuksia tai muita kommentteja mahdollisimman aktiivisesti, sillä hanke on vielä kehittämisvaiheessa.

Lähetä palautetta ja kysymyksiä verkon kautta palautekanavalla.

Useiden menetelmäosion tekstien lopuista löytyy linkkejä aihetta koskeviin verkkosivuihin. Painettuna samat linkit ja kirjallisuusviitteet löytyvät tämän julkaisun lopusta.

MOTV:n käytön muistilista

- soveltuu kontaktiopetukseen ja itseopiskeluun
- soveltuu käytettäväksi erilaisissa opetustiloissa, eri oppiaineissa ja erityyppisissä oppilaitoksissa
- opettajat tekevät käyttöilmoituksen tietoarkistolle sähköpostitse
- muusta kuin kursseihin liittyvästä käytöstä ei tarvitse tehdä ilmoitusta
- koostuu kahdesta osasta: menetelmä- eli aiheosiosta sekä SPSS-osiosta
- osioita voi hyödyntää toisistaan riippumatta
- tietovaranto sisältää runsaasti esimerkkejä ja harjoituksia sekä muuta oppimateriaalia
- SPSS-osion käyttö edellyttää, että ohjelmiston riittävän tuore versio on asennettu käyttäjän mikrotietokoneelle
- Internet-yhteys ei ole välttämätön käytön edellytys -- tietovaranto on saatavana myös rompulla ja menetelmäosion tekstit sisältävänä painettuna julkaisuna.

MOTV:n yhteystiedot

FSD Yhteiskuntatieteellinen tietoarkisto
33014 TAMPEREEN YLIOPISTO
puh. 03-215 8519
fax 03-215 8520
Email: fsd@uta.fi
URL: http://www.fsd.uta.fi/

MOTV-hanke:
Suunnitelija Tuomas J. Alaterä, tuomas.alatera@uta.fi, 03-215 8533
Johtaja (FSD) Sami Borg, sami.borg@uta.fi
Kirjoittajat:
Mikko Mattila, Eija Paaso ja Jouni Sivonen yhteys FSD:n kautta.

Hankkeen ohjausryhmä:
Ari Haukkala (HY, sosiaalipsykologia)
Vilma Hänninen (TaY, sosiologia)
Juha Kääriäinen (TaY, sosiaalipolitiikka)
Jukka Mäkelä (LaY, menetelmätieteet)

Tutkimusprosessi

Yhteiskuntatieteissä tehdään usein ero **kuvailevan** ja **selittävän** analyysin välille. Kuvaileva analyysi vastaa muun muassa kysymyksiin 'mitä', 'minkälainen' tai 'kuinka paljon'. Tällöin on tarkoituksena kuvata minkälainen tutkimuksen kohteen ilmiö on tai kuinka yleisestä ilmiöstä on kyse. Esimerkiksi tutkija voi analysoida, mitä 'työttömyydellä' oikeastaan tarkoitetaan eri yhteyksissä ja kuinka paljon tietynlaista työttömyyttä esiintyy eri ihmisryhmissä tai maantieteellisillä alueilla.

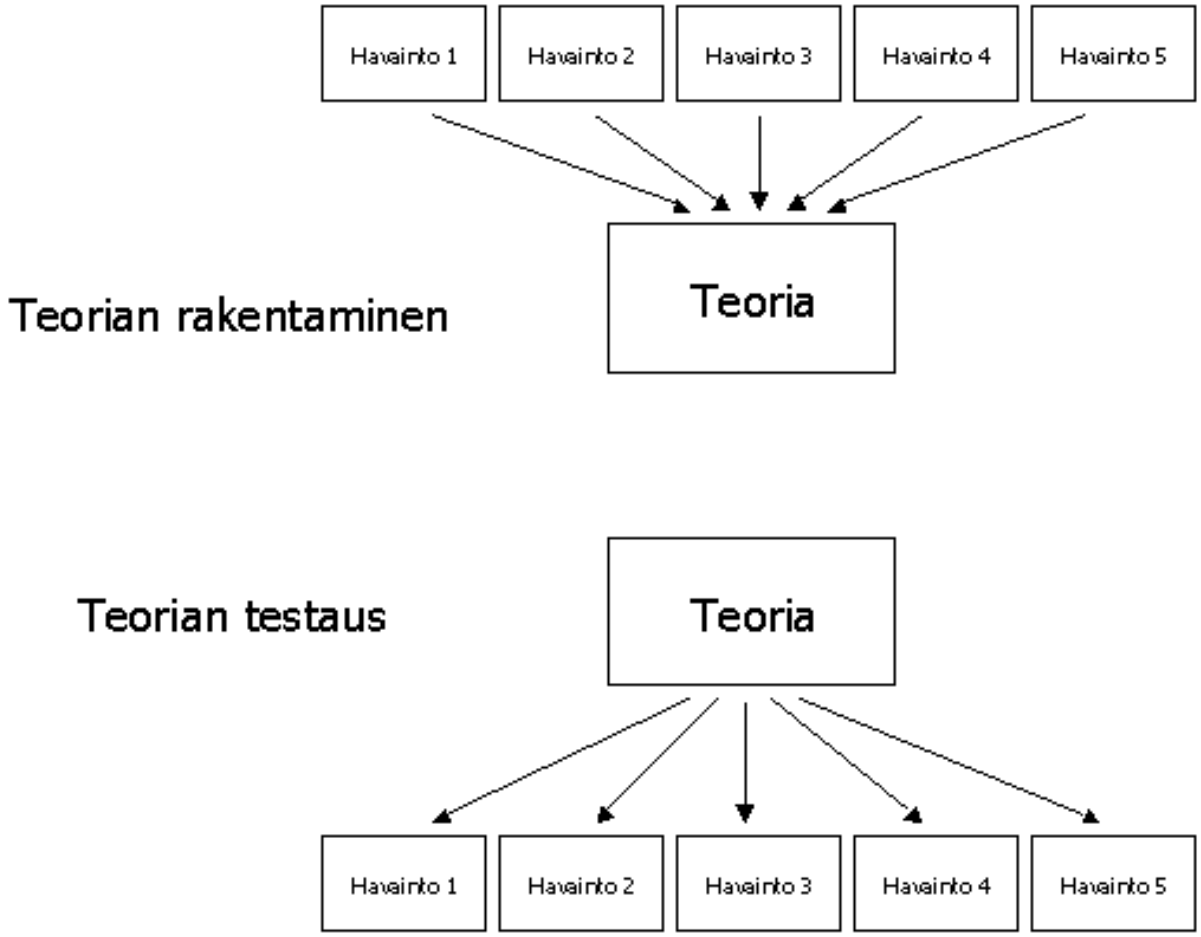
Selittävä analyysi pyrkii vastaamaan miksi -kysymyksiin. Tällöin tutkimusongelmana voi olla kysymys työttömyyden syistä. Tutkija voi olla kiinnostunut siitä, mitkä tekijät johtavat joidenkin ihmisryhmien osalta työttömyyteen tai hän voi olla kiinnostunut siitä, mikä vaikutus eri maiden verotuskäytännöillä on työttömyyden laajuuteen. Tällaisessa selittävässä analyysissa teorian osuus on keskeinen.

Teoria on systemaattinen käsitteellisen tason kuvaus ilmiöiden välisistä riippuvuussuhteista. Otetaan esimerkiksi väite 'korkea työttömyysturvan taso aiheuttaa työttömyyttä, koska turva vähentää työttömien kannustimia ottaa vastaan uusi työpaikka'. Tämä väitelause esittää kausaalisen yhteyden kahden ilmiön välillä ('työttömyysturva' ja 'työttömyys'). Ennen kuin väitteen todenperäisyyttä voidaan tutkia täytyy ilmiöt 'työttömyys' ja 'työttömyysturvan taso' jotenkin määritellä, operationalisoida ja mitata. Tämän jälkeen voidaan empiirisesti tutkia onko työttömyysturvan tasolla ja työttömyyden laajuudella yhteyttä toisiinsa.

Yhteiskuntatieteelliset teoriat pyrkivät selittämään ilmiöiden välisiä suhteita yleisellä tasolla. Etenkin määrällisessä tutkimuksessa on tarkoituksena löytää yhdenmukaisuuksia, jotka voidaan yleistää johonkin tutkimuskohteiden ryhmään (esimerkiksi suomalaiset, OECD-valtiot, jonkin ammattiryhmät edustajat, nuoret, television katsojat jne.). Selitysten yleisyystaso riippuu teoriasta ja tutkimusongelmasta. On kuitenkin tärkeää huomata, että pyrkimyksenä ei ole kuvata yksittäistapauksia, vaan nimenomaan tutkimuskohteista löydettäviä yhdenmukaisuuksia.

Teorian rakentaminen ja teorian testaus

Edellisen esimerkin tarkoituksena oli erottaa teorian ja empirian tasot. Selitysten tulisi perustua hyvin määriteltyihin teoriaan liittyviin käsitteisiin, mutta selitysten pätevyyden tutkimiseksi on tutkijan siirryttävä empiiristen havaintojen tasolle. Näin tutkimusprosessi onkin käytännössä useimmiten empirian ja teorian vuoropuhelua. Kuviossa 1 on tehty analyyttinen ero kahden erillisen prosessin, teorian rakentamisen ja teorian testaamisen, välille.



Kuvio 1. Teorian rakentaminen ja teorian testaus (De Vaus 1994, 12).

Yhteiskuntatieteellisten teorioiden rakentamiselle ei ole tarkoin määriteltyjä sääntöjä. Yhtenä lähtökohtana uuden teorian muodostamiselle voivat olla havaitut ilmiöt, joita olemassa olevat teoriat eivät pysty selittämään. Usein teorian rakentaminen alkaakin havainnoista. Ne voivat olla enemmän tai vähemmän systemaattisesti kerättyjä mielenkiinnon kohteen empiirisiä kuvauksia. Työttömyyden tutkija voi etsiä selityksiä työttömyyden laajuudelle käymällä läpi eri maiden työttömyystilastoja yrittäen samalla löytää yhdenmukaisuuksia korkean ja matalan työttömyyden maista. Näiden yhdenmukaisuuksien perusteella hän voi sitten sommitella alustavan teorian työttömyyteen vaikuttavista asioista.

Tällaista tieteellisen päättelyn tapaa kutsutaan **induktioksi**. Induktiiviselle päättelylle on ominaista, että tutkija tekee yleistyksiä perustuen rajalliselle empiiristen havaintojen määrälle. Työttömyystutkija voi esimerkiksi käydä läpi kaikki EU-maiden työttömyystilastot ja todeta, että kaikissa korkean työttömyyden maissa on löydettävissä jokin yhteinen piirre ja että matalan työttömyyden maissa tätä piirrettä ei ole. Tämän jälkeen hän tekee induktiivisen yleistyksen, että kaikissa korkean työttömyyden maissa ilmenee sama piirre. Induktiivisen päättelyn ongelmana on, että johdettu yleistys ei välttämättä päde. Ehkä EU-maiden ulkopuolelta löytyy maita, joissa ko. työttömyyden piirre esiintyy, mutta niissä työttömyys onkin matala.

Yksi seuraus induktion ongelmasta on, että teoriat, jotka väittävät jotain empiirisestä todellisuudesta on alistettava testattavaksi, ennen kuin ne voidaan hyväksyä tieteelliseksi teorioiksi. Teorian testauksen periaate on tiivistetty kuvion 1 alalaidassa. Testauksen yleinen ajatus on, että teoriasta voidaan johtaa hypoteeseja, joiden totuudellisuutta voidaan arvioida empiiristen havaintojen avulla. Tällaista päättelyä kutsutaan deduktioksi. Hypoteesi on teoriasta

loogisesti johdettu väitelause, jonka avulla voidaan epäsuorasti tutkia teorian pätevyyttä. Ajatuksena on, että jos teoriasta voidaan johtaa hypoteesi, joka osoittautuu empiiristen havaintojen valossa epätodeksi, on myös teoria epätosi. Tämä johtopäätös perustuu siihen, että tosista väitteistä ei voida johtaa loogisesti epätosia väitteitä. Käytännössä hypoteeseja tarvitaan teoriatason ja empiriatason yhtymäkohdaksi, koska teoriatasolla tarkastellaan käsitteellisten ilmiöiden yhteyksiä ilman suoraa viittausta empiriaan. Tällaista tieteellisen tutkimuksen menetelmää on kutsuttu **hypoteettis-deduktiiviseksi menetelmäksi**.

Probabilistinen selittäminen

Erityisen yleistä määrällisen yhteiskuntatieteellisen tutkimuksen alueella on edellä mainittujen induktiivisen ja deduktiivisen päättelyn lisäksi niin sanottu **probabilistinen (tai tilastollinen) selittäminen**. Probabilista selitystä tarvitaan etenkin silloin, kun teorioiden pätevyyttä tarkastellaan empiirisen aineiston valossa.

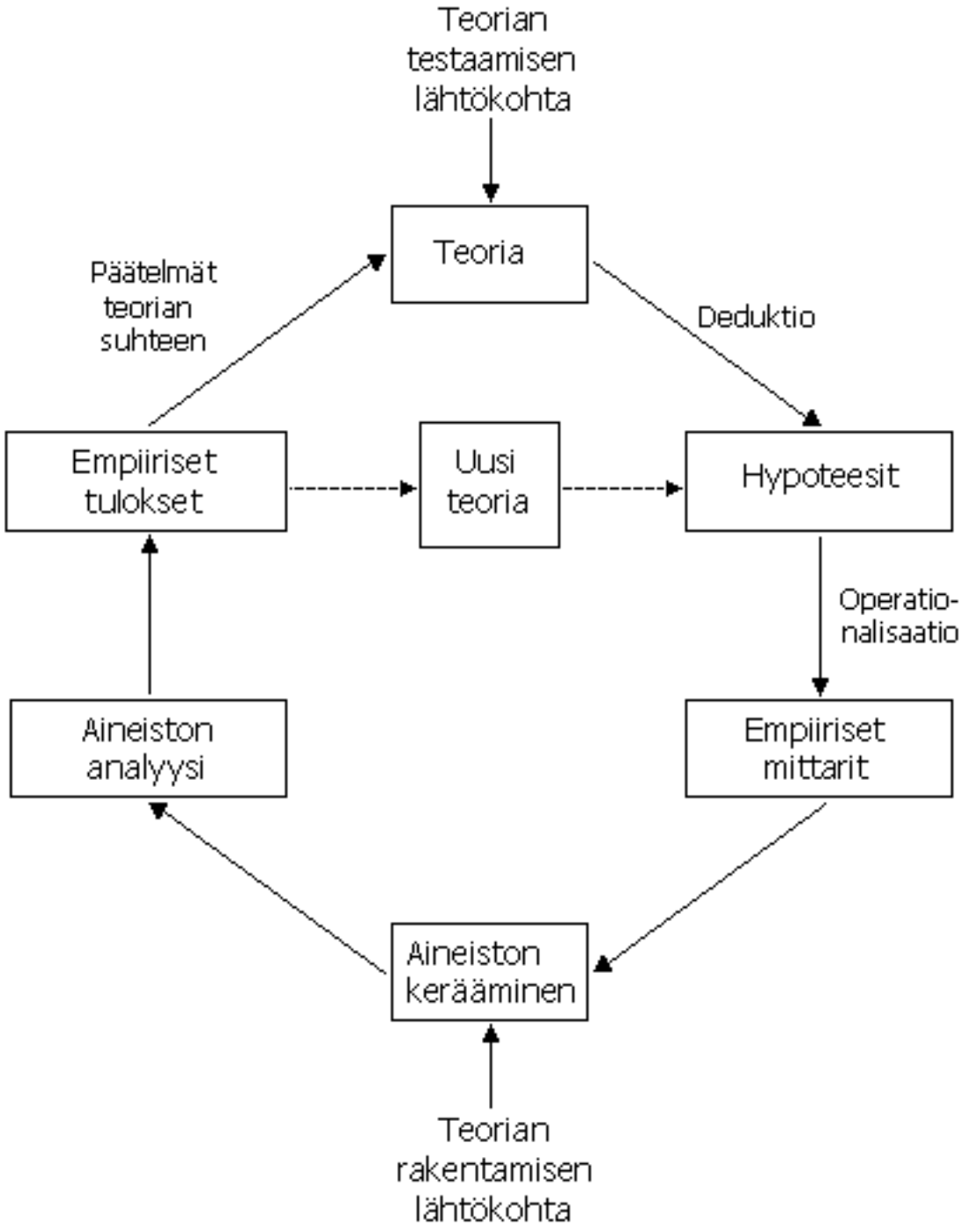
Probabilistinen selitys perustuu todennäköisyyden käsitteelle ja käyttää hyväkseen tilastollisen päättelyn menetelmiä. Probabilistisessa päättelyssä johtopäätös ei seuraa loogisesti perusteista (premisseistä) kuten deduktiivisessa päättelyssä, vaan se on ainoastaan mahdollinen jollain tietyllä todennäköisyydellä. Jos käytetään jo edellä mainittua esimerkkiä hyväksi, työttömyyden tutkija voi saada tulokseksi, että "lähes kaikissa korkean työttömyysturvan maissa on myös korkea työttömyys". Saman asian voi sanoa myös toisin: "jos maassa A on korkea työttömyysturvan taso, on hyvin todennäköistä, että maassa A on myös korkea työttömyyden taso". Selitys on probabilistinen, koska se ei ehdottomasti väitä, että maassa A olisi korkea työttömyys, mutta että todennäköisyys sille on suuri.

Edellä esitetty induktion, deduktion, probabilistisen päättelyn ja hypoteettis-deduktiivisen menetelmän esittely oli hyvin yksinkertaistettu. Tieteenfilosofian piirissä induktiiviseen ja deduktiiviseen päättelyyn liittyviä ongelmia on pohdittu laajasti. Hyvä suomenkielinen lähdeteos edellä esitettyyn keskusteluun on Ilkka Niiniluodon (1983) teos "Tieteellinen päättely ja selittäminen".

Käytännön tutkimusprosessi

Käytännön tutkimustyössä teorian rakennus- ja testausprosessit lomittuvat usein toisiinsa. Tutkimusprosessi on esitetty yksinkertaistettuna kaaviona kuviossa 2. Kuvio esittää edellä mainittujen teorian rakentamisen ja teorian testaamisen lähtökohdat. Teorian testaaja ottaa lähtökohdakseen jo olemassa olevan teorian ja muotoilee sen pohjalta tarpeelliseksi katsomansa hypoteesit. Teoria voi olla peräisin alan tutkimuskirjallisuudesta tai se voi olla tutkijan itse muotoilema. Hypoteesien valitsemisen jälkeen tutkija määrittää käsitteilleen empiiriset mittarit (ks. mittaaminen ja operationalisointi) ja kerää mittareiden avulla tarvitsemansa empiirisen aineiston. Aineiston analyysin pohjalta hän voi tehdä päätelmiä siitä, kuinka hyvin tulokset tukevat teoriaa.

Tutkimusprosessi ei useimmiten kuitenkaan lopu tähän. Tutkija voi saada empiirisen analyysinsa pohjalta uusia ideoita siitä, miten alkuperäistä teoriaa voisi kehittää paremmaksi. Tuloksena voi olla jopa aivan uudenlainen teoria. Näkemys, jonka mukaan teoriaa testaava lähestymistapa ei voi ikinä tuottaa uutta teoreettista tietoa tutkimuskohteestaan, on liian yksioikoinen.



Kuvio 2. Tutkimusprosessi (muokattu De Vaus 1994, 21 pohjalta).

Teorian rakentajan näkökulma lähtee olemassa olevista empiirisistä havainnoista. Näitä havaintoja analysoimalla tutkija kehittää teorian, joka mahdollisesti antaa tyydyttävän selityksen tutkimuskohteesta. Tutkimusprosessi ei tässäkään vaihtoehdossa lopu tähän. Uusi teoria on jotenkin koeteltava empiriaa vasten. Tällöin päädytään taas kuvion 2 ylälaitaan, eli teorian testaamisen lähtökohtaan.

Teorian testaamisen ja teorian rakentamisen erot näyttäytyvät myös varsinaisten kvantitatiivisten menetelmien tasolla. Joskus tehdään ero **eksploratiivisten** ja **konfirmatoristen** menetelmien välillä. Esimerkiksi faktorianalyysissa tehdään tällainen erottelu. Eksploratiivinen faktorianalyysi perustuu aineistolähtöiselle lähestymistavalle. Siinä katsotaan millaisia piileviä (latentteja) ulottuvuuksia aineistosta löytyy rajaamatta etukäteen mitenkään niiden määrää tai

luonnetta. Konfirmatorisessa faktorianalyysissa tutkijalla on jo ennen aineiston analyysia teoreettinen käsitys siitä, millaisen tai millaisia ulottuvuuksia hän olettaa aineistoa löytyvän. Tämän jälkeen konfirmatorinen faktorianalyysi tuottaa tilastollisia tunnuslukuja, joiden pohjalta tutkija voi päättää saivatko hänen odotuksensa tukea aineistosta vai ei. Eksploratiivinen faktorianalyysi ei näitä tunnuslukuja voi tuottaa, koska tarkoituksena on etsiä aineiston sisältämiä ulottuvuuksia ja perustaa tulkinta näihin tuloksiin.

Regressioanalyysi on myös perusluonteeltaan konfirmatorinen. Siinä tutkija päättää etukäteen, mitkä muuttujat analyysiin sisällytetään. Tuloksena on joukko tilastollisia tunnuslukuja, joiden perusteella arvioidaan muuttujien selitysvoimaa. Tosin myös regressioanalyysin voi tehdä niin, että lisätään malliin kaikki saatavilla olevat mahdollisesti asiaan vaikuttavat muuttujat ja katsotaan, mitkä niistä sitten sattuu olemaan tilastollisesti merkitseviä. Tällainen regressioanalyysin käyttö lähenee eksploratiivista analyysia. Aina tätä lähestymistapaa ei pidetä kovin suositeltavana.

Lähteet

- De Vaus, D.A. (1994): *Surveys in Social Research*. Third edition. UCL Press, Guildford.

Tutkimusasetelma

Käytännön tutkimusprosessi jakautuu useaan vaiheeseen. Näistä ensimmäinen tulisi olla tutkimusongelman muotoilu. Tutkimusongelman selkiydyttyä täytyy valita siihen soveltuva tutkimusasetelma. Oikean asetelman valinta on tärkeää muun muassa siksi, että tutkimuksella pyritään erittelemään ilmiöiden välisiä riipuvuussuhteita. Yhteiskuntatieteissä riippuvuussuhteet ovat usein monimutkaisia ja useiden eri tekijöiden kausaalisia vaikutuksia on vaikeaa erotella toisistaan. Jos esimerkiksi tarkoituksena on tutkia, miten muuttuja X vaikuttaa muuttujaan Y, on jotenkin pystyttävä kontrolloimaan muiden muuttujaan Y vaikuttavien tekijöiden osuus. Tätä ongelmaa helpottaa oikean tutkimusasetelman valinta.

Termillä 'tutkimusasetelma' tarkoitetaan joskus eri asioita. Yleisesti ottaen tutkimusasetelman tehtävänä on luoda tutkimusaineistolle mielekäs konteksti, jossa tulosten mahdollisimman yksikäsitteinen tulkinta on mahdollinen. Laajassa mielessä tutkimusasetelma käsitetään niin, että siihen liittyy tutkimusongelman muotoilu, muuttujien valinta, muuttujien operationalisointi, otantatekniikat ja aineiston keruutavat. Suppeammassa mielessä tutkimusasetelmalla tarkoitetaan empiirisen aineiston rakennetta. Seuraavassa tutkimusasetelma -termillä viitataan juuri suppeaan merkitykseen. Tämä tarkoittaa esimerkiksi sellaisia valintoja kuin missä tutkimuksen vaiheissa aineisto kerätään ja kerätäänkö aineisto samasta havaintoyksiköstä useita kertoja eri aikoina, useasta havaintoyksiköstä samanaikaisesti vai useasta havaintoyksiköstä useana eri aikana.

Taulukossa 1 on esitetty yksi tutkimusasetelmien luokittelu, joka kuvaa erilaisten asetelmien perusvaihtoehtoja. Ensimmäinen luokitteluperuste on se, liittyykö tutkimusasetelmaan yksi vai useampia mittauskertoja. Esimerkiksi tavanomainen kyselytutkimus perustuu yhteen mittaukseen. Jos myöhemmin halutaan selvittää, onko vastaajien mielipide kysytyn asian suhteen muuttunut, voidaan kysely toistaa.

Toinen luokittelu-ulottuvuus koskee havaintoyksikköjen määrää. Tutkimuksessa voi olla kohteena vain yksi havaintoyksikkö, jolloin kyseessä voi olla aikasarja- tai tapaustutkimus. Esimerkki aikasarja-aineistosta on rikollisuuden kehitys Suomessa eri vuosina. Havaintoyksikkönä on Suomi, josta on rikollisuuden osalta tehty vuosittaisia mittauksia. Tapaustutkimus voisi liittyä esimerkiksi yrityksen uuden tietojärjestelmän käyttöönotosta aiheutuviin ongelmiin. Jos tutkimus koskisi useita yrityksiä ja niiden ongelmia, voisi kyseessä olla poikkileikkausaineisto.

Taulukko 1. Esimerkkejä tutkimusasetelmista.

	Useita havaintoyksikköjä	**Yksi havaintoyksikkö**
Useita mittauksia	A Klassinen koeasetelma Paneeliaineisto	B Aikasarja-aineisto
Yksi mittaus	C Poikkileikkausaineisto	D Tapaustutkimus

A. Useita havaintoyksikköjä - useita mittauksia

Klassinen koeasetelma

Klassista koeasetelmaa pidetään joskus tieteellisen tutkimusasetelman ideaalimallina. Se antaakin erinomaisen mahdollisuuden kausaalisuhteiden olemassaolon ja voimakkuuden arviointiin, koska siinä pyritään eristämään mahdollisimman hyvin kaikkien muiden muuttujien vaikutus selitettävään muuttujaan. Asetelman ongelmana on, että se on etenkin yhteiskuntatieteissä usein vaikea toteuttaa.

Yksinkertaisimmassa muodossaan klassisessa koeasetelmassa havaintoyksiköt on jaettu kahteen ryhmään: testi- ja kontrolliryhmään. Ideaalitapauksessa havaintoyksiköt on jaettu näihin ryhmiin satunnaisesti, mutta joskus on tarpeen tehdä jako harkinnanvaraisesti (esimerkiksi jos tutkija haluaa varmistaa, että molemmissa ryhmissä on tarpeellinen määrä tietyn ikäisiä henkilöitä ja ryhmät ovat niin pieniä, että satunnainen ryhmäjako ei pysty tätä varmistamaan). Tutkimusprosessin aikana näistä kahdesta ryhmästä mitataan halutut asiat vähintään kahdesti eri aikoina. Lisäksi koeasetelman olennainen piirre on **interventio**, joka kohdistetaan testiryhmään, mutta ei kontrolliryhmään. Interventiolla tarkoitetaan sitä, että kiinnostukseen kohteena olevan kausaalisen muuttujan annetaan vaikuttaa testiryhmään. Esimerkiksi jos kyseessä on lääketieteellinen koe, annetaan testiryhmälle tutkimuksen kohteena olevaa lääkitystä, mutta kontrolliryhmälle vain plaseboa. Yhteiskuntatieteissä interventio voi tarkoittaa myös tutkijasta riippumatonta muutosta, jonka vaikutusta halutaan tutkia. Esimerkiksi muuttunut lainsäädäntö voi olla tällainen interventio.

Tutkimuksen ensimmäinen mittauskerta tehdään aina ennen interventiota, jotta molemmasta ryhmästä pystytään mittaamaan selitettävän muuttujan lähtötaso. Seuraava tai seuraavat mittauskerrat tehdään intervention jälkeen, minkä jälkeen saatuja tuloksia verrataan ensimmäisen mittauskerran tuloksiin. Näin saadaan selville muutoksen suuruus sekä testi- että kontrolliryhmässä. Jos muutos on merkittävästi erilainen testiryhmän osalta kuin kontrolliryhmässä, voidaan päätellä, että selittävällä muuttujalla (interventio) oli kausaalinen yhteys selitettävään muuttujaan.

Esimerkki yhteiskuntatieteellisestä tutkimuksesta, jossa hyödynnetään koeasetelmaa on Gerberin ja Greenin (2000) tutkimus, jossa analysoitiin erilaisia tapoja kannustaa ihmisiä äänestämään vaaleissa. Tutkijat valitsivat erään USAn kaupungin äänestäjäluettelosta neljä eri otosta, joista yksi toimi kontrolliryhmänä. Kolmeen muuhun ryhmään kohdistettiin kuhunkin interventio. Yhtä ryhmää kannustettiin äänestämään soittamalla heille kotiin, yhdelle ryhmälle lähetettiin postissa äänestämiseen kannustavaa materiaalia ja kolmannen ryhmän äänestämistä käytiin kannustamassa henkilökohtaisella vierailulla. Tuloksia analysoimalla tutkijat pystyivät päättelemään, miten eri tavat saada äänestäjät osallistumaan onnistuivat tavoitteissaan.

Yleisesti ottaen klassisen koeasetelman käyttö yhteiskuntatieteissä on vähäistä. Syyt tälle ovat sekä käytännöllisiä että eettisiä. Usein on vaikea keksiä, miten asetelmaa voisi soveltaa käytännön yhteyksissä. Jos tutkija haluaisi tietää, miten television katselu vaikuttaa kulutustottumuksiin, ei tutkimusta voitane tehdä niin, että satunnaisesti valittu ryhmä ihmisiä lahjottaisiin (tai pakotettaisiin!) katsomaan paljon televisiota ja kontrolliryhmän kodeista takavarikoitaisiin kaikki televisiot pois. Monissa tutkimusongelmissa koeasetelman käyttäminen olisi eettisesti kyseenalaista. Jos haluttaisiin tutkia koulukiusattuna olemisen vaikutusta oppilaiden koulumenestykseen, intervention estäisi tutkimusetiikan lisäksi myös lainsäädäntö.

Paneeliasetelma

Myös paneeliasetelmaan kuuluu useiden havaintoyksiköiden käyttö sekä ainakin kaksi eri mittauskertaa. Erona klassiseen koeasetelmaan on se, että paneeliasetelma ei edellytä kontrolliryhmän käyttöä. Samoin kuin klassisessa koeasetelmassa paneeliasetelmassa ensimmäinen mittauskerta suoritetaan ennen interventiota. Seuraava tai seuraavat mittauskerrat

tapahtuvat intervention jälkeen, minkä jälkeen tutkitaan kuinka suuri muutos interventiosta seurasi mielenkiinnon kohteena olevassa muuttujassa. Koska paneeliasetelmaan ei kuulu kontrolliryhmää, ongelmana on se, että tutkija ei voi olla aivan varma siitä johtuuko havaittu muutos juuri interventiosta vai vaikuttiko siihen jokin muu tekijä, jonka osuutta ei etukäteen osattu ottaa huomioon.

Esimerkki paneeliasetelmasta on tutkimus, jossa tarkasteltaisiin opiskelun vaikutusta opiskelijoiden uskonnollisiin mielipiteisiin (De Vaus 1994, 38). Ensimmäinen mittauskerta tehtäisiin uusien opiskelijoiden ilmoittautuessa ensimmäistä kertaa yliopistoon. Toinen mittauskerta voisi sijoittua esimerkiksi kolmen opiskeluvuoden päähän, jolloin samoilta opiskelijoilta kysyttäisiin uudelleen samat kysymykset. Tulosten erojen perusteella tutkija voisi tehdä päätelmiä siitä, miten opiskelu vaikuttaa uskonnollisiin mielipiteisiin. Ongelmana on, että mielipiteiden muutokseen voi vaikuttaa myös joku muu asia, (esimerkiksi tutkimukseen osallistuvien henkilöiden vanhentuminen, siteiden heikkeneminen kotiin jne.)

Paneeliasetelma on myös hyvin tyypillinen tutkimuksissa, joissa havaintona on joukko valtioita ja niistä on useita mittauksia eri aikoina. Esimerkiksi budjettivajeiden tutkimuksessa on tavallista, että aineistossa selitettävänä muuttujana on budjettivajeen vuosittainen suuruus joltain tietyltä aikaväliltä (esim. 1970-1995) tietyssä ryhmässä maita (esim. OECD-maat).

Kvasipaneeliasetelma

Kvasipaneeliasetelma eroaa varsinaisesta paneeliasetelmasta siinä, että mittauksen kohteena olevat ihmiset eivät ole samoja eri mittauskerroilla. Käyttäen edellistä esimerkkiä hyväksi, kvasipaneeliasetelmassa ensimmäinen mittauskerta suoritettaisiin satunnaiselle otokselle ensimmäisen vuoden opiskelijoita ja toinen mittauskerta kolmen vuoden kuluttua uudelleen satunnaiselle otokselle kolmannen vuoden opiskelijoita. Tähän asetelmaan liittyy samat ongelmat kuin paneeliasetelmaan yleisesti sekä lisäksi vielä se, että mittauskertojen väliseen eroon voi vaikuttaa myös tutkittavien ryhmien erilainen koostumus.

Retrospektiivinen paneeli ja retrospektiivinen koeasetelma

Edellä mainittujen asetelmien lisäksi voidaan vielä tarkastella retrospektiivistä paneeliasetelmaa ja retrospektiivistä koeasetelmaa. Nämä asetelmat ovat hyvin lähellä poikkileikkausaineistoja. Näille kahdelle asetelmalle on ominaista, että ne pyrkivät jäljittelemään koeasetelmaa ja paneeliasetelmaa, vaikka niissä käytetään vain yhtä mittauskertaa.

Sekä koeasetelmassa että paneeliasetelmassa on ongelmana, että tutkimuksen teko voi kestää vuosikausia, koska intervention vaikutukset voivat näkyä vasta pitkällä aikavälillä. Retrospektiivisen koeasetelman ideana on, että siinä tehdään vain yksi mittaus intervention jälkeen ja pyydetään sekä testi- että kontrolliryhmän vastaajia vastaamaan myös menneisyyttä koskeviin kysymyksiin. Näin ensimmäinen lähtötason mittaus tehdään samanaikaisesti intervention jälkeisen mittauksen kanssa. Samoin retrospektiivinen paneeliasetelma perustuu yhteen mittauskertaan, jossa kartoitetaan sekä menneisyyden että nykyisyyden tapahtumia, mutta siinä ei ole kontrolliryhmää mukana. Näiden asetelmien ilmeinen ongelma on ihmisten "valikoiva" muisti, joka voi johtaa siihen, että heidän menneisyyttä koskevat vastauksensa eivät ole samat, kuin jos ne olisi aikanaan kysytty tuoreeltaan.

Käyttäen edelleen opiskelijoiden uskonnollisia mielipiteitä esimerkkinä, retrospektiivinen paneeliasetelma olisi sellainen, missä kolmannen vuoden opiskelijoilta kysytään heidän uskonnollisia mielipiteitään juuri vastaushetkellä ja sitä, miten he muistinsa mukaan suhtautuivat samoihin asioihin opiskelunsa alkuvaiheilta.

B. Yksi havaintoyksikkö - useita mittauksia: aikasarja-aineisto

Tyypillisessä aikasarja-asetelmassa yhdestä havaintoyksiköstä on useita mittaustuloksia eri aikapisteissä. Aikasarja-aineistoja käytetään laajasti mm. kansantaloustieteen piirissä. Yksi esimerkki aikasarja-aineistosta voisi olla autovarkauksien määrä Suomessa vuosina 1960-1995. Selittävänä muuttujana analyysissä voisi olla lainsäädännössä tapahtuneet muutokset tai taloudellisen tilanteen muutokset. Politiikan tutkimuksen alalla tyypillinen aikasarja-aineisto kuvaa jonkin puolueen vaalikannatusta jollain aikavälillä. Aikasarja-aineistojen tutkimukseen käytetään yleensä regressioanalyysia, joskaan tämä ei ole ainoa soveltuva menetelmä (ks. Dale & Davies 1994).

C. Useita havaintoyksikköjä - yksi mittaus: poikkileikkausaineisto

Poikkileikkausasetelma on ehkä kaikkein yleisin määrällisen yhteiskuntatieteen asetelma. Se koostuu yhdestä ainoasta mittauskerrasta, joka kohdistetaan useaan havaintoyksikköön. Havaintoyksiköt voivat olla esimerkiksi ihmisiä (kuten useimmiten kyselytutkimuksissa on) tai kuntia, joista selvitetään halutut muuttujien arvot.

Poikkileikkausaineistoihin ei sisälly useiden mittauskertojen sallimia mahdollisuuksia muutostarkasteluihin ajan suhteen, joten kausaalisuhteiden tunnistamiseen ja mittaamiseen täytyy käyttää muita menetelmiä. Käytännössä voidaan pyrkiä tarkastelemaan kiinnostuksen kohteena olevaa kausaalisuhdetta niin, että kaikkien muiden asiaan vaikuttavien tekijöiden vaikutus on eliminoitu analyysin ulkopuolelle.

Oletetaan, että tutkija haluaa poikkileikkausotoksen avulla tietää vaikuttaako sukupuoli uskonnollisten mielipiteiden määrään tai vahvuuteen suomalaisissa, täytyy hänen ottaa myös huomioon se, että luultavasti ikä vaikuttaa myös asiaan niin, että vanhemmat henkilöt ovat keskimäärin uskonnollisempia kuin nuoremmat. Koska Suomessa naiset elävät selvästi pidempään kuin miehet, on luultavaa, että otoksessa naisten keski-ikä on suurempi kuin miesten. Jos alustavat tutkimuksen tulokset osoittavat, että naiset ovat hiukan miehiä uskonnollisempia, tulos voi olla tavallaan harhaanjohtava siitä syystä, että naisvastaajat olivat keskimäärin vanhempia kuin miesvastaajat. Luotettavampien tulosten saamiseksi täytyy iän vaikutus kontrolloida ennen päätelmien tekoa. Tässä voi käyttää apuna esimerkiksi osittaiskorrelaatiokertoimia, regressioanalyysia tai ristiintaulukointia ikäryhmittäin (elaboraatio).

D. Yksi havaintoyksikkö - yksi mittaus: tapaustutkimus

Tapaustutkimuksessa keskitytään jonkun tietyn ainutkertaisen tapahtuman tutkimiseen. Tapaustutkimukselle on ominaista, että tutkimuksen kohdetta tarkastellaan sen luonnollisessa ympäristössä ja että aineistona on useita erilaisia lähteitä. Tarkoituksena on luoda mahdollisimman kattava kuvaus tapahtuman ymmärtämiseksi. Tapaustutkimuksen yleinen määritteleminen on vaikeaa ja joskus rajanveto muihin asetelmiin on ongelmallista (esimerkiksi Yin (1990) puhuu usean tapauksen tapaustutkimuksesta).

Lähteet

- Dale, Angela & Davies, Richard B. (1994): *Analyzing Social & Political Change. A Casebook of Methods*. Sage, Lontoo.
- Gerber, Alan S. & Green, Donald P. (2000): *The Effects of Personal Canvassing, Phone Calls, and Direct Mail on Voter Turnout: A Field Experiment*. American Political Science Review 94: 653-663.
- De Vaus, D.A. (1994): *Surveys in Social Research*. Third edition. UCL Press, Guildford.
- Yin, Robert K. (1990): *Case Study Research. Design and Methods*. Revised Edition. Sage, Newbury Park.

Mittaaminen

Mittaamiseen voidaan käyttää hyvin monenlaisia apuvälineitä: joku mittaa katseellaan, toinen kirjevaa'alla. Perinteisesti mittaaminen ymmärretään määrälliseksi eli siihen liitetään suureet paino, pituus, matka yms. Mittauksen kohteet ovat hyvin erilaisia eri tieteenaloilla: fyysikko tutkii elektronien ominaisuuksia, yhteiskuntatieteilijä ihmisten. Mitattavana voivat olla spektrien aallonpituudet tai etniseen ryhmään kuuluminen. Yksittäisessä tutkimuksessakin tarvitaan useita erilaisia mittareita.

Tieteellisesti pätevällä mittarilla on tietyt vaatimukset. Mittarin määrittäminen lähtee siitä, että ensin määritellään asia tai ilmiö, jota halutaan mitata. Tämä edellyttää ilmiön täsmällistä käsitteellistämistä. Sitten on kyettävä määrittämään konkreettinen mittari eli tutkittava ilmiö on operationalisoitava. Mittari voidaan kehittää itse, mutta usein voidaan käyttää valmiita mittareita. Operationalisoinnin tuloksena syntyy siis mittareita.

Mittaria konstruoitaessa on tärkeää huomioida tutkimuksen kohderyhmä. Esimerkiksi lapsille voidaan harvoin käyttää samanlaista mittaria kuin aikuisille. Valmista mittaria käytettäessä on erityisen tärkeää selvittää, mitä se tarkkaan ottaen mittaa, ja mikä on ollut alkuperäinen kohderyhmä (kulttuuri, kieli, ajankohta, vastaajien ikä yms.)

Mittarin täytyy olla kohteeseensa sopiva: kirjevaa'alla ei voida mitata henkilöiden painoja. Kunnan varallisuuden ja vanhuksista huolehtimisen arviointiin ei voida käyttää samaa mittaria. Myös esimerkiksi tämän hetken työssä jaksamisesta saadaan osuvampi kuva ajanmukaisella mittarilla kuin 20 vuotta sitten konstruoidulla operationalisoinnilla. Valmista mittaria käytettäessä on ehdottomasti tiedettävä, miten mitattava asia on määritelty; sopiiko mittari omiin tutkimustavoitteisiin. Jos vastaaja ei ymmärrä kysymystä, syntyy mittausvirheitä. Mittarin on mitattava sitä asiaa, mitä sillä halutaan mitata. Tällöin puhutaan mittarin validiteetista.

Mittarin on oltava myös luotettava mittauksia toistettaessa eli sillä on oltava pysyvyyttä. Mittarin eduksi luetaan se, että se mittaa kokonaisuudessaan samaa asiaa ts. mittarin konsistessi on hyvä. Nämä ominaisuudet liittyvät mittarin reliabiliteettiin.

Mittaamista suunniteltaessa on hyvä esittää ainakin kysymykset 'mitä', 'mistä' (kohderyhmä), 'millä', 'miten mitataan'. Mittarin tarkkuudesta on hyvä pitää mielessä, että mittaustulos on ikään kuin maisemataulu. Se kertoo jollakin tavalla rajatut piirteet todellisesta maisemasta. Tavoitteena on mahdollisimman realistinen kuva.

Katso verkosta mittaamiseen liittyvistä lisäesimerkeistä kohdat 1-3
(*http://www.fsd.uta.fi/menetelmaopetus/mittaaminen/esimerkit.html*)

Esimerkkitapaus: Kielenkäyttöindeksi kielenvaihtoprosessin kuvaajana

Tiedonkeruuvaiheessa käytettyjä mittareita on joskus hyödyllistä yhdistää toimivampaan muotoon aineiston analysointia varten. Tällaisesta on hyvä esimerkki Marjut Aikion (1988) kehittämä kielenkäyttöindeksi, jonka avulla saadaan kuvattua pitkän aikavälin kielenvaihtoprosessia. Aikio on tutkinut väitöstyössään "Saamelaiset kielenvaihdon kierteessä" saamen kielen vaihtumista suomenkieleksi viidessä saamelaiskylässä. Haastatelluista vanhin oli syntynyt vuonna 1898 ja nuorin 1954 ja heiltä on kysytty kielenkäyttöä vuosikymmenittäin.

Kielenkäytössä erotettiin neljä eri vaihtoehtoryhmää kuvaamaan henkilön puheessa käyttämää kieltä:

1.00 lp (lappi) = vain saamea
0.75 lp/sm = saamea ja suomea, mutta saamen osuus on selvästi suurempi
0.25 sm/lp = saamea ja suomea, mutta suomen osuus on selvästi suurempi
0.00 sm = vain suomea

Vaihtoehdoista käytettiin oheisia lukuarvoja. Lukuarvot on valittu siten, että ne korostavat kummankin pään siirtymävaihetta, jonka aikana hallitseva kieli vaihtuu toiseksi.

Kieli-indeksi laskettiin kullekin haastatellulle kutakin tarkasteltavaa vuosikymmentä kohti. Se perustuu haastatellun erilaatuisten kielikontaktien lukumäärään ja ilmaisee henkilön keskimääräisen saamen kielen käytön eri aikoina. Eri kielikontaktien (lp, lp/sm, sm/lp, sm) arvioidut esiintymiskerrat laskettiin kunakin vuosikymmenenä eli monenko kanssa haastateltu oli arvioinut puhuneensa esim. enimmäkseen saamea 1950-luvulla. Kieli-indeksi on siis painotettu keskiarvo eri vaihtoehdoista (lp, lp/sm, sm/lp,sm), joista kukin vaihtoehto painotetaan kyseisillä kielikontaktien lukumäärällä.

Kieli-indeksin kaava voidaan kirjoittaa seuraavasti:

$KKI = [a*lp + b*(lp/sm) + c(sm/lp) + d*sm] / (a+b+c+d),$

missä kertoimet a, b, c ja d ovat kyseisiä puhekontakteja vastaavien henkilöiden lukumäärä tarkasteltavana vuosikymmenenä.

Mittaaminen: Tilastoyksikkö ja muuttujat

Yhteiskuntatieteissä tiedonkeruun kohteena ovat usein yksittäiset ihmiset. Tällöin mittauksen kohteena olevat henkilöt ovat kvantitatiivisen tutkimuksen käsitteillä ilmaistuna **tilastoyksikköjä** tai **havaintoyksikköjä**. Tilastoyksikkö voi myös olla jokin muu konkreettinen tai abstrakti kohde.

Ihmisillä on tilastoyksikköinä erilaisia tutkimuksellisia rooleja: asiakas, osallistuja, matkustaja, äänestäjä, kuluttaja, työtön, vanhus, potilas, tiettyyn uskonnolliseen yhteisöön kuuluva, opiskelija, asiantuntija, ruotsinsuomalainen tai saamelainen. Tilastoyksikkö voi olla myös organisaatio tai yhteisö: valtio, kansa, perhe, koulu, yritys, yliopisto, sairaala. Sanomalehti tai siinä oleva uutinen voi olla perusyksikkö, josta kerätään tietoa. Toisaalta tilastoyksikkö voi olla abstrakti, kuten asiakassuhde. Tällöin konkreettinen tieto voidaan kerätä useista lähteistä: asiakassuhteessa olevilta henkilöiltä, tilannehavainnointina, rekistereistä ja asiapapereista. Yksittäiset asiakirjat, esimerkiksi perukirja tai mielentilalausunto, voivat myös olla tutkimuksen kohteina tai tutkimuskohteen edustajina.

Tarkoituksenmukaisen tilastoyksikön valitseminen ei ole aina yksinkertaista. Ongelma, joka liittyy siihen, onko tilastoyksikkö mielentilatutkittava vai häntä koskeva mielentilalausunto, ei liene kovin vakava. Sen sijaan tilastoyksikön tarkempi, sisällöllinen määritteleminen voi olla vaikeaa, kun tilastoyksikkönä on esimerkiksi perhe. Tutkimuksessa on lähdettävä perhekäsitteen määrittelemisestä kyseisessä tutkimuksessa ja kenties mietittävä muita mahdollisia tilastoyksikkövaihtoehtoja; voisiko yksittäinen perheenjäsen edustaa perhettä tutkimuksen tilastoyksikkönä? Joskus tutkimusta palvelevien johtopäätösten saamiseksi voidaan yhdessä tutkimuksessa tarkastella rinnakkain erilaisia tilastoyksikkövaihtoehtoja, esimerkiksi perheenjäsen ja perhe.

Tutkimuksessa **populaatio** eli **perusjoukko** on kohdejoukko, josta tutkimuksessa halutaan tehdä päätelmiä. Joskus on mahdollista tehdä kokonaistutkimus, jossa kerätään tietoja kaikista perusjoukkoon kuuluvista tilastoyksiköistä. Otantatutkimuksessa perusjoukkoa edustaa otos, josta saatuja tuloksia voidaan yleistää perusjoukkoon. Joskus on tarkoituksenmukaisempaa

kerätä ns. näyte, joka ei edusta kattavasti perusjoukkoa, mutta jonka avulla saadaan käytössä olevilla resursseilla tarkoituksenmukaisemmin tietoa tutkittavasta asiasta.

Tutkimusta varten tieto on saatettava sellaiseen muotoon, että sitä voidaan johdonmukaisesti ja jäsennellysti käsitellä. Kvantitatiivisessa tutkimuksessa tämä tapahtuu tilastollisten muuttujien avulla. Kun esimerkiksi ihmisiltä kysytään jotakin asiaa, kysymykseen annetut vastaukset eroavat. Yleisesti ottaen tilastoyksiköiden tiedot eroavat toisistaan riippumatta siitä, millä tiedonkeruumenetelmällä ne on tuotettu tai minkä tyyppistä ilmiöaluetta mitataan. Tällaisesta tilastoyksiköihin liittyvästä asiasta tai ominaisuudesta voidaan luoda **tilastollinen muuttuja**. Muuttujia voivat olla esimerkiksi sukupuoli, ikä tai mielipide, kun tilastoyksikkönä on henkilö. Yrityksen ollessa tilastoyksikkönä muuttujia voivat olla liikevaihto tai henkilöstön määrä. Silloin, kun tietoa kerätään kunnista, muuttujina voivat olla myös asukasluku, kuntamuoto tai henkirikosten määrä. Termi 'tilastollinen' korostaa sitä, että mittauksessa saatu muuttujan arvo on tietyllä hetkellä tilastoitu tieto, useimmiten luku, joka tiivistää mahdollisesti hyvinkin moniulotteisen ominaisuuden.

Kyselytutkimuksessa tilastollinen muuttuja voidaan muodostaa kysymyksestä tai väittämästä. Tavallisesti vastauksista saadaan muuttujan arvoiksi numeroita. Kysymyksiin annetut vastaukset ovat **muuttujan arvoja**. Jos tutkittaisiin täysi-ikäisiä työelämässä olevia henkilöitä, ikämuuttujan **mahdolliset arvot** olisivat välillä 18 - 65 vuotta. Kun tietyn henkilön ikä on 50 vuotta, tämä voidaan ilmaista tilastotieteen kielellä seuraavasti: tilastoyksikön a_i saama muuttujan $\underline{x}$ arvo on 50 vuotta. Sukupuolimuuttujan mahdolliset arvot ovat 'nainen' ja 'mies'. Tällaiset sanalliset vaihtoehdot käsitellään tilasto-ohjelmilla yleensä numeroiksi muutettuina, jolloin kutakin vaihtoehtoa vastaa yksikäsitteinen **numeerinen koodi**. Esimerkiksi sukupuoli voidaan koodata siten, että numero 1 tarkoittaa naista ja numero 2 miestä.

Tutkimuksen näkökulmasta muuttujat voidaan jaotella mm. **taustamuuttujiin** ja varsinaisiin **tutkimusmuuttujiin**. Tutkimusmuuttujat liittyvät välittömästi tutkittavaan ilmiöön; sen sijaan taustamuuttujat antavat yleisempää tietoa tilastoyksiköstä. Käytännössä jako ei ole välttämättä täysin yksiselitteinen. Eniten käytettyjä taustamuuttujia ovat sukupuoli, siviilisääty, syntymävuosi, ikä ja koulutus. Joskus taustamuuttujiksi voidaan lukea hiukan spesifimpiäkin muuttujia. Esimerkiksi tilastokeskuksen vaalimenestystutkimuksissa (ks. esim. *http://www.stat.fi/tk/he/vaalit/vaalit99/tilastoanalyysi.html*) taustamuuttujina on käytetty alueittaisista muuttujista työllisyyttä, elinkeinorakennetta, kaupungistumisastetta, eläkeläisten osuutta ja puolueiden kannatuspohjaa (KESK - KOK - SDP). Lääkärien työoloja ja kuormittuneisuutta selvittävässä tutkimuksessa taustamuuttujiksi on nimetty tyypillisten sukupuolen ja iän lisäksi päätoimipaikka, nimike päätoimessa, työsuhteen vakinaisuus sekä erikoistuminen. Varsinaisia tutkimusmuuttujia ovat mm. työaika, työtahdin kokeminen ja työtahdin kiristymisen syyt.

Muuttujia voidaan jaotella myös sillä perusteella, kuinka välittömästi ne mittaavat tutkittavia asioita. Useat henkilöitä koskevat taustatiedot lukeutuvat muuttujiin, joilla on tarkoitus mitata vain ja ainoastaan kysymyksessä mainittua asiaa esim. sukupuoli. Kun taas kysytään "Oletko puolueen jäsen?" ei välttämättä ollakaan kiinnostuneita pelkästään puolueen jäsenyydestä, vaan kysymyksen tarkoituksena voi olla selvittää henkilön poliittista aktiivisuutta. Tällöin puolueen jäsenyys *indikoi* poliittista aktiivisuutta ja siitä voidaan käyttää nimitystä **indikaattorimuuttuja**. Myös syntymävuosi voidaan ymmärtää indikaattorimuuttujaksi. On eri asia olla kiinnostunut syntymävuodesta, johon liittyy erilaisia historiallisia tapahtumia, kuin olla kiinnostunut vain henkilön iästä. Myös kysymyspatteristot koostuvat usein indikaattorimuuttujista, joilla operationalisoidaan erilaisia käsitteitä.

Mittaaminen: Havaintomatriisi

Tilastoyksikköjä koskevat tiedot on koottava järkevään muotoon aineiston käsittelyä varten. Yleensä havainnot tallennetaan tilasto-ohjelmissa matriisiksi. Sen kullakin vaakarivillä

on yhden tilastoyksikön saamat muuttujien arvot. Tietyn muuttujan arvot puolestaan sijaitsevat samoissa sarakkeissa kaikilla tilastoyksiköillä. Kuvion 1 havaintomatriisissa on sukupuolta ja ikää koskevat tiedot seitsemältä henkilöltä. Lisäksi ensimmäisessä sarakkeessa on vastauslomakkeen numero. Tilasto-ohjelmistot käyttävät **havaintomatriisia** perusdatana, josta ne laskevat mm. tunnuslukuja ja piirtävät graafisia esityksiä.

	lomake	sukupuol	ika
1	1	1	23
2	2	1	24
3	3	2	50
4	4	1	70
5	5	2	18
6	6	2	40
7	7	2	37

Kuvio 1. Havaintomatriisi

Tilastotieteen kaavoissa tilastoyksiköitä merkitään yleensä kirjaimella a, johon liitetään alaindeksi, esim. a_3. Alaindeksi kertoo, millä rivillä kyseisen tilastoyksikön tiedot ovat. Yleisemmin merkitään a_i, i= 1, ..., 7.
Tämä tarkoittaa, että aineistossa on seitsemän havaintoyksikköä, a_1, a_2, a_3, a_4, a_5 , a_6, a_7.

Formaalisissa matemaattisissa esityksissä, kuten kaavoissa, muuttujien merkitsemiseen käytetään kirjaimia $\underline{x}$, $\underline{y}$ ja $\underline{z}$ sekä samoja kirjaimia alaindeksillä varustettuina, esim. $\underline{x}_1$, $\underline{x}_2$, $\underline{x}_3$, ... tai yleisesti $\underline{x}_i$ tai $\underline{x}_j$. Viiva kirjaimen alla viittaa siihen, että muuttujan arvo vaihtelee satunnaisesti, tiettyjen sallittujen arvojen puitteissa. Kun viiva jätetään pois, merkintä tarkoittaa satunnaismuuttujan saamaa numeerista arvoa; voidaan siis kirjoittaa $\underline{x}_3 = x$, missä x on yleisesti jokin muuttujan $\underline{x}_3$ numeerinen arvo. Merkintä $\underline{x}_i$, i=1,2,3...20, tarkoittaa, että muuttuja vaihtuu, kun alaindeksi vaihtuu.

Havaintomatriisi voidaan kirjoittaa yleisessä muodossa käyttämällä kirjaimia. Kirjain n on tilastoyksiköiden määrä aineistossa ja kirjain k muuttujien määrä.

	$\underline{\mathbf{x}}_1$	$\underline{\mathbf{x}}_2$	**…**	**…**	**…**	$\underline{\mathbf{x}}_k$
$\mathbf{a_1}$	x_{11}	x_{12}	…	…	$x_{1,k-1}$	…
$\mathbf{a_2}$	x_{21}	x_{22}	…	…	$x_{2,k-1}$	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
$\mathbf{a_{n-1}}$	$x_{n-1,1}$	$x_{n-1,2}$	…	…	$x_{n-1,k-1}$	$x_{n-1,k}$
$\mathbf{a_n}$	$x_{n,1}$	$x_{n,2}$	…	…	$x_{n,1-1}$	$x_{n,k}$

Merkintä x_{ij} tarkoittaa solussa ij olevaa muuttujan arvoa.

-- HARJOITUSTEHTÄVÄ --

Tehtävä 1. Määrittele tilastoyksikkö, perusjoukko ja tilastolliset muuttujat tutkittaessa

A. Suomen väestön ikärakennetta
B. vallitseeko sukupuolen ja alkoholin käytön välillä riippuvuutta
C. koulussa menestymisen ja sosiaaliryhmän välistä riippuvuutta
D. miten kunnan työttömyysaste vaikuttaa rikollisuuteen

Mittaaminen: Muuttujien ominaisuudet

Muuttujilla mitataan mm. mielipiteitä, olettamuksia, arvoja, asenteita, tietämistä tai taustatietoja. Muuttujien sisällölliset ominaisuudet ovat tärkeitä. On syytä pohtia, millaisella muuttujalla mitäkin asiaa kuvataan; mitä ominaisuusvaatimuksia tilastolliselle muuttujalle asetetaan. Näitä asioita on syytä pohtia jo lomakkeen suunnitteluvaiheessa. Näihin vaatimuksiin vaikuttaa myös yhtäältä se, millaisia tilastomenetelmiä halutaan myöhemmin soveltaa ja toisaalta se, miten tutkija uskoo saavansa vastaajilta mahdollisimman relevanttia tietoa. Myös tutkimustuloksia tulkittaessa on syytä tiedostaa, millaisen kysymyksen tai väittämän avulla kyseiseen tilastolliseen muuttujaan on päädytty.

On olemassa **numeerisia** ja **ei-numeerisia** muuttujia. Ei-numeeriset muuttujat mielletään usein nimenomaan mahdollisiksi kuvaamaan tilastoyksikön laatua. Kuitenkaan laatu ja määrä eivät ole toisistaan täysin erillisiä: sukassa olevien reikien määrä kertoo sukkien laadusta! Sairaalan osastolla toimivien hoitajien määrä kertoo hoidon laadun potentiaalisista mahdollisuuksista. Vaaleissa annetut äänimäärät kertovat ehdokkaiden suosiosta, mikä myös on eräänlaista laadun mittaamista. Myöskään numeerisen ja ei-numeerisen raja ei ole välttämättä täsmällinen: Esimerkiksi palkkaluokkamuuttuja on tietyssä mielessä numeerinen, sillä palkkaluokat voidaan määritellä numeroilla; toisaalta palkkaluokilla ei voi esimerkiksi suorittaa laskutoimituksia samoin kuin tavanomaisilla numeerisilla muuttujilla.

Muuttuja on **jatkuva**, kun sen kahden arvon välissä on ääretön määrä arvoja. Konkretisoituna se tarkoittaa, että lukuarvon perään voidaan aina lisätä desimaaleja; lukuarvo voidaan ilmoittaa aina tarkemmin ja tarkemmin. Lukusuoralla ajateltuna: kahden pisteen välistä voidaan aina osoittaa uusi piste. Esimerkiksi 1 km:n ja 1,1 km:n välissä on 1,01 km; edelleen 1 km:n ja 1,01 km:n välissä 1,005 km jne. Aika voidaan myös ilmoittaa tarkemmin ja tarkemmin: 1/10 000 sekuntia, 1/100 000 sekuntia tai vaikkapa miljardisosa sekuntia jne. Mittaustarkkuus on kuitenkin rajoitettu. Yhteiskunnallisia ilmiöitä tarkasteltaessa ajasta usein riittää vuoden tai puolen vuoden tarkkuus. Matkoissa riittää kilometrien tai kymmenien kilometrien tarkkuus, riippuen tutkimusaiheesta. Joskus muuttujan jatkuvuusominaisuutta halutaan korostaa esimerkiksi siten, että aikasarjoja kuvataan viivadiagrammilla tai pylväskuviota käytettäessä pylväät ovat kiinni toisissaan.

Muuttuja on **epäjatkuva** eli **diskreetti**, kun sen mitta-asteikolla siirrytään hyppäyksittäin arvosta toiseen. Tyypillinen epäjatkuva muuttuja on lukumäärä, esimerkiksi lasten lukumäärä: perheessä on 2 tai 3 lasta, ei 2,7456 lasta. Mielenkiintoista kuitenkin on, että suomalaisilla on Tilastokeskuksen mukaan vuonna 1999 keskimäärin 1,82 alle 18-vuotiasta lasta! Muuttujan jatkuvuus tai epäjatkuvuus voidaan huomioida graafisissa esityksissä: Esimerkiksi histogrammissa, jossa pylväät ovat kiinni toisissaan, jatkuvuusominaisuus korostuu verrattuna pylväsdiagrammiin. Jatkuvuusominaisuus liitetään yleensä kvantitatiivisiin muuttujiin, mutta tarkemmin ajateltuna myös monet kvalitatiiviset muuttujat ovat jatkuvia. Voisivatko esimerkiksi 'väri' tai 'hyvinvoinnin taso' olla jatkuvia muuttujia?

Muuttujan sanotaan olevan **dikotominen**, jos se saa kaksi arvoa: ominaisuus on olemassa kyseisellä tilastoyksiköllä tai sitä ei ole olemassa; henkilö on Suomen kansalainen tai ei ole. Mikäli muuttujan dikotomiaominaisuutta nimenomaan halutaan hyödyntää, se koodataan 0-1-muuttujaksi (0=ei, 1=kyllä). Tällaisia muuttujia kutsutaan **dummy-muuttujiksi**. Esimerkiksi

muuttujan 'siviilisääty' alkuperäisistä arvoista, naimisissa - naimaton - eronnut - leski, voidaan muodostaa neljä dummy-muuttujaa: naimisissa (0=ei, 1=kyllä), naimaton (0=ei, 1=kyllä), eronnut (0=ei, 1=kyllä) ja leski (0=ei, 1=kyllä). Dummy-muuttujia käytetään mm. regressioanalyysissä.

Hyvin tärkeä muuttujan ominaisuus on se, millä tasolla muuttuja määrittelee tilastoyksiköiden väliset erot. Tällöin on kysymys esimerkiksi siitä, onko muuttujan tehtävä pelkästään *luokitella* tilastoyksiköt jonkin ominaisuuden mukaan vai panna tilastoyksiköt ominaisuuden mukaiseen järjestykseen (ei lainkaan, vähän, paljon ominaisuutta) vai määritelläänkö tilastoyksiköiden erot muuttujalla jopa tarkkoina lukuarvoina ominaisuuden määrän mukaan. Tällaisessa **muuttujan mittaustason määrittämisessä** käytetään mitta-asteikkoja.

Mitta-asteikot ja mittaustaso

Jokaisen tilastollisen muuttujan mahdolliset arvot ovat peräisin tietyltä mitta-asteikolta. 'Mitta-asteikko' on helppo mieltää numerisiin muuttujiin. Kuitenkin kvantitatiivista tutkimusta tehtäessä puhutaan mitta-asteikosta myös ei-numeeristen eli sanallisten, ns. merkkitietomuuttujien yhteydessä. Mittaustaso riippuu mitta-asteikon tarkkuudesta ja mitattavasta asiasta; esimerkiksi henkilön tulot voidaan ilmoittaa valinnan mukaan joko täsmällisinä markkamäärinä tai palkkaluokkana, mutta onnellisuuden mittaaminen on epätarkempaa, sillä siihen ei ole mittayksikköä. Mitta-asteikkojen luokittelu neljään eri tyyppiin perustuu niiden ominaisuuksiin. Usein termejä 'mittaustaso' ja 'mitta-asteikko' käytetään synonyymeinä: puhutaan esimerkiksi sekä luokittelu**tason** että luokittelu**asteikon** muuttujasta.

Käytettävän mittaustason määrääminen kullekin muuttujalle on tärkeää jo tiedonkeruuta suunniteltaessa. Pääsääntönä voidaan pitää, että tieto kerätään mahdollisimman alkuperäisenä, siis tarkimmalla mahdollisella mittaustasolla. Toisinaan luotettavuusseikat edellyttävät tarkkuusvaatimusten lieventämistä. Viimeistään aineistoa analysoitaessa tutkija törmää mittaustasoihin, sillä tilastomenetelmät on kehitetty siten, että niillä on tiettyjä mitta-asteikkovaatimuksia. Käytännössä näistä matemaattisista vaatimuksista joudutaan usein tinkimään. Tämä on yleisesti hyväksyttyä silloin, kun ratkaisu palvelee tutkittavan ilmiön tutkimista ja esilletuomista. Vaikka mitta-asteikkovaatimuksissa joustetaan, on hyvä tietää, milloin käytetyn menetelmän vaatimukset toteutuvat ja milloin eivät.

Numeerinen mittaaminen: Välimatka-asteikko ja suhdeasteikko

Numeerisessa mittaamisessa on perinteisesti erotettu kaksi tasoa: välimatka-asteikko ja suhdeasteikko. Suhdeasteikko on vaativampi taso, sillä se edellyttää että muuttujan arvoilla on välimatka-asteikon ominaisuuksien lisäksi absoluuttinen nollapiste. Tätä vaatimusta ei kuitenkaan käytännössä yleensä tarvitse huomioida. Tällöin voidaan mittaustasojen hierarkiaan perustuen puhua **vähintään välimatka-asteikon** muuttujista. Numeerisilla muuttujilla voidaan luonnollisestikin suorittaa laskutoimituksia.

Välimatka-asteikosta käytetään myös nimityksiä välimatkatason asteikko ja intervalliasteikko. Välimatkan mittaaminen on määrällistä ts. numeerista mittaamista. Nimessä oleva 'välimatka' tai 'intervalli' viittaa siihen, että välimatka-asteikon muuttujan arvot ovat säännöllisen välimatkan päässä toisistaan. Siirryttäessä edellisestä seuraavaan asteikon pisteeseen, siirrytään aina täsmälleen saman verran. Tyypillinen välimatka-asteikon muuttuja on esimerkiksi syntymävuosi, jonka mittayksikkö on gregoriaanisen kalenterin vuosi. Asteikolla liikuttaessa mittayksikön verran, siirrytään aina yhtä pitkä aika. Numeerisiin muuttujan arvoihin liittyy yleensä aina mittayksikkö, kuten aikayksikkö vuosi. Kouluarvosana on sikäli harvinainen numeerinen muuttuja, että siihen ei liity varsinaista mitta-yksikköä (ellei se ole 'arvosana'). Suhdeasteikolla on kaikki välimatka-asteikon ominaisuudet, mutta lisäksi sillä on

"suhdeominaisuus": muuttujan arvojen suhde (eli toinen jaettuna toisella) pysyy samana, vaikka mittayksikköä muutetaan. Tämä tarkoittaa myös sitä, että muuttujalla on olemassa "absoluuttinen nollapiste", esimerkiksi 0 euroa. Suhdeasteikosta käytetäänkin myös nimitystä absoluuttinen asteikko. Mielenkiintoista on havaita, että ikämuuttujalla on nollapiste eli se on suhdeasteikon muuttuja, mutta syntymävuosimuuttujalla sitä ei ole. Tällä erottelulla ei kuitenkaan ole käytännön merkitystä. Myös lukumäärämuuttujat ovat suhdeasteikon muuttujia.

Yhteiskuntatieteellisessä tutkimuksessa suhdeominaisuudella on harvoin todellista merkitystä. Joskus jopa välimatkaominaisuuden merkitystä voidaan spekuloida. Mittaustarkkuuden määrääminen, onko se ajalle esimerkiksi kuukausi, päivä vai sekunti, ja mittausvirheiden olemassaolo, ovat joka tapauksessa huomioitava tiedonkeruuta suunniteltaessa. Mietittäessä, kuinka todellisia muuttujien käytetyt mitta-asteikot ja mittaustasot ovat sisällöllisesti, päästään mielenkiintoisiin pohdintoihin. Esimerkiksi vaaleissa laskettavien äänimäärien yksikkö on joko "ääni" tai "kpl". Yhden äänen lasketaan sisältävän saman verran kannatusta, antoipa sen aktiivipoliitikko tai puolipakolla äänestyspaikalle tuotu "nukkuva". Voidaan kysyä, mittaako "ääni" siis absoluuttista, todellista kannatusta. Toinen henkilö voi kannattaa ehdokasta enemmän kuin toinen. Yhden ihmisen sisäistämän kannatuksen määrää ei voida mitata kovin helposti. Virallisesti yksi Suomen kansalainen saa kannattaa yhden äänen verran yksissä vaaleissa ja kunkin äänen tulkitaan olevan yhtä paljon. Monien vähintään välimatka-asteikollisina pidettyjen muuttujien tarkempi sisällöllinen pohtiminen panee miettimään "välimatkojen" todellista merkitystä. Onko esimerkiksi kouluarvosanojen 4 ja 5 sisällöllinen välimatka sama kuin arvosanojen 7 ja 8?

Sanallinen mittaaminen

Vaikka mittaaminen mielletäänkin yleensä numeroihin, survey-tutkimuksissa mittaaminen on hyvin usein myös sanallista. Ollaan kiinnostuneita sellaisista tilastoyksikön ominaisuuksista, jotka on ilmaistava sanallisesti. Tällaisten sanallisten muuttujien mittaustaso on aina luokitteleva, mutta joissakin tapauksissa sillä on myös järjestysominaisuus.

Luokitteluasteikko

Luokitteluasteikosta käytetään myös nimityksiä luokittelutason asteikko, luokitusasteikko, laatueroasteikko ja nominaaliasteikko. Muuttuja, jolla on luokitteluominaisuus eli joka jakaa tilastoyksiköt tietyn ominaisuuden mukaisiin ryhmiin tai luokkiin, on luokittelutasoa. Tällaisia "ominaisuuksia" voivat olla esimerkiksi sukupuoli, siviilisääty ja kansalaisuus. Näiden muuttujien arvoilla (luokilla) ei ole mitään yksiselitteistä järjestystä.

Järjestysasteikko

Järjestysasteikosta käytetään myös nimityksiä järjestystason asteikko ja ordinaaliasteikko. Mikäli muuttujan arvot voidaan panna jonkin ominaisuuden mukaiseen järjestykseen, muuttujan mittaustaso on järjestysasteikko. Järjestämiseen ei tarvita tarkkaa mittayksikköä, millä välimatkoja mitattaisiin. Esimerkiksi Likert-asteikko muodossa 'täysin eri mieltä - jokseenkin eri mieltä - jokseenkin samaa mieltä - täysin samaa mieltä' on järjestysasteikko: mahdollisilla muuttujan arvoilla on yksiselitteinen järjestys; toiseen suuntaan samanmielisyys kasvaa ja toiseen vähenee.

Käytännössä sanallisten järjestysasteikkojen laatiminen kyselylomakkeisiin on joskus hankalaa: "Aina - usein - silloin tällöin - joskus - harvoin - ei koskaan". Sanalliset ilmaisut tarkoittavat eri ihmisille eri asioita: toiselle 'joskus' on harvemmin kuin 'harvoin', toiselle ilmaisujen merkitykset ovat päinvastaiset; jollekin taas 'silloin tällöin', 'joskus' ja 'harvoin' merkitsevät suunnilleen samaa. Vastausvaihtoehtojen järjestys kyselylomakkeessa kuitenkin kertonee tutkijan tarkoituksen. Lisäksi osa vastaajista haluaa ilmaista mielipiteensä hillitysti,

toiset antavat mielellään äärivastauksia. Vastautilanteessa on siten hyvä, että on tarpeeksi vaihtoehtoja. Aineiston käsittelyvaiheessa "liioista" luokista päästään yhdistämällä luokkia.

Mitta-asteikoiden yleisominaisuus on hierarkisuus niiden vaativuuden perusteella: suhdeasteikko, välimatka-asteikko, järjestysasteikko, luokitteluasteikko. Tämä tarkoittaa sitä, että vaativammalla asteikolla on myös vähemmän vaativan asteikon ominaisuudet: Järjestyasteikko järjestysominaisuuden lisäksi myös luokittelee. Suhdeasteikolla on erityisten suhdeasteikon ominaisuuksiensa lisäksi välimatka-asteikon, järjestysasteikon ja luokitteluasteikon ominaisuudet. Kysymykseen, mitä mittaustasoa tai mitta-asteikkoa muuttuja on, vastauksena ilmoitetaan vaativin mitta-asteikko eli se, jolla on eniten ominaisuuksia. Se siis "vaatii" muuttujalta eniten. Ikämuuttuja on suhdeasteikon muuttuja, koska sillä on absoluuttinen nollapiste kaikkien muiden mitta-asteikkojen ominaisuuksien lisäksi. Ikä-muuttuja on myös välimatka-asteikollinen, sillä iän mittaukseen on olemassa mittayksikkö, vuosi. Edelleen vuodet voidaan panna aikajärjestykseen (järjestysasteikko) ja iän avulla voidaan myös luokitella; puhekielessäkin käytetään ilmaisua "ikäluokka".

Pääsääntöisesti sanalliset muuttujat koodataan numeerisessa muodossa havaintomatriisiin. Sanoilla on vaikea tehdä laskutoimituksia, mutta numeeriset koodit mahdollistavat "sanoilla laskemisen". Matemaattisen eksaktisti tämä ei ole luvallista, mutta käytännön tutkimuksissa siitä on todettu olevan hyötyä ja se on yleisesti hyväksytty silloin, kun se on sisällöllisesti perusteltua, johdonmukaista ja tulkittavissa olevaa. Tyypillinen esimerkki tällaisesta mittaustasovaatimusten lieventämisestä on summamuuttujan laskeminen asenneväittämistä. Voidaan todeta, että jopa luokittelutason muuttujien välisistä suhteista on mahdollista saada informaatiota korrelaatiokertoimia laskemalla, muistaen kuitenkin, että luokittelutason riippuvuustarkasteluihin on olemassa selkeä ja kiistaton menetelmä, ristiintaulukointi.

Asenteita mitataan usein Rensis Likertin (1932) kehittämällä asteikolla, joka järjestää vastaajat "samanmielisyyden" määrän mukaan. Likert-asteikon vastausvaihtoehdot ovat 'täysin samaa mieltä', 'jokseenkin samaa mieltä', 'jokseenkin eri mieltä', 'täysin erimieltä'. Vastausvaihtoehtoihin voidaan lisätä vaihtoehtoja, jolloin asteikko voi olla esimerkiksi seuraavanlainen: 'täysin samaa mieltä', 'jokseenkin samaa mieltä', 'ei samaa eikä eri mieltä', 'jokseenkin eri mieltä', 'en osaa sanoa', 'en halua sanoa'. Analysointivaiheessa 'en osaa sanoa (eos)' ja 'en halua sanoa' vaihtoehdot voidaan määritellä puuttuvaksi tiedoksi.

-- HARJOITUSTEHTÄVIÄ --

Tehtävä 1. Mitä mittaustasoa ovat seuraavat muuttujat:

A. luokan oppilasmäärä
B. suuntautumisvaihtoehto (tai opintojakso)
C. vastaus muotoa: usein - harvoin - ei koskaan
D. ulospäinsuuntautuneisuus (ulkopuolinen tarkkailija havainnoi näkymättömissä 20 henkilön ryhmätilannetta ja antaa heille pisteitä 1-10 heidän käyttäytymisensä perusteella
E. juostu matka Cooperin testissä
F. kunto mitattuna Cooperin testin avulla
G. kunnassa tehtyjen rikosten määrä suhteessa kunnan asukaslukuun

Tehtävä 2. Mieti, missä 'en osaa sanoa' -vaihtoehdon paikka on Likert-asteikolla. Jos 'en osaa sanoa' -vaihtoehto on mukana, onko mittaustaso edelleen järjestysasteikko?

Tehtävä 3. Millaisia dummy-muuttujia tekisit seuraavasta osaWVS-aineiston muuttujasta? Mitä teet neljälle viimeiselle vaihtoehdolle?

"Entä mitä puoluetta Te ette äänestäisi missään nimessä?"

1. SDP (Suomen Sosialidemokraattinen puolue)
2. KESK (Suomen Keskusta)
3. KOK (Kansallinen Kokoomus)
4. VASEMMISTOLIITTO
5. RKP (Ruotsalainen Kansanpuolue)
6. Vihreä liitto
7. SKL (Suomen Kristillinen Liitto)
8. NUSU (Nuorsuomalainen Puolue)
9. PS (Perussuomalaiset)
10. Jokin muu ryhmittymä
11. En äänestäisi
12. En halua sanoa
13. En osaa sanoa

Mittaaminen: Mittarin luotettavuus

Operationalisointi

Monet yhteiskunta- ja käyttäytymistieteellisissä tutkimuksissa tarkasteltavat käsitteet ovat varsin abstrakteja. Tällaisia ovat älykkyys, onnellisuus, tasa-arvo, poliittinen aktiivisuus, suvaitsevaisuus jne. Mitä sinun mielestäsi on suvaitsevaisuus? - Jos viisi eri tutkijaa määrittelevät toisistaan tietämättä 'suvaitsevaisuuden', saamme viisi eri määritelmää. Kvantitatiivinen tutkimus edellyttää käsitteiden määrittelemistä sellaisiksi analyyttisiksi käsitteiksi, joita voidaan mitata. Tällaista käsitemäärittelyä ja mittareiden luontia kutsutaan **operationalisoinniksi**. Abstrakteista käsitteistä luodut mittarit ymmärretään yleensä kysymys- tai väittämäpatteristoiksi.

Alkula ym. (1995, ss. 75-76) erottavat neljä eri vaihetta operationalisoinnissa:

1. Käsitteen yleinen hahmottaminen ja määrittäminen
2. Käsitteen osa-alueiden määritteleminen
3. Siirtyminen teoreettisesta kielestä konkreettiseen arkikieleen ja indikaattoreihin
4. Operationalisoinnin tarkka kuvaaminen

Tutkijan on siis osoitettava selvästi, mitä tarkasteltava käsite hänen tutkimuksessaan tarkoittaa. Määrittelyprosessin alkuvaiheessa perehdytään aiheeseen liittyviin aikaisempiin tutkimuksiin ja muuhun kirjallisuuteen. Myös aiheesta käytävät keskustelut auttavat jäsentämään käsitettä, sen osa-alueita ja konteksteja.

Tutkimuksen luotettavuutta lisäävät operationalisoinnin vaiheiden esittäminen jäsennellysti ja konkreettisesti sekä lopullisten indikaattoreiden valinnan ja muotoilun huolellinen perusteleminen. Tämä helpottaa myös mittareiden ja samalla kokonaisten havaintoaineistojen uudelleenkäyttöä. Uudiskäyttäjän on tärkeää selvittää, missä viitekehyksessä käsitettä on käytetty ja mikä on ollut alkuperäisen tutkimuksen kohderyhmä.

Ajan kuluessa yhteiskunta muuttuu ja "samojen" ilmiöiden vertaaminen eri ajankohtina vaikeutuu. Se luo tarvetta myös mittareiden muuttamiseen. Jos halutaan tietää, uuvuttavatko työntekijöitä nyt samat asiat kuin 20 vuotta sitten, voidaan ehkä käyttää samoja mittareita kuin aiemmin. Jos sen sijaan tutkimuksen kohteena on ihmisten nykyinen kokemus työuupumuksesta, mittaria on uudistettava.

Operationaalistamisprosessissa on tärkeää pitää mielessä validiteetti- ja reliabiliteettivaatimukset.

Esimerkki 1.

Raigo Liiman (2000) on tutkinut pro gradu -työssään uskonnollisuutta virolaisten ja Virossa asuvien venäläisten keskuudessa. Tutkimusaineistot ovat vuosilta 1991, 1992, 1994 ja 1998. Hän on määritellyt uskonnollisuuden uskontoon kuuluvaksi tai perustuvaksi, sille ominaiseksi riitiksi, toimitukseksi tai vakaumukseksi. Tutkimuksessa on sovellettu Rodney Starkin ja Charles Y. Glockin (1968) kehittämää uskonnollisten ulottuvuuksien teoriaa, ja uskonnollisuus on jaettu teoreettisesti neljään eri ulottuvuuteen: ideologinen ulottuvuus, rituaalinen ulottuvuus, institutionaalinen ulottuvuus ja seurausten ulottuvuus. Edelleen esimerkiksi rituaalinen ulottuvuus jaetaan yksityiseen ja julkiseen hartauden harjoitukseen. Yksityistä hartauden harjoitusta ovat mm. Raamatun lukeminen, rukoileminen ja mietiskely; julkista ovat mm. kirkossa ja häissä käynti sekä sakramentteihin osallistuminen.

Valmiiden aineistojen mittarit eivät aina ole niin kattavia kuin tutkija haluaisi. Tällöin voidaan käyttää useampia aineistoja, kuten tässä työssä on tehty. Vuoden 1998 aineistossa (Suomen Akatemia) rituaalista uskonnollisuutta on mitattu julkiseen uskonnonharjoitukseen liittyvällä kysymyksellä "Kuinka usein käytte jumalanpalveluksessa?" ja yksityiseen uskonnonharjoitukseen liittyvällä kysymyksellä "Kuinka usein rukoilette Jumalaa?" Vuonna 1994 Virossa tehdyssä kyselyssä rituaalista ulottuvuutta on selvitetty kysymyksillä: "Seuraatteko uskonnollisia ohjelmia televisiosta tai radiosta?" ja "Kuinka usein te luette Raamattua?" ja "Onko teillä kotona Raamattu tai Uusi testamentti?"

Esimerkki 2.

Kun halutaan tutkia sosiaalista toimintaa, pitää määritellä, mitä sosiaalinen toiminta on. Operationalisoinnin perustana on käsitteen perinpohjainen määritteleminen. Esimerkin tämän erittäin laaja-alaisen ja abstrahoidun käsitteen määrittelemisen vaikeudesta saa Ilpo Vilkunan (1998) artikkelista "Sosiaalisen taidon metsästys - Lähtökohtia sosiaalisen kyvykkyyden ymmärtämiseen", jossa hän käy läpi erilaisia tapoja ymmärtää 'sosiaalinen toiminta' ja nähdä sen osa-alueita. Hän esittää mm. Ziglerin ja Trickettin (1977) määrittelemät sosiaalisen kompetenssin ulottuvuudet: fyysinen terveys ja hyvinvointi, formaalinen kognitiivinen kyky, suorituskyky, motivaation/emootion taso. Jopa fyysinen terveys voidaan siis nähdä 'sosiaaliseen toimintaan' kuuluvana osa-alueena.

-- HARJOITUSTEHTÄVIÄ --

Tehtävä 1. Millaisiksi kysymyksiksi operationalisoisit työuupumuksen? Käy huolellisesti läpi operationalisoinnin vaiheet. Käytä Alkulan ym. operationalisoinnin vaihekuvausta ja aiheeseen liittyvää kirjallisuutta hyväksesi. Rajaa kuvitellun tutkimuksen kohderyhmä.

- Mitä työuupumus on?
- Millaisia osa-alueita voidaan työuupumuksessa nähdä erityisesti tässä kohderyhmässä? Voisiko näitä osa-aluejaotteluita olla useita?
- Millaisissa konkreettisissa tilanteissa työuupumus näkyy?
- Millaisilla kysymyksillä tai väittämillä konkreettisia työuupumuksen merkkejä voidaan selvittää?
  Laadi mittari.

Arvioi mittaria. Mieti mm. onko jokin työuupumuksen osa-alue painottunut liikaa. Onko jotakin jäänyt huomaamatta? Ovatko mittarissa käytettävät sanat ja lauseet yksiselitteisiä ja ymmärrettäviä?

Tehtävä 2. Mieti, millaisilla konkreettisilla kysymyksillä sinä mittaisit tutkimuksessasi uskonnollisuuden rituaalista ulottuvuutta. Lue ensin esimerkki 1. Tutki, mitä varannon verkkoversiosta löytyvän World Values Survey osa-aineiston kysymyksiä voisit käyttää rituaalisen ulottuvuuden mittareina.

Mittarin validiteetti

Mittarin **validiteetilla** tarkoitetaan sen pätevyyttä eli sen hyvyyttä mitata juuri sitä, mitä sen on tarkoitus mitata - tarpeeksi kattavasti ja tehokkaasti. Mittaria on osattava käyttää oikeaan kohteeseen, oikealla tavalla ja jotta se tavoittaa kohteen, myös oikeaan aikaan. Esimerkiksi epäonnistunut otanta, mittauksen ajankohta tai jopa haastateltavan ja haastattelijan välinen henkilökemia voivat aiheuttaa "epäpätevyyttä" mittarin käytössä. Lähtökohdiltaan virheellinen tutkimusasetelma vaikuttaa ratkaisevasti tutkimuksen kokonaisvaliditeettiin. Yksittäisen mittarin hyvä validiteetti onkin välttämätöntä tutkimuksen kokonaisvaliditeetin kannalta.

Validiteetin käsitettä on kirjallisuudessa luokiteltu. Esimerkiksi jos valintakoe ennustaa hyvin opinnoissa menestymistä, sen ennustevaliditeetti on hyvä. Validiteetille on määritelty myös muita "erityisnimiä", kuten sisällöllinen validiteetti, samanaikaisvaliditeetti, rakennevaliditeetti ja prosessivaliditeetti (ks. Alkula ym. 1995, s. 91-92 ja Nummenmaa ym.1997, s. 203-204). Nämä validiteetin lajit voidaan nähdä sekä yksittäisten mittareiden validiteettia että koko tutkimuksen validiteettia arvioitaessa. Esimerkiksi sisällöllisen validiteetin käsite korostaa, että mittari todella mittaa sisällöllisesti sitä, mitä sillä halutaan mitata. Jos valintakokeen ennustevaliditeetti on hyvä, voitaisiin myös sanoa, että sen sisällöllinen validiteetti on hyvä, koska sen tarkoituksena on nimenomaan toimia opintomenestyksen ennustajana. Tällöin mittariin on osattu valita sisällöllisesti oikeita asioita. Vastavasti muutkin validiteetit tarkoittavat itse asiassa samaa asiaa. Maxwellin ja Dalaneyn sanoilla: "Validiteetti tarkoittaa pohjimmiltaan totuutta tai virheettömyyttä, vastaavuutta todellisuuden ja siitä tehtyjen väittämien välillä." (Maxwell & Delaney in Designing Experiments and Analyzing Data).

Validi mittari on tulos onnistuneesta operationalisoinnista. Käsiteanalyysin loogisella ja täsmällisellä argumentoinnilla vahvistetaan operationalisoinnin uskottavuutta, sillä tutkittavat ilmiöt voidaan käytännössä operationalisoida hyvinkin erilaisiksi mittareiksi. On hyvä, jos lukija voi prosessia seuraten itse arvioida mittarin pätevyyttä ja vakuuttua siitä.

Operationalisoinnissa voidaan epäonnistua. Hankaluuksia voi aiheuttaa itse käsitteen määrittely. Sanojen valinta lopullisiin mittareihin tuottaa päänvaivaa. Voi olla, että vastaaja ei esimerkiksi iästään tai sosiaalisesta asemastaan johtuen ymmärrä lainkaan tai ymmärrä samalla tavalla kysymyksiä kuin tutkija. Tutkijan kieli voi olla abstraktia, yksittäisillä täytesanoilla voi olla erilainen painoarvo eri ihmisille, valmiista mittareista lainatut sanat tai käsitteet voivat olla vahentuneita. Kulttuurin huomioiminen voi unohtua kokonaan: Yhdysvalloissa laadittu mittari ei olekaan kulttuurierojen vuoksi pätevä Suomessa tai Suomessa käytettävä mittari ei välttämättä toimi Ruotsissa.

Mittarin validiteettin testaamiseen on yritetty löytää erilaisia keinoja. Esimerkiksi "äärivastaajien" suhteen voidaan saada jotakin informaatiota Likert-asteikollisen mittarin validiteetista, kun lasketaan samojen vättämien kieteisten ja myönteisten versioiden korrelaatiot. Erisuuntaiset väittämät korreloivat keskenään voimakkaasti, jos ne todella mittaavat samaa asiaa.

Joissakin tilanteissa mittarin validiteettia on mahdollista testata kriteerimuuttujan tai -muuttujien avulla. Esimerkiksi valintakoemittareilla saatuja tuloksia voidaan verrata myöhempiin opintosuorituksiin, jos halutaan osoittaa, että valintakokeella on saatu opinnoissaan menestyviä opiskelijoita. Mittarin pätevyyttä ei voida tällöinkään todeta täydellisesti, sillä asetelmassa jäävät havaitsematta mm. ne hylätyt pyrkijät, jotka olisivat menestyneet opinnoissaan. Tällainen jälkikäteen tapahtuva mittarin vertaaminen kriteerimuuttujaan on hyödyllistä silloin, kun kehitetään toistuvasti käytettäviä mittareita. Mittareiden kehittäminen vie aikaa ja muita resursseja. Ajan kuluminen voi asettaa mittarille myös muutospaineita. Procter (1998, s. 129) toteaa, että on lähes mahdotonta laskea kvantitatiivista mittaa mittarin validiteetille, minkä vuoksi on vain parasta pitää ongelma mielessään ja etsiä keinoja validiteetin parantamiseksi.

-- HARJOITUSTEHTÄVÄ --

Tehtävä 3. Mieti konkreettisia esimerkkejä. Millaisia asioita ei voida mitata samoilla mittareilla esimerkiksi Albaniassa ja Suomessa?

Mittarin reliabiliteetti

Reliabiliteetti-sana voidaan suomentaa sanoilla 'luotettavuus', 'käyttövarmuus' ja 'toimintavarmuus'. Kvantitatiivisen tutkimuksen kielessä sillä tarkoitetaan mittarin johdonmukaisuutta; sitä, että se mittaa aina, kokonaisuudessaan samaa asiaa. Arkikielen 'luotettavuus' on tutkimuksen kielessä validiteetti. Mittarilla tarkoitetaan tässä yhteydessä samaa asiaa mittaavaa asenneväittämä- tai kysymysjoukkoa. Jos mittari on täysin reliaabeli, siihen eivät vaikuta satunnaisvirheet eivätkä olosuhteet.

Reliabiliteetissa erotetaan kaksi osatekijää: **stabiliteetti** ja **konsistenssi**. Stabiliteetissa on kysymys mittarin pysyvyydestä ajassa. Epästabiilissa mittarissa näkyvät olosuhteiden ja vastaajan mielialan ynnä muiden satunnaisvirheiden vaikutukset helposti. Mittarin pysyvyyttä voidaan tarkastella vertaamalla useampia ajallisesti peräkkäisiä mittauksia. Tällöin aikavälin pituus tulisi osata optimoida: Sen pitää olla tarpeeksi pitkä, jotta vastaaja ei muista vastauksiaan, mutta toisaalta niin lyhyt, ettei todellisia muutoksia asioissa ole ehtinyt tapahtua. Monissa tapauksissa tämä reliabiliteetin mittaustapa ei ole toteuttamiskelpoinen, sillä huono reliabiliteettikerroin voidaan usein helpommin selittää ajassa tapahtuneilla todellisilla muutoksilla kuin epästabiililla mittarilla (Wright 1979, s. 47).

Mittarin konsistenssilla eli yhtenäisyydellä tarkoitetaan sitä, että kun useista väittämistä koostuva mittari jaetaan kahteen joukkoon väittämiä, kumpikin väittämäjoukko mittaa samaa asiaa. Tällöin molempien väittämäjoukkojen kokonaispistemäärien välinen korrelaatiokerroin saa suuren arvon. Koska ei ole mitään ulkoista kriteeriä, jolla testattaisiin mittarin reliaabeliutta, on tyydyttävä edellä kuvatulla tavalla "sisäisiin" kriteereihin eli samaan tutkimusjoukkoon ja mittariin itseensä. (Procter 1998, s. 128). Tämän toteamiseksi yleisesti käytetään mm. Cronbachin alfakerrointa, joka perustuu väittämien välisiin korrelaatioihin. On kuitenkin todettava, että on mahdollista luoda väittämäpatteristo, joka sisältää täysin eri asioita mittaavia, mutta keskenään voimakkaasti korreloivia muuttujia. Toisaalta saman ilmiön osa-alueita mittaavat muuttujat eivät aina välttämättä korreloi keskenään ja kuitenkin niitä on tarpeen tarkastella yhdessä.

Samalla reliabiliteetin käsitteellä on määritelty siis kaksi varsin erilaista mittarin ominaisuutta. Stabiili mittari ei välttämättä ole konsistentti eikä konsistentti mittari välttämättä stabiili. Käytännössä reliabiliteetti liitetään pääasiassa mittarin konsistenssiin. Vaikka mittari olisi sekä konsistentti että stabiili, se ei riitä. Tutkimuksen mittari voi nimittäin mitata vääräkin asiaa hyvin johdonmukaisesti. Mittarin on oltava myös validi.

Paljon käytetty tunnusluku reliabiliteetin mittaamiseksi on Cronbachin α (alfa). Sillä mitataan nimenomaan mittarin konsistenssia eli yhtenäisyyttä. Cronbachin alfa lasketaan muuttujien välisten keskimääräisten korrelaatioiden ja väittämien lukumäärän perusteella. Mitä suurempi alfan arvo on, sitä yhtenäisempi mittarin voidaan katsoa olevan. Käytännössä kannattaa kokeilla, mikäli mahdollista, eri muuttujakombinaatioita ja verrata saatuja alfan arvoja. (Ks. verkosta lisäesimerkki 4). Reliaabeliutta osoittamaan voidaan laskea alfa-kerroin myös käyttäen ns. puolitusmenetelmää (*Split-Half*), jolloin muuttujat jaetaan kahteen ryhmään ja alfa-kertoimet lasketaan kummallekin osiolle. Guttmanin mukaan voidaan laskea useita alimpia rajoja (*Lower Bounds*).

Alfan standardoitu estimaatti voidaan laskea seuraavalla kaavalla:

$$a = \frac{k * \bar{r}}{1 + (k-1) * \bar{r}}$$

jossa

$\bar{r}$ = väittämien välinen keskikorrelaatio eli väittämien välisten Pearsonin korrelaatiokertoimien keskiarvo.

k = väittämien lukumäärä. (SPSS 1999, s. 362)

Reliabiliteetista ollaan yleensä kiinnostuneita sen vuoksi, että väittämäpatteriston muuttujat halutaan tiivistää summamuuttujaksi. Reliabiliteettia kuvaava tunnusluku lasketaan niille muuttujille, joita on tarkoitus yhdistää. Tällöin väittämien koodaus tulee olla sama, kuin se on summamuuttujaa laskettaessa. Joskus joidenkin muuttujien koodaus pitää kääntää.

Teknisesti reliabiliteettia saadaan parannettua, kun jätetään alfa-kertoimen arvoa alentavia muuttujia pois. Tällöin voi kuitenkin mittarin validiteetti kärsiä eli mittari ei enää olekaan kattava. On siis mietittävä myös sisällöllisesti, mitä poistetaan. Ainakin monitulkintaiset väittämät on syytä jättää pois.

Reliaabeliuden perusajatus voidaan esittää SPSS-ohjelmiston oppaasta suomennetulla yksinkertaisella lauseella: "Reliaabeliin kyselyyn annetut vastaukset eroavat, koska vastaajilla on erilaisia mielipiteitä - ei sen vuoksi, että kysely on hämmentävä tai monitulkintainen." (SPSS 1999, s.362)

Otos ja otantamenetelmät

Määrällinen yhteiskuntatieteellinen tutkimus pyrkii kuvailemaan ja selittämään tutkimuksen kohteena olevia ilmiöitä järjestelmällisten havaintojen avulla. Empiirisen havainnoinnin eli mittauksen kohteita voidaan kutsua **havaintoyksiköiksi** (*unit of observation*). Havaintoyksikkö määräytyy tutkimusongelman perusteella. Esimerkiksi jos tutkimuksella halutaan tietoa suomalaisen aikuisväestön mielipiteistä, havaintoyksikköinä ovat siihen kuuluvat henkilöt. Jos tutkija haluaa tietoa suomalaisista kunnista, havaintoyksikköinä ovat Suomen kunnat jne.

Havaintoyksikön valinnan jälkeen tutkijan tulee ratkaista, kuinka monesta havaintoyksiköstä hän kerää tietoa. Kaikkien havaintoyksiköiden muodostamaa kokonaisuutta kutsutaan tutkimuksen **perusjoukoksi** (*population*). Varmin tapa saada määrällistä tietoa tutkimuskohteesta on mitata halutut ominaisuudet jokaisesta tutkimuksen perusjoukkoon kuuluvasta havaintoyksiköstä. Käytännön syistä tämä on kuitenkin usein mahdotonta. Esimerkiksi kaikkien suomalaisten haastatteleminen olisi lähes mahdoton tehtävä muun muassa sen vaatimien resurssien takia. Tämän vuoksi tutkimuksessa useimmiten keskitytään perusjoukkoa pienemmän, satunnaisesti valitun havaintoyksikköjoukon eli **otoksen** tutkimiseen. Tilastollisen päättelyn avulla otoksesta saatuja tietoja voidaan käyttää hyväksi tehtäessä päätelmiä koko perusjoukosta. Tilastollisten päätelmien pätevyys riippuu muun muassa siitä, kuinka hyvin otoksen valinta eli otanta on suoritettu. Tämän vuoksi on tärkeää ymmärtää otannan perusperiaatteet ja erilaisten otantamenetelmien luonne.

Otos ja näyte

Tutkimuksen kohteena olevat perusjoukkoa pienemmät havaintoyksikköjoukot voidaan jakaa otoksiin ja näytteisiin (*probability sample* ja *non-probability sample*). Otos on sellainen havaintoyksikköjen joukko, johon kaikilla havaintoyksiköillä on tiedossa oleva nollaa suurempi todennäköisyys tulla valituksi. Näytteessä havaintoyksikköjen valinta on usein harkinnanvarainen, eikä havaintoyksikköjen todennäköisyyttä tulla valituksi tiedetä. Yleensä määrällisessä tutkimuksessa suositaan otosaineistoja, koska niiden avulla pystytään tekemään paremmin tilastollisia yleistyksiä perusjoukkoon.

Tarkasti perusjoukon ominaisuuksia kuvastavaa otosta kutsutaan **edustavaksi otokseksi** (*representative sample*). Edustavan otoksen saamiseksi täytyy varmistaa, että mitään havaintoyksikköjen ryhmää ei systemaattisesti suosita tai suljeta otoksen ulkopuolelle. Jos käyntikysely tehtäisiin niin, että haastattelijat pyrkisivät tavoittamaan haastateltavat kotiosoitteista vain keskellä päivää, ei tuloksena olisi edustava otos, koska työssäkäyvien osuus otoksesta olisi huomattavasti pienempi kuin heidän osuutensa perusjoukosta. Varmin tapa saada otoksesta edustava on käyttää satunnaisuutta hyväksi otosta valittaessa. Käytännössä tämä tarkoittaa sitä, että otokseen valikoidut havaintoyksiköt "arvotaan" satunnaisesti.

Joissakin tapauksissa satunnaisotoksen saaminen perusjoukosta on mahdotonta. Varsin usein tutkijalla ei ole käytettävissään tietoja kaikista havaintoyksiköistä, jolloin niiden satunnainen valinta koko perusjoukosta on mahdotonta. Tällöin tutkijan on tyydyttävä harkinnanvaraiseen **näytteeseen**. Tällaiseen tilanteeseen joudutaan usein esimerkiksi sosiologian alalla tutkittaessa erilaisten alakulttuurien jäseniä kuten huumeiden käyttäjiä tai prostituoituja. Huumeiden käyttäjistä ei ole saatavilla minkäänlaista listaa, josta otanta voitaisiin suorittaa. Itse asiassa edes perusjoukon koosta ei ole kovinkaan tarkkoja tietoja. Tässä tapauksessa tutkija saattaa aloittaa tutkimuksensa muutamasta tuntemastaan huumeiden käyttäjästä, haastatella heitä ja sen jälkeen pyytää heiltä vinkkejä uusista haastateltavista. Toinen vaihtoehto voisi olla huumevieroitusklinikan asiakkaiden haastattelu. Kumpaakin menetelmää käyttäen tuloksena olisi

näyte, koska valittujen havaintoyksiköiden edustavuudesta suhteessa perusjoukkoon ei olisi mitään taetta.

Yksinkertainen satunnaisotanta

Perustavanlaatuinen otantamenetelmä on ns. **yksinkertainen satunnaisotanta** (*simple random sampling*). Siinä kaikilla perusjoukon havaintoyksiköillä on samansuuruinen todennäköisyys tulla valituksi otokseen.

Käytännössä yksikertainen satunnaisotanta etenee vaiheittain. Ensimmäisessä vaiheessa tutkijalla täytyy olla käytettävänään lista kaikista perusjoukon havaintoyksiköistä (eli ns. otantakehikko). Oletetaan, että tutkija haluaa tehdä otokseen perustuvan tutkimuksen Suomen kuntien taloudellisesta tilasta vuonna 2000 ja hänellä on aakkosellinen lista kaikista Suomen kunnista. Vuonna 2000 Suomessa oli 452 kuntaa. Otannan toteuttamisen helpottamiseksi tutkija numeroi havaintoyksikkönsä alkaen numerosta yksi, jonka saa Alahärmän kunta. Sen jälkeen Alajärvi saa numeron kaksi, Alastaro numeron kolme jne. Aakkosissa viimeinen kunta (Äänekoski) saa numeron 452. Kannattaa huomata, että näitä numeroita ei pidä sekoittaa yleisesti käytettyyn viralliseen kuntanumerointiin, jota kannattaa käyttää kuntien tunnuksena aineistossa.

Seuraavaksi tutkijan täytyy päättää haluamansa otoksen koko. Tätä varten on olemassa erilaisia sääntöjä, jotka liittyvät siihen, kuinka tarkasti otoksesta saadut tulokset voidaan yleistää perusjoukkoa koskevaksi. Suomalaisissa valtakunnallisissa tutkimuksissa käytetään yleensä vähintään tuhannen hengen otoksia, jolloin tulosten luottamusväli on muutaman prosenttiyksikön luokkaa (ks. tarkemmin tilastollinen päättely). Yleisesti ottaen otoskoko on suhteutettava tutkimustarpeisiin ja käytettävissä oleviin resursseihin. Jos perusjoukko on pieni, kannattaa tehdä niin sanottu kokonaistutkimus eli kerätä tiedot kaikista perusjoukon jäsenistä.

Oletetaan, että kuntatutkija haluaa otokseensa 50 kuntaa. Otoksen valintaa varten tutkija tarvitsee 50 satunnaislukua välillä 1-452. Nämä satunnaisluvut voidaan poimia esimerkiksi tilastollisten taulukkokirjojen satunnaislukutaulukoista. Kätevä tapa on aloittaa satunnaisesti jostain taulukon osasta ja katsoa, minkä luvun kolme seuraavaa taulukon numeroa muodostavat. Jos tämä luku on välillä 001-452, kirjoitetaan se muistiin ja siirrytään seuraavaan kolmen satunnaisluvun muodostamaan lukuun. Jos luku on suurempi kuin 452, siirrytään suoraan seuraavaan lukuun. Tätä prosessia toistetaan, kunnes tutkijalla on lista 50 satunnaisesta luvusta väliltä 1-452. Satunnaislukujen valinnassa voidaan käyttää hyväksi myös tarkoitukseen soveltuvia tietokoneohjelmia. Otoksen muodostamisen lopuksi kuntalistasta valitaan 50 satunnaislukujen osoittamaa kuntaa, jotka näin muodostavat tutkimuksen otoksen.

Yksinkertainen satunnaisotos on periaatteiltaan helppo ymmärtää ja on usein myös helppo toteuttaa. Monissa tapauksissa ei kuitenkaan ole helppo saada listaa kaikista perusjoukon havaintoyksiköistä, jolloin menetelmän käyttö on mahdotonta. Kyselytutkimuksissa perusjoukko on usein suuri ja laajalle alueelle hajaantunut. Näin on esimerkiksi tilanteessa, jossa tutkitaan henkilökohtaisten haastattelujen avulla suomalaisten kulutustottumuksia. Jos haastateltavien valinta perustuisi yksinkertaiseen satunnaisotantaan, henkilökohtaisten haastattelujen tekeminen vaatisi suuria määriä resursseja, koska haastattelijat joutuisivat matkustamaan ympäri Suomea satunnaisotokseen valikoituneiden henkilöiden asuinpaikkojen mukaan. Tällaisissa tutkimustilanteissa käytetäänkin usein muunlaisia otantamenetelmiä, esimerkiksi ryväsotantaa.

Systemaattinen satunnaisotanta

Systemaattinen eli tasavälinen otanta (*systematic sample*) on tavallaan pelkistetty versio yksinkertaisesta satunnaisotannasta. Myös systemaattista otantaa varten tutkija tarvitsee listan perusjoukon havaintoyksiköistä. Poimintavälin määrittelemiseksi on laskettava otoksen suhteellinen koko perusjoukosta. Jos esimerkiksi oletetaan, että perusjoukkoon kuuluu 500

havaintoyksikköä ja otoskoko on 100, saadaan suhteelliseksi otoskooksi 1/5 (=100/500). Näin ollen havaintoyksikkölistasta poimitaan joka viides havainto otokseen.

Ennen otoksen poiminnan aloittamista täytyy päättää, mistä kohdasta havaintoyksikköjen listaa otoksen valinta aloitetaan. Tässä voidaan käyttää hyväksi satunnaislukutaulukoita. Systemaattinen otanta etenee tämän jälkeen niin, että listasta poimitaan otokseen joka viides havaintoyksikkö aloittaen satunnaisesti valitusta lähtökohdasta. Jos lista loppuu ennen kuin havaintoyksikköjä on saatu poimittua tarpeellinen määrä, jatketaan prosessia taulukon alusta.

Systemaattinen satunnaisotanta on teknisesti erittäin helppo toteuttaa, mutta siihen liittyvät samat ongelmat kuin yksinkertaiseen satunnaisotantaan. Tämän lisäksi ongelmia aiheutuu, jos havaintoyksikkölista noudattaa jotain säännöllistä jaksollisuutta. Jos esimerkiksi tiedot perusjoukosta koostuvat pariskunnista ja poimintaintervalli on parillinen luku, seurauksena voi olla, että otokseen saattaisi valikoitua ainoastaan joko miehiä tai naisia.

Ositettu otanta

Ositetun otannan avulla pyritään varmistamaan, että otos on mahdollisimman edustava tutkimuksen kannalta merkittävien ryhmien osalta. Edustavassa otoksessa tärkeät ryhmät ovat edustettuina otoksessa samassa suhteessa kuin perusjoukossa. Joskus jokin ryhmä voi olla niin pieni, että yksinkertainen satunnaisotanta ei pysty varmistamaan, että ryhmän edustus toteutuisi otoksessa. Esimerkkinä voidaan käyttää jo edellä mainittua kuntatutkijaa, joka haluaa tutkia suomalaisia kuntia otoksen perusteella. Tutkijaa kiinnostaa erityisesti asukasluvultaan suurten kaupunkikuntien ja asukasluvultaan pienten maalaiskuntien erot ja hän haluaa varmistaa, että näiden kaupunkien osuus otoksessa on yhtä suuri kuin niiden osuus kaikkien kuntien joukosta. Suhteellisesti oikean kokoisen edustuksen otoksessa voi varmistaa käyttämällä **ositettua otantaa** (*stratified sampling*).

Ositetussa otannassa käytetään hyväksi etukäteistietoja perusjoukon jakautumisesta ryhmiin. Esimerkiksi vuonna 2000 Suomessa oli Tilastokeskuksen luokittelun mukaan 67 kaupunkimaista kuntaa. Suhteellisesti näitä kaupunkikuntia oli siis noin 15 prosenttia perusjoukosta. Koska tutkija haluaa varmistaa, että kuntaotokseen sisältyy yhtä suuri osuus kaupunkimaisia kuntia kuin muita kuntia, hän jakaa ensin kunnat näihin kahteen ryhmään. Oletetaan lisäksi, että hän haluaa otokseensa yhteensä 100 kuntaa. Varmistaakseen erityyppisten kuntien edustavuuden hän poimii otokseen 15 kuntaa kaupunkikuntalistalta ja 85 kuntaa maalaiskuntalistalta. Tämä menetelmä varmistaa, että lopullisessa otoksessa kaupunkimaisten ja muiden kuntien suhteellinen osuus on sama kuin perusjoukossa. Yksittäisten kuntien poiminta kahdelta listalta voidaan tehdä esimerkiksi käyttäen yksikertaista satunnaisotantaa.

Edellinen esimerkki ositetusta otannasta on hyvin yksinkertainen. Käytännössä luokittelevia muuttujia voi olla useita, jolloin perusjoukko täytyy jakaa useampaan ryhmään ennen otannan suorittamista. Kuntatutkija voisi esimerkiksi haluta, että otoksessa toteutuu myös kuntien maantieteellinen jakauma edustavasti. Tämä varmistuu jakamalla kunnat kuntamuodon lisäksi läänien mukaan ja poimimalla näistä ryhmistä oikea määrä kaupunkimaisia ja muita kuntia.

Ositetun otannan käyttöön suurissa kyselytutkimuksissa liittyy samoja ongelma kuin yksinkertaiseen ja systemaattiseen satunnaisotantaan. Otokseen valikoituneet vastaajat voivat olla levittäytyneinä suurella maantieteelliselle alueella ja näin heidän haastattelemisensa vaatii paljon matkustamista ja siihen liittyviä kuluja.

Ryväsotanta

Ryväsotantaa (*cluster sampling*) käytetään yleensä suuria haastattelututkimuksia tehtäessä. Tavoitteena on vähentää tietojen keruun aiheuttamia kustannuksia samalla varmistaen, että otos on kuitenkin mahdollisimman edustava. Ryväsotantaa voidaan hyödyntää myös silloin, kun tutkijalla ei ole käytettävissään kattavaa listaa kaikista havaintoyksiköistä.

Ryväsotanta koostuu useasta eri otoksesta. Ajatuksena on, että ensin tehdään otanta havaintoyksikköjä suuremmista kokonaisuuksista, jonka jälkeen valitaan näistä kokonaisuuksista varsinaiseen otokseen tulevat havaintoyksiköt. Oletetaan, että tutkimustehtävänä on selvittää sairaalapotilaiden tyytyväisyyttä heidän saamansa hoitoon. Kaikilla sairaaloilla on omat potilasrekisterit, mutta tutkijalla ei ole käytettävissään kattavaa tietoa kaikista maan potilaista. Hänellä on kuitenkin apunaan lista kaikista Suomen sairaaloista.

Ryväsotanta etenee niin, että ensin tutkija ottaa haluamansa kokoisen otoksen sairaaloista. Tässä vaiheessa voidaan käyttää muita edellä esitettyjä otantamenetelmiä, esimerkiksi yksinkertaista satunnaisotantaa. Tämän jälkeen tutkija voi pyytää valituista sairaaloista listat heidän potilaistaan ja poimia varsinaisen otoksen näistä listoista. Menetelmän ilmeisenä etuna on se, että potilashaastattelut voidaan rajoittaa valittuun määrään sairaaloita, mikä vähentää tiedonkeruun kustannuksia. Samaa menetelmää voidaan käyttää esimerkiksi tutkittaessa jonkin kaupungin asukkaiden mielipiteitä. Ensimmäisessä vaiheessa valitaan otos kaupungin alueista, ja sen jälkeen varsinainen otos poimitaan näistä valituista alueista.

Postikyselyaineiston kokoaminen

Satunnaisesti valittuihin vastaajaotoksiin perustuvia joukkohaastattelututkimuksia on tehty yli sadan vuoden ajan. Tällaisten tutkimusten suosio kasvoi merkittävästi 1900-luvun puolivälin jälkeen. Tähän vaikutti tietokoneiden keksiminen ja käyttöönotto mm. markkinatutkimusten ja myös yhteiskunta- ja käyttäytymistieteiden apuvälineenä.

Joukkohaastattelututkimukset ovat alusta saakka enimmäkseen perustuneet malliin, jossa otokseen kuuluvat henkilöt vastaavat kohtuullisen haastatteluajan puitteissa valmiiksi laadittuihin kysymyksiin niiden vastausvaihtoehtojen pohjalta. Rakenteensa vuoksi niiden tiedonkeruutapaa voidaan kutsua strukturoiduksi. Englannin kieltä mukaillen alan tutkimusta on nimitetty suomeksi 'survey-tutkimukseksi', jonka käypiä vastineita ovat ainakin kysely- tai lomaketutkimus. Nämä termit ovat suositeltavia myös siksi, että sosiaalitutkimuksessa 'haastattelu'-termi yhdistetään nykyisin vahvasti laadullisiin tutkimusmenetelmiin.

Yhteiskunta- ja käyttäytymistieteellisessä tutkimuksessa toteutetaan strukturoituja kyselyjä monin eri tekniikoin. Niistä yleisimpiä ovat käynti-, puhelin-, posti- ja Internet-kyselyt. Usein tekniikoita myös yhdistellään vastaajien ajan ja tiedonkeruun kustannusten säästämiseksi. Yksi metodologisesti tärkeä ero koskee sitä, vastataanko kysymyksiin konkreettisessa haastattelijan ja haastateltavan välisessä vuorovaikutustilanteessa (käyntikyselyt, muut "face-to-face" -kyselyt ja esimerkiksi puhelinkyselyt) vain antaako vastaaja kyselyvastauksensa omatoimisesti (*self completion*).

Tiedonkeruutekniikka vaikuttaa paljon siihen, millaisia kysymyksiä kyselyssä voidaan esittää ja millaiset tekijät vaikuttavat vastauksiin ja tutkimustulosten luotettavuuteen. Haastattelijan ja haastateltavan vuorovaikutuksesta on etua tietopohjaisia kysymyksiä esitettäessä mutta kontakti saattaa vääristää esimerkiksi arkaluonteisiin kysymyksiin annettuja vastauksia. Tällaisia voivat olla alkoholin käyttöön, terveystietoihin tai seksuaaliseen käyttäytymiseen liittyvät aiheet. Käynti- ja puhelinkyselyjä kustannuksiltaan edullisemmissa posti- ja Internet-kyselyissä epävarmuustekijät liittyvät muun muassa siihen, että haastattelija ei ole avustamassa ja valvomassa vastaamista. Joihinkin kysymyksiin ei tällöin ehkä osata vastata teknisesti oikealla tavalla, niihin voidaan jättää vastaamatta kokonaan tai saatetaan valita herkästi 'en osaa sanoa' -vaihtoehto. Aina vastaajana ei myöskään täysin varmasti ole tarkoitettu henkilö.

Menetelmäopetuksen tietovarannossa ei toistaiseksi tarkastella vertailevasti erilaisille tiedonkeruutekniikoille ominaisia ratkaisuja kysymysten ja vastausvaihtoehtojen laadinnassa. Koska tietovaranto on suunnattu erityisesti opiskelijoille, on tarkoituksenmukaista keskittyä edullisimpiin ja tekniseltä toteutukseltaan yksinkertaisimpiin tiedonkeruutapoihin. Tästä syystä tietovaranto keskittyy kyselyjen toteuttamista ja lomakesuunnittelua koskevissa osuuksissa ja niiden esimerkeissä lähinnä postikyselyihin. Niitä koskeva ohjeistus on kuitenkin sovellettavissa varsin yleisesti muihinkin kyselyjen toteuttamistekniikoihin.

Tutkimusongelmien hahmottaminen ja esitutkimus

Kuten mikä tahansa tieteellinen tutkimus, postikyselytutkimuskin alkaa aikaisempaan tutkimukseen perehtymisellä. Alaan liittyvien julkaisujen ohella kannattaa perehtyä tutkimuksissa käytettyihin aineistoihin niiltä osin kuin tietoa on saatavissa. Muun muassa Yhteiskuntatieteellisen tietoarkiston verkossa olevat aineistokuvaukset sisältävät runsaasti linkkejä arkistoitujen data-aineistojen kyselylomakkeisiin.

Aiheeseen perehtyminen auttaa hahmottamaan ja kehittämään tutkimuksen ongelmanasettelua sekä asetettavien kysymysten ratkaisun edellyttämiä operationalisointeja. Mikäli tutkimuksen pääkysymyksiin vastaaminen edellyttää uuden kyselyaineiston keräämistä, on määritettävä aineiston perusjoukko, havaintoyksikkö ja mahdolliset otokseen tai näytteeseen

liittyvät yksityiskohdat. Edelleen on eriteltävä tutkimusongelmiin liittyviä selitettäviä ja selittäviä tekijöitä siten, että ne voidaan ottaa kattavasti huomioon tutkimuslomakkeen varsinaisten sisältökysymysten ja vastaajien taustatietoja koskevien muuttujien laatimisessa. Etenkin uusien tutkimuskysymysten kohdalla tämä saattaa vaatia pitkäaikaistakin esitutkimusta aikaisempaan sivuavaan tutkimukseen perehtymisen ohella.

Kyselyn esitutkimusvaiheen voi toteuttaa hyvin eri tavoin. Yleensä on suositeltavaa on kerätä jonkinlainen esiaineisto testaamalla ideoita kyselylomaketta varten. Tällöin on usein paikallaan esittää myös avoimia kysymyksiä niistä aihealueista. joita tutkimuksen on määrä koskea. Luokittelemalla esitutkimuksesta saatuja vastauksia tutkimuksen tekijä voi arvioida erilaisten kysymysrakenteiden mahdollisuuksia ja vastausvaihtoehtojen alaa. (Vrt. avointen kysymysten käsittely myöhemmin tässä kappaleessa.) Esitutkimusvaiheessa saatu palaute voi myös nostaa havaittujen virheiden lisäksi esiin joitakin tutkijalta unohtuneita tärkeitä kysymyksiä ja aihealueita.

Kyselylomakkeen laatiminen ja viimeistely

Riittävän aihepiiriin perehtymisen jälkeen alkaa kyselylomakkeen luonnostelu versio version jälkeen. Tarvetta suunnittelun huolellisuuteen ja kiireettömyyteen ei voine ylikorostaa: kyselytutkimuksissa lomakkeen suunnittelu on korvaamattomin osa koko tutkimusprosessia. On aivan tavanomaista, että asiantuntijoidenkin suulla valmiiksi arveltua lomaketta joudutaan vielä "viilaamaan". Mainituista syistä menetelmäopetuksen tietovaranto käsittelee kyselylomakkeen suunnittelua laajasti erillisenä kokonaisuutena.

Ennen lomakkeen monistamista on suositeltavaa antaa lomakkeen eri versioita luettavaksi tutkimuksen ohjaajille tai muille alaa tunteville henkilöille. Lisäksi lopulliseksi arvioitu lomake kannattaa täyttää itse ja antaa se vielä pienen koevastaajajoukon vastattavaksi, ja tehdä vielä tarpeen vaatimat viimeiset muutokset ja täydennykset.

Saatteiden laatiminen ja vastausprosentti

Kyselytutkimuksiin liittyvien saatekirjeiden laatiminen voi pikaisesti ajatellen tuntua välttämättömältä pahalta silloin kun kysymyslomake on vihdoin saatu valmiiksi ja on yleensä kiire lähettää se "kentälle". Erilliselle paperille tai kyselylomakkeeseen liitetty saate on kuitenkin suunnattoman tärkeä dokumentti tutkimuksen onnistumisen ja myös aineiston mahdollisen uudiskäytön kannalta.

Jäljempänä esitettävien saatteiden laatimisen muistilistojen huomioon ottaminen vaikuttaa suoraan kyselyjen vastausprosentteihin. Monet kohdat liittyvät vastaajien motivointiin, mutta joukossa on myös lainsäädännön kannalta tärkeitä näkökohtia. Tietojen kerääjä on nimittäin velvollinen selittämään tutkimuksen kohteelle syyt tietojen keräämiseen. Lisäksi hänen on selvitettävä kokoamiensa tietojen käyttötarkoitus.

Tieteellinen kyselytutkimus tuleekin toteuttaa aineiston suunnittelusta tulosten raportointiin ja aineiston säilyttämiseen saakka tutkimuseettisesti kestävällä tavalla. Tämä edellyttää

- täsmällisesti muotoiltua ja tutkimuseettisesti hyväksyttävää tutkimus- ja aineistonkäyttöstrategiaa (tietojen ja aineiston käyttötarkoitus),
- kustannustehokkuutta (kerätään järkevästi vain tarvittavat tiedot mahdollisimman tehokkaasti ja pienin kustannuksin) sekä
- tietojen luovuttajien riittävää informointia ja motivointia, jotta suuren henkilömäärän oikeudet turvataan eikä vastaajia vaivata turhaan.

Tärkeiden periaatteiden soveltaminen käytäntöön ei suinkaan ole aina helppoa. Kyselyn huolellisesta suunnittelusta huolimatta suuri osa otoksiin kuuluvista henkilöistä ei syystä tai toisesta halua tai ehdi osallistua tutkimuksiin.

Postikyselyissä joudutaan normaalisti lähettämään myös vastausmuistutuksia ja/tai ns. "karhulomakkeet" otokseen kuuluville henkilöille, koska ensimmäiseen kyselykierrokseen vastanneiden määrä ei tavallisesti kohoa tarpeeksi suureksi. Kohtuullinen tai tyydyttävä vastausprosentti riippuu paljolti vastaajajoukosta ja kyselyn aihepiiristä. Näin ollen ei liene mahdollista määrittää yleispätevästi 'riittävää' postikyselyn vastausprosenttia. Valtakunnallisissa aikuisväestön satunnaisotoksiin liittyvissä postikyselyissä joudutaan nykyisin tyytymään vain hieman 50 prosentin yläpuolelle nouseviin vastausprosentteihin.

Saatekirjeiden sisältöön, ulkoasuun ja kieleen kannattaa siis kiinnittää erityistä huomiota. Saatteen tulee herättää luottamusta ja vastausmotivaatiota. Lisäksi sen pitää selvittää ainakin seuraavat asiat:

Ensimmäisen kyselykierroksen saate:

a. mikä kysely/tutkimus
b. kuka tekee tutkimuksen, kuka teettää (jos teettäjä), keihin kysely kohdistuu (ei välttämättä kannata mainita, kuinka moneen henkilöön kohdistuu)
c. tutkimuksen tarpeellisuuden perustelu
d. maininta tutkimustulosten ja -aineiston käytöstä sekä vastaajien anonymiteetin säilymisestä
e. jokaisen vastaajan vastausten tarpeellisuus tutkimuksen onnistumiseksi
f. milloin lomake on viimeistään palautettava takaisin (ei 1-2 viikkoa pidempää vastausaikaa lomakkeen saamisesta, ellei erityisen painavaa syytä)
g. etukäteiskiitokset vastauksista/yhteistyöstä
h. tekijän ja teettäjän edustajan nimet ja allekirjoitukset (opinnäytteissä käytetään teettäjän edustajana usein työn ohjaajan nimeä)

Muistutuskierroksen saate:

a. mikä ja kenen tutkimus, milloin edellinen lähetys lähetettiin
b. miksi lähetetään muistutuskortti vastaamattomille tai kokonainen muistutuskierros: (aina ei ole tiedossa, ketkä ovat vastanneet ja ketkä eivät)
c. hyvin näkyvä maininta siitä, että kyselyyn jo vastanneiden ei tarvitse enää vastata uudelleen
d. vetoomus vastaamisen ja kyselyn onnistumisen puolesta
e. vastausten viimeinen palautuspäivämäärä väljästi määriteltynä (esim. viikon kuluessa)
f. mahdollisesti uudestaan tekijöiden ja teettäjien nimet ja allekirjoitukset; lisäksi mahdollisia uusia suosituksia ja suosittelijoita

Muistutuskierros tulee toteuttaa mahdollisimman pian ensimmäisen vastauskierroksen vastausajan umpeuduttua. Joissakin tapauksissa karhukierroksia toteutetaan useampia kuin yksi, mutta mikäli muistutukset menevät lomakkeineen kaikille, kustannus-/hyötysuhde voi jäädä pieneksi.

Todettakoon lisäksi, että joissakin kyselyissä vastaajien tunnistamattomuutta ei ole tarpeen suojella niin voimakkaasti, että ensimmäisen kyselykierroksen jälkeen ei tiedetä kuka on jo vastannut kyselyyn. Näin voi olla tapauksissa, joissa kerättävät tiedot eivät ole mitenkään arkaluonteisia, tietojen kerääjä ja luovuttajat tuntevat toisensa, vastaaminen on esim. sopimuksiin perustuvaa "virkatyötä", vastaajat toimivat julkisissa tehtävissä jne. Mikäli tieto jo vastanneista on käytössä, se tietenkin säästää työtä ja muistutuskierroksen kustannuksia, kun jo vastanneille ei tarvitse lähettää uudestaan vastausmateriaalia.

Postikyselyn painotyöt, lähettäminen ja karhuaminen

Lopullisen lomakkeen ja saatteen valmistuttua alkaa paino- ja lähetysvaihe. Vähänkään suuremmissa, ainakin tuhansien vastaajien, tutkimuksissa monistamiseen kannattaa käyttää kirjapainopalveluja. Muistutuskierrokseen varautumisen vuoksi lomakkeita, lähetyskirjekuoria ja palautuskirjekuoria painetaan tavallisesti, edellä mainittuja esimerkkejä lukuun ottamatta, ainakin kaksi kertaa niin paljon kuin otokseen kuuluu vastaajia. Ensimmäisellä keruukierroksella mahdollisesti tarvittavaa erillistä saatetta painetaan yhteen keruukertaan riittävä määrä, sillä karhu- eli muistutuskierrokselle on tehtävä oma saate, muistutuskortti tai vastaava. Kirjekuorien osalta kannattaa ottaa selvää postin lähettämistä ja palauttamista helpottavista painatus- ja lähetysmahdollisuuksista välittömien tutkimuskustannusten ja tutkijan työajan minimoimiseksi.

Vastaajille on annettava mahdollisuus vastauslomakkeen maksuttomaan palauttamiseen lähetyskirjeeseen liitettävällä kirjekuorella. Yhteiskuntatieteellisissä tutkimuksissa ei yleensä käytetä vastauspalkkioita kuten arvontoja vastanneiden kesken, koska ne saattavat vinouttaa otoksen rakennetta.

Aineiston saattaminen käyttökuntoon

Yleistä lomakkeiden koodauksesta

Lomakkeiden palautuksen ja karhuamisen jälkeen palautetut lomakkeet käydään läpi ja joukosta poistetaan kokonaan tyhjät tai liian puutteellisesti täytetyt lomakkeet. Tässä vaiheessa on syytä pitää kirjaa hyväksymisen ja hylkäämisen kriteereistä ja määristä, koska niitä koskevia tietoja tarvitaan tutkimuksen raportoinnissa käsiteltäessä otoksen rakennetta ja kato-ongelmia.

Aineiston ulkopuolelle jättämisen kriteereitä on mahdotonta määritellä yleispätevästi, koska jo muutamiin kysymyksiin vastaaminen saattaa joskus olla, aineiston koosta riippuen, vahva aineistoon mukaan ottamisen peruste. (Sinänsä joidenkin tyhjähköjen lomakkeiden puuttuvien tietojen tallentaminen havaintomatriisiin puuttuviksi tiedoiksi ei ole ongelma, koska toimenpide ei muuta tutkimuksen tuloksia tai johtopäätöksiä.)

Tallennettavaksi hyväksyttyihin lomakkeisiin on ehdottomasti merkittävä juokseva numerointi esimerkiksi etusivun yläreunaan. Käsin merkitty tai tarkoitukseen soveltuvalla leimasimella tehty lomakenumero tallennetaan myöhemmin havaintomatriisiin havaintoyksiköt toisistaan erottavaksi tunnistemuuttujaksi. Sen avulla aineiston havaintoyksiköt voidaan myöhemmin yhdistää lomakkeisiin, mikä on välttämätöntä mm. virheellisesti syötettyjen tietojen korjaamiseksi. Lisäksi havaintoyksikköjen lomakkeet identifioiva muuttuja mahdollistaa erilaisten lisätietojen liittämisen havaintomatriisiin jälkikäteen.

Lomakkeiden numeroinnin jälkeen seuraa varsinainen koodausvaihe, jonka kesto riippuu paljon koodausta vaativien muuttujien määrästä ja laadusta. Lisäksi asiaan vaikuttaa se, onko osa koodausvaiheeseen normaalisti kuuluvasta tarkistustyöstä siirrettävissä yhtäaikaisesti tallennuksen kanssa suoritettavaksi. Jos aineiston tallentaja on kokematon tai kyseessä on pienikokoinen aineisto, on kaikkien lomakkeiden kaikki vastausmerkinnät yleensä syytä tarkistaa ja tarpeen vaatiessa koodata etukäteen ennen tallennusta. Suurissa aineistoissa tai kokeneiden tallentajien tapauksessa strukturoitujen kysymysten vastausmerkintöjä ei välttämättä tarvitse tarkistaa ennen tallennusta.

Viimeistään tallennusvaiheessa on kuitenkin siis päätettävä matriisiin tallennettavasta informaatiosta. Opiskelijan on suositeltavaa tarkistaa ennen tallennusta kaikki vastaukset ja koodata merkinnät vastausohjeiden mukaisiksi. Usein tietokoneohjelmissa on mahdollista etukäteen määritellä hyväksyttävien koodien joukko, mikä pienentää koodauksesta mahdollisesti seuraavaa virhettä.

Jos aineiston tallennuksessa käytetään yleensä samaa puuttuvan tiedon koodia, esimerkiksi nollaa, on ainakin tästä linjasta poikkeavat puuttuvan tiedon koodit syytä kirjoittaa lomakkeisiin.

Kaikissa kysymyksissähän nollaa ei välttämättä voi käyttää puuttuvan tiedon merkkinä, koska se voi olla jonkin muuttujan validi arvo.

Työllistävin luokittelu ja koodaus liittyy usein avoimiin kysymyksiin annettujen vastausten muuntamiseen numerokoodeiksi. Avoimiin kysymyksiin annetut vastaukset voidaan tallentaa myös tekstimuotoisesti, mutta mm. tietosuojanäkökohtien kannalta voi olla tarkoituksenmukaista koodata ja tallentaa avoimet kysymykset numerokoodein.

Avointen kysymysten koodaus

Avoimiin kysymyksiin annetut vastaukset saattavat olla useista virkkeistä koostuvia tarinoita, ranskalaisille viivoille tiivistettyjä vastauksia tai vain tärkeintä asiaa kuvaavia yksittäisiä sanoja. Vastaukset ovat yleensä myös sisällöltään hyvin kirjavia eikä niiden luokitteleminen numerokoodausta varten ole useinkaan helppoa. Lomaketutkimuksissa tutkijan tehtävänä on kuitenkin "pakottaa" vastauksia erikseen päätettäviin sisältöluokkiin. Luokitus voi olla ennalta määrätty, mutta tavanomaisinta on laatia luokitus avoimeen kysymykseen saatujen vastausten pohjalta.

Tällöin on ensiksi muodostettava vastauksiin sopiva sisältöluokitus, jonka jälkeen vastaukset voidaan koodata sen mukaan. Tämän ns. luokitusrungon vaihtoehdot numeroidaan juoksevalla numerolla. Käytännössä luokitus muodostetaan siten, että aluksi kirjataan lomakkeista yksittäisiä vastauksia ja hahmotellaan vähitellen niiden pohjalta vastausluokkia. Vastausten sisältöä voi pyrkiä jakamaan eri luokkiin käyttäen apuna esimerkiksi tukkimiehen kirjanpitoa. Työtä jatketaan niin kauan kunnes uudentyyppisiä vastauksia ja tarvetta uusiin vastausluokkiin ei enää kerry merkittävässä määrin.

Sisältöluokitusta laadittaessa on hyvä pitää mielessä, että kyse ei ole tutkimuksessa käytettävästä lopullisesta luokittelusta ja että yhtä avointa kysymystä kohden voidaan luoda ja koodata useita vastausten sisältöä kuvaavia muuttujia.

Muun muassa koodaustapojen vaihtelevuuden vuoksi avoimen kysymyksen sisältöluokitukselle ei ole mahdollista antaa yleispäteviä onnistuneisuuskriteereitä. Laatu riippuu sekä asiasisällön rakenteesta että tutkijan tavoitteista luokituksen suhteen. Luokituksille ja koodaustyölle on silti mahdollista asettaa käytännön kokemuksen kautta joitakin perusperiaatteita.

Näistä tärkein liittyy tietojen yksityiskohtaisen tallentamisen yleisperiaatteeseen. Sekä aineiston kerääjä että sen mahdollinen uudiskäyttäjä saattavat myöhemmin käyttää luokiteltavia tietoja muihinkin kuin tiedonkeruuvaiheessa tunnistettuihin käyttötarkoituksiin. Tästä syystä sisältöluokituksen ja sen mukaisen koodauksen tulee olla riittävän hienojakoinen. Yksityiskohtaisia tietoja on sitten mahdollista myöhemmin luokitella eri tavoin vaihteleviin tarkoituksiin.

Toisaalta melkein jokaiselle asialle oman koodin antaminen johtaa siihen, että monet vastauskategoriat keräävät hyvin pieniä vastausosuuksia. Kooltaan täysin mitättömiä vastausluokkia kannattaa välttää, ellei niiden käyttöön ole painavia sisällöllisiä syitä.

Koodausteknisistä syistä luokitusrungon viimeisenä vastausluokkana kannattaa käyttää 'jokin muu' -ryhmää. Siihen koodattavien vastausten osuuden ei tulisi nousta liian suureksi (esim. korkeintaan 10-20 %).

Lopuksi on päätettävä kutakin avointa kysymystä varten koodattavien muuttujien lukumäärästä. Siihenkin on mahdotonta antaa yleispäteviä sääntöjä. Yhden avoimen kysymyksen koodaukseen kannattaa käyttää useita muuttujia, jos suuri osa vastauksista näyttää sisältävän monia aspekteja.

Laajimmassa koodaustavassa kunkin vastausluokan voi koodata omaksi muuttujakseen, jolloin muuttujaan koodataan dikotominen tieto siitä, mainitsiko vastaaja asian vastauksessaan vai ei (koodataan esimerkiksi 0=puuttuva tieto, 1=kyllä ja 2=ei). Tällaisen koodaustavan käyttö on perusteltua lähinnä silloin, kun vastausten sisältöalue on suppea.

Usein on tarkoituksenmukaisempaa koodata vastaukset esim. yhdestä kolmeen muuttujaan, joihin vastaukset tallennetaan luokitusrungon luokkia vastaavin numeroin. Vastausta kuvaavat

numerokoodit merkitään tavallisesti lomakkeen jompaan kumpaan reunaan. Huomattakoon, että ns. tyhjät vastaukset tulee merkitä puuttuvan tiedon koodilla.

Käytännössä monet eivät vastaa avoimiin kysymyksiin lainkaan. Myös monissa avoimiin kysymyksiin vastanneiden lomakkeissa saadaan jotakin luokiteltua vastaussisältöä vastaava koodi vain osaan kysymyksen koodaukseen varatuista muuttujista. Tulosten raportointivaiheessa käytetään usein vain ensimmäiseen muuttujaan koodattuja vastauksia, mutta kaikista kysymystä varten koodatuista muuttujista saadut tiedot voidaan myös yhdistää yhdeksi kokonaisjakaumaksi. Tämä voidaan toteuttaa tilasto-ohjelmistojen soveltuvilla komennoilla.

Vastaajien antamien tietojen tarkistus ja korjaus

Tutkijan on tarpeellista kaikin puolin varmistua lomakkeiden täytön ja koodauksen moitteettomuudesta, jotta tallennusvaihe sujuu kitkatta. Nämä rutiinit ovat aina tärkeitä tutkimuksen reliabiliteetin kohottamiseksi. Erityisen tärkeitä ne ovat pienehköissä kyselyissä.

Parhaatkaan vastausohjeet eivät nimittäin poista sitä ongelmaa, että osa vastaajista vastaa lomakkeen kysymyksiin teknisesti väärällä tavalla. Vaikka heitä on ohjattu rengastamaan mieleisiään vaihtoehtoja vastaavat numerot, jotkut käyttävät tästä huolimatta "rukseja" tai ympyröivät vastausvaihtoehdon tekstin. Tämänkaltaiset virheet eivät ole tutkimuksen kannalta ongelmallisia, joskin ne saattavat lisätä tallennusvaiheen virheitä. Hankalampaa on sen sijaan koodata teknisesti oikeaan muotoon sellaisia vastauksia, joissa annettu vastaus ei lainkaan vastaa annettuja ohjeita tai vaihtoehtoja. Joskus kyselyn laatija jopa jää osaavan vastaajan koukkuun ja huomaa laatineensa selvästi epäonnistuneen kysymyksen.

Useissa tapauksissa sellaisenaan tallennettavaksi kelpaamattomat vastaukset voidaan kuitenkin koodata tallennukseen kelpaaviksi muuttamalla niitä suurimman sisällöllisen hyödyn ja varsinaisten sisältövaihtoehtojen tasapuolisen kohtelun periaatteiden mukaisesti. Esimerkiksi strukturoiduissa kysymyksissä vastaukset sijoitetaan silloin tällöin kahden sijoittumista aidosti kuvaavan vaihtoehdon väliin. Tällaiset vastaukset voitaneen sijoittaa vahvempaa mielipidettä/toimintaa/tms. kuvaavaan luokkaan, koska puuttuvan tiedon koodaaminen muuttaisi tässä kohden enemmän vastausta. Tällaiset tapaukset ovat kuitenkin aika harvinaisia.

Toinen tavallinen ongelma koskee monivastauskysymyksiä, joissa vastaajia on pyydetty nimeämään korkeintaan niin ja niin monta kohtaa tai esimerkiksi asettamaan joitakin asioita tai ominaisuuksia tärkeysjärjestykseen. Jos vastauksia on liikaa, yksi hyvä tapa on esimerkiksi tallentaa tällaiset muuttujat ylhäältä alas parillisella lomakenumeroilla, ja alhaalta ylös parittomalla lomakenumerolla. Tällöin lomakkeessa viimeisinä mainitut kohdat eivät kärsi suhteettomasti sijoittumisestaan listan loppupäähän.

Järjestyssijoja hyödyntävissä monivalintakysymyksissä on ongelmana se, että pieni osa vastaajista ei noudata vastausohjeita tarkkaan, vaan mainitsee vain esim. kolme tärkeintä asiaa, mutta ei aseta niitä tärkeysjärjestykseen. Tällaisessa tapauksessa kaikki kohdat voinee koodata niihin ykkössijaa vastaavalla koodilla tai arvotuilla järjestysnumeroilla, sen sijaan että kaikki tiedot koodattaisiin puuttuvan tiedon koodilla.

Havaintoaineiston tarkistus, varmuuskopiointi ja arkistointi

Toisaalla käsiteltävän aineiston syöttämisen ja tallentamisen jälkeen havaintomatriisin sisältämät tiedot on vielä syytä tarkistaa. Näitä rutiineja esitellään erikseen tietovarannon SPSS-osiossa. Tarkistetusta aineistosta tulee ottaa riittävästi varmuuskopiota, mielellään erilaisille tallennusvälineille. Sama pätee aineistosta tuotettaviin uusiin versioihin, jotka sisältävät aineistosta tuotettuja uusia muuttujia.

Erityisen tärkeää on tallentaa kaikki aineistonkeruuvaiheen keskeiset sähköiset dokumentit (data-aineisto, kyselylomakkeiden ja saatteiden tekstitiedostot yms.) kootusti johonkin hakemistoon/kansioon ja ottaa siitä varmuuskopiot. Tämä palvelee paitsi tutkijaa itseään myös aineiston mahdollista arkistointia myöhemmässä vaiheessa.

Kyselylomakkeen laatiminen

Tähän artikkeliin liittyy useita käytännön esimerkkejä mm. annettujen ohjeiden ja suositusten soveltamisesta. Niihin voi tutustua kootusti alaluvussa lisäesimerkit.

Tieteellisen kyselyn onnistuminen edellyttää, että tutkija osaa ottaa laaja-alaisesti huomioon vastaajien ajan, halun ja taidot vastata kyselyyn. Lomakkeen huolellinen suunnittelu ja testaaminen vaikuttavat ratkaisevasti tutkimuksen onnistumiseen, mutta hyvä lomake ei suinkaan yksin riitä. On kiinnitettävä huomiota myös muihin kyselyn toteuttamiseen liittyviin seikkoihin.

Lomakesuunnittelussa on otettava huomioon monia asioita. Seuraavaan on koottu erilaisia lomakkeen sisältöä ja ulkoasua koskevia vinkkejä. Monet niistä soveltuvat yleisemminkin eri tiedonkeruutavoilla toteutettaviin kyselyihin. Tämä aihealueittainen kokoelma huomionarvoisista seikoista on laadittu etenkin postikyselylomakkeita silmällä pitäen. Monet vinkeistä soveltuvat hyvin myös muihin vastaajien itse täyttämiin lomakkeisiin (verkkokyselyt) sekä haastattelijan täytettäväksi tarkoitettujen lomakkeiden suunnitteluun (käynti- ja puhelinkyselyt).

Lomakkeen laajuus ja ulkoasu

Lomakkeen kohtuullinen pituus ja ulkoasun selkeys ovat erittäin tärkeitä sekä vastaajalle että myöhemmin tietojen tallentajalle. Ylipitkä kysely karkottaa vastaamishalun. Postikyselyissä keskimääräisen vastausajan ei tulisi ylittää 15-20 minuuttia.

Lomakkeen suunnittelussa huomioitavien seikkojen listaus alkaa aivan tarkoituksella lomakkeen pituudesta ja ulkoasusta. Tämä tuntuu ehkä provosoivalta tieteellisen tutkimuksen näkökulmasta, mutta vastaamispäätökset perustuvat postikyselyissäkin paljolti ensivaikutelmaan vastaanotetusta materiaalista. Siihen vaikuttaa saatteen lisäksi ratkaisevasti lomakkeen yleisilme. Huono vastausprosentti voi pilata loistavasti suunnitellun aineiston.

Jotta vastaaja ja tiedon tallentaja huomaavat kaikki kysymykset, on lomakkeen taitto syytä tehdä pääsääntöisesti siten, että kysymykset etenevät ylhäältä alaspäin. Mikäli tästä tehdään poikkeuksia, voi apuna käyttää samantapaisia nuolia kuin sarjakuvissa käytetään.

Yleinen selkeysvaatimus ei saa johtaa siihen, että lomakkeesta tulee suurella kirjasimella kirjoitettu harvarivinen moniste. Kannattaa pyrkiä tiiviiseen ja pienehköllä, mutta selkeällä kirjasimella tehtyyn lomakkeeseen. Kysymykset pitää erottaa toisistaan selkeästi, esimerkiksi viivoin tai laatikoimalla.

Myös palstoittaminen säästää tilaa ja saa lomakkeen näyttämään ja tuntumaan lyhyemmältä. Internetissä on linkkejä lukuisiin kyselylomakkeisiin. Edellä mainituista asioista löytää esimerkkejä ainakin Yhteiskuntatieteellisen tietoarkiston arkistoimien aineistojen sivustoilta, joissa on useimmiten linkit myös tutkimusten kyselylomakkeisiin.

Esimerkissä 1 on malleja lomakkeen laajuudesta ja ulkoasusta.

Luottamuksen herättäminen ja vastaajien ominaisuuksien huomioon ottaminen

Lomaketutkimuksissa on pyrittävä tutkimusongelman kannalta kattavaan, mutta samalla yksinkertaiseen ja helppotajuiseen kysymyksenasetteluun. Varsinkaan tieteellisessä kyselyssä ei pidä harrastaa varmuuden vuoksi kysymistä. On hyvä muistaa, että vastaajajoukko tuntee vain harvoin tutkittavan aihealueen yhtä hyvin kuin kysymysten laatija.

Lomakkeen potentiaalisten palauttajien täytyy paitsi jaksaa, myös osata vastata kyselyyn. Standardoiduissa kyselyissä vastaajien tulee ymmärtää kysymykset mahdollisimman samalla tavalla ja myös vastata niihin yhteismitallisin arviointiperustein. Tämä edellyttää kauttaaltaan yksinkertaista, tarkoituksenmukaista ja täsmällistä kieltä kysymysten laadinnassa.

Yksinkertaisuuden vaatimus koskee myös kysymysten pituutta: hyvä kysymys on aina kohtuumittainen.

Esimerkissä 2 on malli vastaamisen kannalta epäselvästä kysymyksestä.

Myös kyselyn kohdejoukkoon kuuluvat kielivähemmistöt tulee ottaa käyttettävissä olevien resurssien puitteissa huomioon. Suomen koko aikuisväestöön kohdistuvissa satunnaisotoksiin perustuvissa valtakunnallisissa kyselyissä on erittäin suositeltavaa kääntää kyselylomake myös ruotsiksi.

Tietosuojanäkökohtia ja vastaamishalua silmällä pitäen kysymyslomake on laadittava siten, ettei vastaajien tarvitse huolehtia antamiensa tietojen väärinkäyttömahdollisuuksista. Satunnaisotokseen kuuluville henkilöille lähetettäviin lomakkeisiin ei pidä merkitä epäilyjä herättäviä identifikaatiotunnuksia. Tiettyjä poikkeuksia voi olla, kuten julkisissa tehtävissä toimivia vastaajaryhmiä tai vastaajaryhmiä, joiden edustajien kanssa asiasta on sovittu etukäteen. Vastaajan anonymiteetin säilyminen tulee jatkuvasti ottaa huomioon myös kysymysten laadinnassa. Lisäksi vastaajien taustatietojen kartoittamisen alussa on hyvä mainita, että taustatietoja tiedustellaan vastausten tilastollista käsittelyä varten.

Luottamuksen herättämistä ja vastaajaa kohtaan tunnetun arvostuksen osoittamista voi olla sekin, että kyselyn laatija teitittelee vastaajaa kautta lomakkeen. Valinta teitittelyn ja sinuttelun välillä on kuitenkin tehtävä aineistokohtaisesti lähinnä vastaajaryhmän ominaisuuksien perusteella. Myös kyselyn yleinen luonne ja mahdollisen teettäjän suhde vastaajaryhmään vaikuttaa asiaan. Sinutteleva kieli on yleistymässä ja aikuisväestöstä poimittuihin satunnaisotoksiin kohdistuvissa kyselyissä käytetään jo silloin tällöin sinuttelumuotoa. Kumpi tahansa käy, kunhan vain kysymysten laatija on linjasssaan johdomukainen.

Esimerkissä 3 on malli sinuttelevista ja teitittelevistä lomakkeista.

Lomakkeen kokonaisrakenne ja sisällön loogisuus

Yleensä lomake kannattaa aloittaa kysymyksillä, joihin on varmasti helppoa vastata. Selittävinä muuttujina käytettävät ns. taustakysymykset saattaa kannattaa jättää joko kokonaan tai ainakin pääosin kyselyn loppuun, koska niiden kysyminen laajasti heti alussa voi ehkä herättää negatiivisia tuntemuksia vastaajassa (anonymiteetti).

Kyselyyn on helpompaa vastata, kun kysymykset ovat loogisessa järjestyksessä. Sama lomake voi sisältää sisällöllisesti hyvinkin erilaisia asioita, mutta samaan asiaan liittyvät kysymykset on sijoitettava loogiseen järjestykseen peräkkäin. Sama koskee aihealueesta toiseen siirtymistä.

Kysymysten onnistuneisuus - tasapainoisuus ja sisällöllinen kattavuus sekä yleinen selkeys - ovat niin ikään erittäin tärkeitä sisällön jäsentyneisyyden kannalta. Näihin kysymyksiin palataan myöhemmin tässä artikkelissa.

Kysymyksenasettelun tarkkuustaso ja avointen kysymysten harkittu käyttö

Pääsääntö on, että kaikkea kysytään kohtuullisen tarkasti. Analyysivaiheessa liian hienojakoiseksi havaittua informaatiota on helppo tiivistää. Karkeajakoisesti kerättyjä vastauksia ei sitä vastoin voi enää muuttaa hienojakoisemmiksi. Muun muassa vastaajien ikää ei ilman erityisen painavia syitä pidä kysyä luokiteltuna. Mieluummin kannattaa kysyä syntymävuotta (tai ikää) vuoden tarkkuudella, jolloin muuttuja sallii tarpeen vaatimat ikäluokitukset analyysivaiheessa. Toisaalta kysymysten ja vastausvaihtoehtojen liiallinen spesifisyys voi tuottaa näennäistä mittaustarkkuutta, jos esimerkiksi tiedustellaan hankalia muistinvaraisia asioita liian tiuhalla seulalla.

Esimerkissä 4 on mittaamisen näennäistarkkuudesta.

Kysymysten tarkkuustasoon liittyvistä kysymyksistä yksi tavanomaisin koskee sitä, laaditaanko kysymykseen valmiit vastausvaihtoehdot (strukturoitu kysymys) vai riittääkö avoin

kysymys. Täysin avoimia kysymyksiä on suositeltavaa sisällyttää lomakkeeseen harkiten ainoastaan silloin kun niiden käyttöön on painava syy. Postikyselyjen kaikki vastaajat eivät vastaa niihin ja vastaustavatkin vaihtelevat paljon, eikä vastauksista saatu informaatio aina täytä tutkijan odotuksia. Mutta jos vastaajajoukko tiedetään aktiiviseksi ja helposti myös kirjallisesti kantaa ottavaksi, avointen kysymysten käyttö voi olla hyvinkin perusteltua.

Esitutkimusvaiheessa avoimia kysymyksiä kannattaa käyttää aihepiirin eri ulottuvuuksien kartoittamiseksi. Monesti tästä kannattaa edetä strukturoituun kysymyksenasetteluun lopullisessa tutkimuslomakkeessa.

HUOM 1. Joitakin kvantitatiivisia muuttujiakin kannattaa kysyä avointa kysymystä muistuttavalla tavalla, jos vastaajan voi olettaa kykenevän vastaamaan tarkalla määrällä eikä kyse ole kovin usein toistuvasta toiminnasta (esim. montako kertaa olette osallistuneet ...)

HUOM 2. Avoimiin kysymyksiin annetut vastaukset voidaan tallentaa datatiedostoon kirjoitettuina tai numeroiksi koodattuina. Koodaus on yleensä tarpeen suorittaa ennen aineiston tallentamista. Avointen kysymysten koodaustapoja käsitellään toisessa tietovarannon artikkelissa.

Esimerkissä 5 on malleja avoimista ja puoliavoimista kysymyksistä.

Vastausohjeet

Lomakkeeseen kannattaa aina merkitä mahdollisimman yksityiskohtaisia vastausohjeita. Niitä kannattaa käyttää sekä yksittäisten kysymysten lopussa että lomakkeen alussa, jossa pitää ainakin ilmoittaa seuraavaa ohjetta vastaava sisältö: "Ellei toisin mainita, rengastakaa oikeaa vaihtoehtoa vastaava numero, tai kirjoittakaa vastauksenne sille varattuun tilaan."

Lomakkeen kysymyksiin voi kuulua ja usein kannattaakin sisällyttää sekä varsinainen kysymys että vastausohje. Kyselyn alussa oleva yleinen vastausohje ei aina riitä, jos kysymys rakenteensa puolesta vaatii lisäohjeita teknisesti oikean vastaamisen turvaamiseksi. Kysymysten edetessä itseään toistavia vastausohjeita voi jättää pois, mikäli vastaajien voi olettaa jo oppineen vastaustavan.

Kaikkia kysymyksiä ei myöskään voi tai ole pakko kysyä kaikilta vastaajilta. Tällöin kysymykseen vastaava joukko on ilmoitettava selkeästi lomakkeeseen tehtävillä merkinnöillä ja sanallisilla huomautuksilla. Lisäksi on opastettava seikkaperäisesti, mihin lomakkeen kohtaan kysymyksen sivuuttavat vastaajat siirtyvät.

Kyselyn yleisohjeiden, kysymyskohtaisten teknisten vastausohjeiden ja oikeiden vastauspaikkojen osoittamisen lisäksi on kysymyksiin usein tarpeen tehdä sisällöllisiä täsmennyksiä. Muistin- tai arvionvaraisia tietoja kysyttäessä voi olla luontevaa käyttää vastaamista helpottavia väljennyksiä; esimerkiksi osallistumiskertoja tiedusteltaessa saattaa olla paikallaan antaa vastaajan vastata muotoiluun "noin ____ kertaa".

Abstrakteja tai yleisiä asioita kysyttäessä on joskus tapana sisällyttää arvioitavaan asiaan esimerkkejä. Usein ne merkitään suluin kysymyslauseen loppuun. On kuitenkin tiedostettava, että esimerkit saattavat rajata vastaajan ajattelua pelkästään mainittuihin asioihin, vaikka niiden varsinainen tarkoitus on tehdä ymmärrettäväksi muutoin vaikeaselkoista asiaa. Tästä syystä esimerkkejä tulee käyttää ainoastaan hyvin painavista syistä. Ensisijaisesti kysymykset tulee laatia niin selkeiksi, etteivät ne kaipaa mahdollisesti ohjailevia esimerkkejä.

Esimerkissä 6 on malleja vastausohjeista.

Kysymysten rakennevaihtoehtoja

Yksi vaikeimmista lomakkeen laatimisen ongelmista koskee sitä, kysytäänkö kysymykset yksittäin vai sarjoissa. Kun halutaan selvittää samaan asiaan liittyviä tekijöitä tai

vastausvaihtoehdoiltaan yhteneviä kysymyksiä, on kysymyssarjojen eli kysymyspatteristojen käyttö hyödyllistä.

Samoihin asiakokonaisuuksiin liittyviä yksittäisiä seikkoja kannattaa kysyä erikseen esimerkiksi luetteloin. Kysymyssarjoihin vastaaminen on vastaajille usein helpompaa kuin lukea joko monimutkaisia vaihtoehtoja tai itseään toistavia kysymyksiä.

Joissakin tapauksissa yksittäisten kysymysten ja vaihtoehtojen raja on kuitenkin veteen piirretty viiva. Näin on silloin, kun joihinkin asioihin liittyviä ominaisuuksia ei ole esimerkiksi arvioinnin vaikeuden vuoksi tarkoituksenmukaista kysyä kovin seikkaperäisesti. Tällöin voidaan turvautua ns. monivastauskysymyksiin:

- Listata tiettyyn asiakokonaisuuteen liittyvät asiat ja pyytää vastaajia merkitsemään kaikki hänen kohdallaan kyseeseen tulevat vastaukset. *Ks. esimerkki 7.*
- Listata asiat ja pyytää vastaajia mainitsemaan niistä tärkeimmät (esim. korkeintaan 3-5 asiaa, riippuen tietenkin kohtien kokonaislukumäärästä). *Ks. esimerkki 8.*
- Listata asiat ja pyytää vastaajia asettamaan ne tärkeysjärjestykseen merkitsemällä tärkeintä merkinnällä '1.', toiseksi tärkeintä merkinnällä '2.', jne. *Ks. esimerkki 9.*

Vastausvaihtoehdoissa huomioitavia seikkoja

Yleensä lomakkeesta tallennetaan havaintomatriisiin numeroita ja siksi vastausvaihtoehdotkin kannattaa ehdottomasti luetella numeroilla, ei kirjaimilla. Tämä vähentää virheitä tietojen tallentamisessa ja kohentaa tutkimuksen reliabiliteettia.

Strukturoitujen kysymysten vastausvaihtoehtojen tulee periaatteessa aina olla toisensa poissulkevia. Joitain poikkeuksia on ja useimmiten ne liittyvät joko preferenssikysymyksiin (pyydetään nimeämään esimerkiksi ensisijainen vastaus) tai monivalintakysymyksiin. Vastausvaihtoehtojen päällekkäisyys on valitettavan yleinen ongelma ja siksi tähän liittyvissä erimerkeissä käsitellään asiaa melko yksityiskohtaisesti.

Esimerkissä 10 on malleja toisensa poissulkeviksi tarkoitetuista vaihtoehdoista.

Sanalliset skaalat ja niitä vastaavat vastausvaihtoehtojen numerot antavat enemmän mahdollisuuksia tutkimustulosten kuvailuun raportointivaiheessa (esimerkiksi Likert-asteikolliset muuttujat). Toisaalta tietyt tilastolliset menetelmät edellyttävät tarkkaa mittaustasoa, joka voidaan saavuttaa poistamalla vastausvaihtoehtoja tarkasti vastaavat sanamuodot ja laventamalla vastausskaalaa (esim. mielipiteen sijoittuminen asteikolla nollasta sataan).

Kummassakin tavassa on myös haittapuolia. Skaalojen keinotekoinen muokkaaminen tilastollisen analyysin kriteereitä vastaaviksi vie vaihtoehdot kauaksi ihmiskielelle ja -mielelle tavanomaisesta ajattelusta ja vaihtoehdoiltaan laajoja skaaloja voidaan herkästi käyttää eri tavoin. Toisaalta sanoiksi puetut vastausskaalat ovat tosiasiassa korkeintaan järjestyslukuasteikollista mittaamista, koska tuntuu mielivaltaiselta ajatella, että lukuarvosta yksi lukuarvoon kaksi olisi yhtä paljon "matkaa" kuin esimerkiksi "täysin samaa mieltä" olemisesta "jokseenkin samaa mieltä" olemiseen. (Ks. muuttujien mittaustaso.) Useimmilla yhteiskuntatieteellisillä tieteenaloilla sanalliset vastausskaalat ja niihin liittyvä vastausvaihtoehtojen numerointi kuuluvat kuitenkin vahvasti tutkimusperinteeseen.

Joskus lomaketutkimuksissa näkyy käytettävän paljon dikotomisia vaihtoehtoskaaloja. Tällaiset kahden vaihtoehdon kysymykset ovat helppoja vastata, mutta niiden tarjoama informaatio ei ole ainakaan asenteita tai käyttäytymisaikomuksia mitattaessa kovin "rikasta". Joskus dikotomisia muuttujia on kuitenkin tarkoituksenmukaista käyttää siten, että niistä voidaan luoda yhdistettyjä muuttujia ja tarkasteluja.

Huomautettakoon myös, ettei uusia ja kovin omintakeisia vastausskaaloja kannata käyttää, jos tarjolla on muita testattuja ja toimivia ratkaisuja. Ensisijaisesti kannattaa turvautua aiemmin käytettyihin ja tiedeyhteisön jo omaksumiin vastausskaaloihin. Ne voivat usein myös olla vastaajille entuudestaan tuttuja. "Vanhojen" kysymysten käyttö on suositeltavaa senkin vuoksi, että niiden reliabiliteettia ja validiteettia on jo tutkittu aiemmin.

Myös arvioitavien asioiden tutuksi tulemista vastaamisen yhteydessä voi yrittää hyödyntää. Tarpeen vaatiessa samojen aiheiden eri puolia kannattaa pyytää arvioimaan eri asteikoilla, jolloin lomakkeessa säästyy tilaa ja vastaaminenkin saattaa helpottua.

"En osaa sanoa", "en tiedä", "vaikea sanoa" tai "en halua sanoa" -vaihtoehtojen käyttöön ei ole olemassa yksiselitteisiä ohjeita. Niitä kannattaa käyttää tarpeen mukaan, mutta ei kuitenkaan tarjoilla vastaajille liian herkästi. Ns. EOS-vastausta käytetään tavallisesti skaalan lopussa, jolloin se kerää vähemmän vastauksia kuin keskelle sijoitettuna. Joissakin kyselyissä tai kysymyksissä ei käytetä näitä vaihtoehtoja ollenkaan, mutta tällöin riskinä on vastausten reliabiliteetin näennäinen kohottaminen tai joidenkin vastaajien turhautuminen.

"Vaikea sanoa" sijoitetaan usein esimerkiksi mielipideskaalan keskelle, jolloin se saattaa kerätä paljonkin vastauksia. Sitä lähellä on vaihtoehto, jossa vastaaja kieltää ajattelevansa skaalan kummankaan pään mukaisesti ("en ole samaa enkä eri mieltä".)

Usein on myös järkevää sisällyttää monivalintakysymysten viimeiseksi arvioitavaksi kohdaksi tai joissakin tapauksissa vaihtoehtoskaalan loppuun "muu, mikä" -vaihtoehto. Tällöin vastaaja pääsee "sanomaan sanottavansa" sellaisesta asiasta, jota hänen mielestään olisi pitänyt kysyä kysymyksessä tai kertomaan syystä tai toisesta täysin vastausskaalasta poikkeavan sijoittumisensa.

Lisäksi mainittakoon, että etenkin postikyselyissä on erittäin tärkeää kyetä aina erottamaan, onko vastaaja ylipäänsä vastannut kysymykseen vai tarkoittaako tyhjä vastaus "nollaa" tai esim. "ei kertaakaan". Jos nolla voi olla "oikea" vastaus, on hyvä ilmoittaa kysymyksen vastausohjeessa, millä tavoin vastaajan tulee merkitä nollaa tarkoittavat vastaukset (esim. merkitsemällä viiva tai kirjoittamalla "ei kertaakaan").

Näkökulmia kysymysten sisältöön ja tyyliin

Lomakekysymysten laatimisohjeissa (esim. Jyrinki 1976) viitataan usein hyvin perustellusti siihen, että vastaajien on usein helpompaa vastata omakohtaisiksi koettuihin kysymyksiin. On myös tunnettua, että vastaajille yleisessä muodossa esitetyistä kysymyksistä voidaan saada kovin erilaisia tuloksia kuin jos asian yleistä tilaa oltaisiin mitattu yhdistämällä vastaajien arvioita asiaan liittyvästä omakohtaisesta tilanteesta.

Esimerkiksi vuoden 1994 presidentinvaalien 1. kierroksen jälkeen tehdyssä mittauksessa 21 % valtakunnallisen otoksen vastaajista katsoi kannatusmittauksilla olleen yleensä ottaen paljon vaikutusta äänestäjien ehdokasvalintoihin, mutta vain yksi prosentti vastaajista myönsi mittausten vaikuttaneen paljon omaan ehdokasvalintaan! (Ks. FSD:n aineisto nro FSD1019.)

Joskus kyselyn alussa tai yksittäisten kysymysten yhteydessä onkin tarpeen erityisesti korostaa sitä, että tutkimuksessa ollaan kiinnostuneita juuri vastaajan omasta mielipiteestä. Kysymysten laatija voi myös tukea tätä pyrkimystä välttämällä johdattelevia kysymyksiä, joita näkee esitettävän valitettavan usein. Läheskään aina kyse ei ole tietoisesta harhaanjohtamisesta, vaan pieniä suuntimia voi syntyä miltei tiedostamatta.

Kysymysten tasapainoisuuteen vaikuttaa useita tekijöitä. Sinänsä tasapuolinen ja hyvä kysymys on mahdollista pilata epätasapainoisella skaalalla tai vastausvaihtoehdoiltaan moitteeton kysymys voi olla kohtuuttoman yksipuolinen.

Kysymyksenasettelu onkin helposti johdatteleva, jos kysymykseen sisällytetään esimerkiksi mielipiteiden tai toimenpiteiden valittuun suuntaan ohjaavia sanavalintoja. Valitettavan tavanomaista on kysyä "kuinka tärkeinä pidätte seuraavia asioita?" ja saada vastaukseksi kaikkien asioiden tärkeyttä!

Esimerkissä 11 on malleja johdattelevista ja epätasapainoisista kysymyksistä.

Kysymyksen jakaminen osiin tai asettaminen kysymyssarjaksi vähentää usein "tasapaino-ongelmia". Tämä neuvo sopii kysymyksiin, joissa tiedustellaan useampia asioita kuin on tarkoitus, tai joissa tarjoillaan vastausvaihtoehdoiksi asioita, jotka eivät todellisuudessa ole toisensa poissulkevia. (*Vrt. esimerkki 10e*.)

Oman ongelma-alueensa muodostavat arkaluonteiset kysymykset, joiden kokonaissuunnitteluun on panostettava riittävästi. Asian koskee muun muassa kyselyn saatetekstien muotoiluja, varsinaisten kysymysten tietosisältöjä, taustamuuttujien sijoittelua lomakkeessa, vastausvaihtoehtojen luokittelua ja yleensä mittaamiseen sopivaa tarkkuustasoa. Mikäli kyselylomake sisältää useita arkaluonteisia kysymyksiä, sitä tärkeämpää on testata ja muokata tutkimusinstrumenttia tältä osin ennen varsinaista tiedonkeruuta.

Tutkimuseettisiä näkökohtia

Tutkimuseettisten näkökohtien huomioon ottaminen sopii kokoavaksi näkökulmaksi myös tieteellisen kyselytutkimuksen lomakesuunnitteluun. Tiedonkeruu tieteen piirissä ja nimissä on toteutettava huolellisesti tieteen objektiivisuutta tukevia välineitä kunnioittaen. Tutkijan on suunniteltava tutkimusinstrumenttinsa tutustumalla riittävästi ja samaa aihetta koskeviin aikaisempiin julkaisuihin ja tutkimusaineistoihin. Näin hän kykenee tunnistamaan tutkimuksen todelliset aukot ja osaa laatia niitä varten tarkoituksenmukaiset tiedonkeruuinstrumentit.

Tutkijan ei myöskään pidä tehdä tieteellistä tutkimusta vain tilaajalleen, omalle organisaatiolle tai itselleen. Tieteen avoimuus, tulosten kontrolloitavuus ja niistä keskusteleminen tiedeyhteisön piirissä ovat luovuttamaton osa prosessia, jonka läpi tulokset muovautuvat tieteellisiksi.

Lisäksi ainakin julkisin varoin operoivien tutkimusaineistojen kerääjien tulisi aina pyrkiä palvelemaan tiedeyhteisön yhteisiä tavoitteita siten, että uusi tutkimusaineisto siirtyy oman aktiivisen ensikäyttövaiheen jälkeen muidenkin tutkijoiden hyödynnettäväksi. Yhteiskuntatieteiden aloilla tätä tarkoitusta palvelee Suomessa Yhteiskuntatieteellinen tietoarkisto, jonne arkistoidut aineistot ovat maksutta käytössä tieteelliseen tutkimukseen ja opetukseen.

Lisäesimerkit lomakesuunnitteluun

1. Lomakkeen laajuus ja ulkoasu

Laajuudeltaan ja ulkoasultaan malliksi kelpaavia lomakkeita on esimerkiksi Elinkeinoelämän valtuuskunnan eli EVA:n kyselyissä. Niissä on tosin alusta pitäen tähdätty tiiviiseen, barometrityyppiseen raportointiin. Esim. FSD:n aineisto nro FSD1051.

Akateemisten tutkimushankkeiden suunnittelemissa kyselyissä lomakkeet laajenevat usein pidemmiksi. Vuoden 2000 Suomen ISSP-aineistossa käytetty lomake on jo lähellä ihannemitan maksimia. Esim. FSD:n aineisto nro FSD0115.

HUOM! Kyselylomakkeiden katselu Internetissä edellyttää, että käyttäjän työasemalta löytyy Adoben Acrobat Reader-ohjelma.

2. Vastaamisen yhteismitalliset arviointiperusteet

Kuvitellaan, että jokin organisaatio haluaisi selvittää sisäisen tiedonkulkunsa toimivuutta, ja että se päätyisi mittaamaan asiaa vain seuraavalla kysymyksellä:

```
Miten arvioit tiedonsaantia organisaatiomme sisäisestä toiminnasta?
(esim. päätösten sisältö, taustat, valmistelu)

Tietoa sisäisestä toiminnasta on tarjolla
1 täysin riittämättömästi
2 jokseenkin riittämättömästi
3 jokseenkin riittävästi
4 täysin riittävästi
5 en osaa sanoa
```

Millaisia ongelmia tähän kysymykseen vastaamisessa ja saaduissa tuloksissa voisi ilmetä? Ensimmäinen pulma voi olla, ajattelevatko vastaajat tilannetta yleensä vai tiedonsaantia vain omalta osaltaan. Yleensä mielipidekyselyissä on tarkoituksenmukaista muotoilla kysymykset siten, että niissä tiedustellaan vastaajan omakohtaista mielipidettä. Mikäli mielipiteet koskevat asioita, joista vastaajalla on omakohtaista kokemusta, kysymys kannattaa kohdentaa tähän kokemukseen.

Esimerkissä olisi myös syytä tiedustella sisäiseen tiedonkulkuun liittyviä asioita aiheryhmiin jaoteltuna (vrt. kysymyksen suluissa mainitut asiat). Olisi voitu kysyä, kuinka paljon vastaajat ovat tekemisissä kunkin osa-alueen kanssa ja mitä mieltä he ovat kuhunkin seikkaan liittyvästä tiedonsaannista.

Toinen ongelma liittyy siihen, millä tavoin kyselyssä pyritään tuottamaan yleistävää tietoa useista osa-alueista koostuvista kokonaisuuksista. Suhteellisen yksityiskohtainen asiakokonaisuuksien osiin jakaminen on suotavaa edellyttäen, että kysely ei laajene liikaa. Yleistävä kokonaiskuva on sitten jälkeenpäin mahdollista koota yhdistämällä yksittäisten kysymysten tuloksia esimerkiksi summamuuttujilla. On huomattava, että nyt esimerkkikysymyksen sisällöllisessä täsmennyksessä esitetään useita aiheita. Ne voivat painottua eri tavoin vastaajien arvioinneissa.

Voidaan myös kysyä, onko sisäisestä toiminnasta tarjolla olevan tiedon määrä ainoa olennainen arvioinnin kohde. Nyt kysymyksen eri osissa puhutaan ensin tiedonsaannista, mutta vastausvaihtoehtoihin johtavassa osassa kapeammin ensin vain 'tarjolla olemisesta' ja sitten vastausvaihtoehdot liitetään vain tietojen tarjolla olemisen riittävyyteen.

Tällaiset epäjohdonmukaisuudet saattavat johtaa siihen, että kysymyksiin vastataan ei-toivotulla tavalla erilaisin arviointiperustein. Vastaamisperusteiden yhteismitallisuus on koetteilla tavanomaisesti myös silloin, kun kysymykset tai vastausvaihtoehdot ovat liian pitkiä

tai vaikeaselkoisia. Ylipitkät kysymyslauseet sisältävät herkästi enemmän kuin yhden kysymyksen ja juuri siksi ne tulisi jakaa osiin tai keskittyä kysymään vain olennaisinta asiaa.

3. Esimerkkejä sinuttelevista ja teitittelevistä lomakkeista

- Sinuttelu: Ks. FSD:n aineisto nro FSD1027.
- Teitittely: Ks. FSD:n aineisto nro FSD1045.

4. Mittaamisen näennäistarkkuus

Eräässä tutkimuksessa vastaajia pyydettiin kuvaamaan lämpömittariasteikolla -50 ... +50 suhtautumistaan yhteiskunnallisiin ryhmiin ja tahoihin. Vastaavantyyppinen asenneskaala esiintyy joissakin muissa tutkimuksissa hieman erilaisella asteikolla nollasta sataan (0-100). Monet vastaajat eivät kuitenkaan käyttäneet lainkaan mielellään asteikon negatiivisia arvoja vaan arvioivat ryhmiä lähinnä välillä 0 ... +50. Toiset vastaajat taas käyttivät skaalaa koko laajuudessaan, mikä heikensi tulosten luotettavuutta.

Joskus suhtautumista joihinkin asioihin pyydetään ilmaisemaan sijoittamalla oma kanta ainakin näennäisesti ei-numeeriselle ulottuvuudelle, jolla on sanoin nimetyt ääripäät, esimerkiksi:

```
Heikko                                                  Vahva

|-------------------------------------------------------|


Luotettava                                    Epäluotettava

|-------------------------------------------------------|
```

Janoille merkityt vastaukset olisi mahdollista koodata jopa millimetrin tarkkuudella. Mutta kertoisivatko itsesijoitukset vastaajan suhtautumisesta tarkemmin kuin vastaukset esimerkiksi viisiportaiseen sanalliseen skaalaan? Ainakin hyvin tiheän skaalan käyttäminen lisää erilaisten vastaamistapojen riskiä.

5. Avoin kysymys

Täysin avoin kysymys voi koskea hyvinkin laaja-alaista aihealuetta, kuten suhtautumisen tai toiminnan vapaamuotoisia perusteluja tai jonkin muun asian vapaamuotoista arviointia. Esimerkiksi:

```
"Mitä mieltä olette Suomen jäsenyydestä Euroopan unionissa?"

"Miksi ette ole osallistunut yhdistyksen toimintaan viimeksi
kuluneiden 12 kuukauden aikana?"
```

Tutkija voi halutessaan lisätä avoimeen kysymykseen myös erilaisia täsmentäviä vastausohjeita; esim. "arvioi asiaa ... siitä ja siitä ... näkökulmasta". Mikäli avoimeen kysymykseen on odotettavissa paljon erilaisia mainintoja samalta vastaajalta, voi täsmennys koskea lueteltavien asioiden määrää, esimerkiksi "mainitse korkeintaan niin ja niin monta kohtaa". Myös itse kysymyslauseeseen voi liittää erilaisia rajauksia, kuten pyynnön mainita vain tärkein tai tärkeimpiä asioita.

Puoliavoimeen kysymykseen voidaan yhdistää strukturoitu ja avoin osuus. Ensin voidaan esimerkiksi kysyä jotakin vastausvaihtoehdoilla kyllä/ei ja pyytää perustelemaan vastaus avoimella kysymyksellä. Perustelujen tivaaminen vain strukturoidun kysymyksen tiettyyn vastauskategoriaan vastanneilta on kuitenkin usein johdattelevaa. Mikäli lomakkeessa pyydetään

esimerkiksi perustelemaan vain kielteisiä vastauksia, ovat kysymykset (ja tutkimuksen tavoitteetkin) herkästi epätasapainoisia. [Ks. kohdan 11 esimerkit]

Avointa kysymyksenasettelua voidaan käyttää myös silloin, kun lomakkeessa ei tilanpuutteen vuoksi ole mahdollista luetella kaikkia vaihtoehtoja (esim. vastaajan asuinkunta). Mikäli vastausvaihtoehtojen listasta voi tehdä liitteen lomakkeeseen, vastaajaa on mahdollista pyytää merkitsemään vastauskenttään oikeaa vaihtoehtoa vastaava numero. Esimerkiksi verkkokyselyihin verrattuna tilaa vievien lisätietojen antaminen on postikyselyissä hankalaa. [Ks. myös avointen kysymysten koodaus]

6. Vastausohjeet

6a Lomakkeen alun yleinen vastausohje

Kyselyyn vastaamiseen liittyviä yleisiä asioita kuten saatteita käsitellään tietovarannon kohdassa postikyselyn toteuttaminen. Kyselylomakkeen alkuun on syytä aina suotavaa liittää seuraavan esimerkin tapainen yleinen vastausohje.

```
"Rengastakaa jokaisen kysymyksen kohdalla omaa näkemystänne parhaiten
vastaavan vaihtoehdon numero. Muutamassa kysymyksessä vastaus kirjoitetaan
sille varattuun tilaan."
```

(Lähde: FSD0115 ISSP 2000: Suomen aineisto.)

6b Sisällölliset ja tekniset täsmennykset

Seuraavassa kysymysluonnoksessa on useita sisällöllisiä ja teknisiä täsmennyksiä. Luottamustehtävissä toimimisen ajankohta ja huomion kiinnittäminen vain varsinaisiin jäsenyyksiin ovat sisällöllisiä täsmennyksiä. Teknisiä vastausohjeita ovat mm. A- ja B-kysymysten erottaminen toisistaan mainitsemalla asteikkojen erot sekä se, että B-kysymykseen vastataan oikeanpuoleisessa sarakkeessa olevilla numeroilla.

```
A) Missä luottamustehtävissä toimitte tai olette toiminut?
(varajäsenyyksiä ei huomioida; asteikko 1-3)
B) Entä missä luottamustehtävissä olisitte kiinnostunut toimimaan?
Rengastakaa oikeanpuoleisesta pystysarakkeesta kaikki ne tehtävät (1 -
10), joissa ette toimi nyt, mutta jotka kiinnostaisivat Teitä.
```

	Toimin tällä hetkellä	Olen toiminut, en toimi enää	En ole toiminut koskaan	Olen kiinnostunut toimimaan
Tehtävä 1	1	2	3	1
Tehtävä 2	1	2	3	2
Tehtävä 3	1	2	3	3
Tehtävä 4	1	2	3	4
Tehtävä 5	1	2	3	5
Tehtävä 6	1	2	3	6
Tehtävä 7	1	2	3	7
Tehtävä 8	1	2	3	8
Tehtävä 9	1	2	3	9
Tehtävä 10	1	2	3	10

Oikean vastauspaikan osoittaminen onnistuu moniosaisen kysymyksen sisällä kysymystekstiin rakennetuin sanallisin täsmennyksin (vrt. edellinen kaksiosainen esimerkkikysymys). Vastaajien siirtäminen kysymysten yli seuraaviin kysymyksiin (eli ns. hypyt) on merkittävä kyselylomakkeeseen riittävän näkyvästi.

Lähtökysymyksessä siirto voidaan merkitä tiettyjen vastausvaihtoehtojen perään...

```
Kysymys 21. Onko sinulla voimassa oleva, auton ajamiseen tarvittava
ajokortti?
1 Kyllä
2 Ei » siirry kysymykseen 28
```

tai kirjoittaa siirtoa koskeva huomautus kysymyksen jälkeen...

```
Kysymys 21. Onko sinulla voimassa oleva, auton ajamiseen tarvittava
ajokortti?
1 Kyllä
2 Ei

HUOM! Ajokortittomat eli kysymykseen 21 'ei' vastanneet siirtyvät
kysymykseen 28 (sivu 7).
```

Ennen tulokysymystä voi olla jälleen tarpeellista huomauttaa lomakkeessa siitä, että kaikki vastaavat seuraavaan kysymykseen. Oikean vastauspaikan selkeä osoittaminen on siis erittäin tärkeätä erityisesti kirjekyselyissä, joissa vastaajalla ei yleensä ole käytettävissään muita kuin kirjallisia vastausohjeita. Käyntikyselyissä asia voidaan hoitaa haastattelijan toimesta ja esimerkiksi puhelin- ja verkkokyselyissä hypyt voidaan automatisoida lomakkeeseen.

7. Monivalintakysymys dikotomisin vastausvaihtoehdoin

Dikotomisia eli kahden vastausvaihtoehdon monivalintakysymyksiä käytetään muun muassa selvitettäessä erilaisten toimintojen käyttöä ja tuntemusta, tai valittaessa lukuisista eri ominaisuuksista soveltuvimpia. Koska vastausvaihtoehdoiltaan dikotomisiin monivalintakysymyksiin on suhteellisen nopeaa vastata, niitä käytetään runsaasti myös erilaisten asioiden karkean tärkeysjärjestyksen kartoittamiseen.

Ideana on, vastaajat arvioivat kutakin yksittäistä asiaa joko kyllä/ei -skaalalla tai merkitsevät esimerkiksi vastausruutuihin kaikki kysymykseen tulevat kohdat. Tuloksia koottaessa useimmin mainitut tulkitaan vastaajille keskimäärin tärkeimmiksi. Esimerkiksi Euroopan Unionin komission teettämät eurobarometrit sisältävät runsaasti seuraavan esimerkin kaltaisia monivalintakysymyksiä dikotomisin vastausvaihtoehdoin:

8. Monivalintakysymys rajatuin vastausmäärin

Kaikista monivalintakysymyksen yksittäisistä kohdista ei ole pakko tehdä data-aineistoon omia muuttujia. On myös mahdollista pyytää vastaajia mainitsemaan suurestakin arvioitavien asioiden joukosta korkeintaan tietty määrä asioita. Tällöin vastaukset voidaan tallentaa korkeintaan siihen määrään muuttujia kuin vastaajia on pyydetty nimeämään asioita. Seuraavan esimerkin vastaukset voitaisiin tallentaa kolmeen muuttujaan arvoilla 1-10. (Käytännössä kukin tavoite siis voitaisiin koodata myös omaan muuttujaan, mutta silloin datatiedoston tallentamistyön määrä kasvaisi.)

```
Rengasta alla olevasta numerosarakkeesta korkeintaan kolme tavoitetta,
joita kuntamme sosiaali- ja terveystoimessa tulisi erityisesti
painottaa seuraavien viiden vuoden aikana.
Tavoite A 1
Tavoite B 2
Tavoite C 3
Tavoite D 4
Tavoite E 5
Tavoite F 6
Tavoite G 7
Tavoite H 8
Tavoite I 9
Tavoite J 10
```

9. Järjestyslukuasteikolliset monivalintakysymykset

Tutkijan on usein saatava monivalintakysymyksellään dikotomisia vastausvaihtoehtoja tarkempaa tietoa. Silloin vastaajia voi pyytää asettamaan arvioitavia asioita tärkeysjärjestykseen. Kun asioita on paljon, esimerkiksi kymmeniä, on harvoin kuitenkaan järkevää pyytää vastaajaa asettamaan kaikki vastaukset preferenssijärjestykseen. Tavanomaista on tiedustella kolmesta viiteen tärkeintä asiaa ja teknisesti tämän voi toteuttaa monin eri tavoin. Joka tapauksessa vastaamisessa käytettävät merkinnät on ohjeistettava riittävän yksityiskohtaisesti.

Pari esimerkkiluonnosta:

```
Mitkä seuraavista asioista ovat sinulle henkilökohtaisesti tärkeimpiä
omassa elämässäsi? Merkitse viivoille tärkein numerolla yksi (1.)
toiseksi tärkein numerolla kaksi (2.) ja kolmanneksi tärkein numerolla
kolme (3.)

Asia A ___
Asia B ___
Asia C ___
Asia D ___
Asia E ___
Asia F ___
Asia G ___
Asia H ___
```

tai

```
Mitkä seuraavista asioista ovat sinulle henkilökohtaisesti tärkeimpiä
omassa elämässäsi? Rengasta riveiltä tärkein, toiseksi tärkein ja
kolmanneksi tärkein. (Kultakin riviltä vain yksi rengastus.)

               Asia A Asia B Asia C Asia D Asia E Asia F Asia G Asia H

 1. tärkein asia   1      2      3      4      5      6      7      8
 2. tärkein asia   1      2      3      4      5      6      7      8
 3. tärkein asia   1      2      3      4      5      6      7      8
```

Vertailevassa sosiaalitutkimuksessa on käytetty täydelliseen tärkeysjärjestykseen perustuvaa mittaamistapaa muun muassa kartoitettaessa ihmisten arvostuksia. Yleensä ottaen preferenssikysymysten käyttö on tarkoituksenmukaista silloin kun arvioitavat asiat ovat vastaajille riittävän tuttuja. Näin lienee asian laita ainakin arvostuksiin ja ajankäyttöön liittyvien asioiden kohdalla. Suhtautumista hyvin abstrakteihin asioihin ei kannattane mitata liian tarkoin preferenssimittauksin.

Vastaajien itse täyttämissä lomakkeissa on ongelmana myös se, että vastaajat eivät aina noudata vastausohjeita; merkintöjä tehdään enemmän kuin on pyydetty tai järjestyslukuja ei käytetä kuten on ohjeistettu. Tämä aiheuttaa ongelmia erityisesti lomakkeiden tallennusvaiheessa ja saattaa vääristää tuloksia osioiden alkupäässä mainittujen asioiden hyväksi. Tällaisia ongelmia voidaan vähentää suunnittelemalla lomakkeiden tallennus siten, että virheellisesti täytettyjä lomakkeita kohdellaan eri asiasisältöjen kannalta tasapuolisesti.

10. Toisensa poissulkevat vastausvaihtoehdot

Strukturoituihin kyselylomakkeisiin perustuvien tutkimustulosten reliabiliteetti riippuu huomattavassa määrin siitä, kuinka onnistuneesti kysymykset ja niiden vastausvaihtoehdot kattavat tutkimuskohteet. Vaihtoehtojen onnistuneisuus riippuu erityisesti kahdesta asiasta: siitä kuinka laaja-alaisesti ja tasapainoisesti vastausvaihtoehdot kattavat tutkittavan ilmiön, ja kuinka hyvin ne tukevat perusteiltaan yhtenevää vastaamistapaa. Viimeksi mainittu edellyttää yleensä

sitä, että vastausvaihtoehdot eivät sisällä vastauksia muihin kuin kysyttyyn kysymykseen, ja että vaihtoehdot ovat toisensa poissulkevia.

Toisensa poissulkevuuden kriteeri jää noudattamatta tyypillisesti silloin, kun numeerisesti määritellyt vastausvaihtoehdot ovat jääneet päällekkäisiksi. Näin on esimerkissä 10a ja tarkkaan ottaen myös esimerkin 10b asukasmääriin perustuvassa kuntakokoluokituksessa.

Esimerkki 10a:

```
Kuinka monena päivänä viikossa seuraat
televisiosta kotimaisia viihdesarjoja?

1   Kerran viikossa
2   2-3 krt viikossa
3   3-5 krt viikossa
4   5-7 krt viikossa
```

Esimerkki 10b:

```
Mikä on kotikuntanne nykyinen asukasmäärä?

1   - 5000 asukasta
2   5000 - 10000 asukasta
3   10000 - 50000 asukasta
4   50000 - 100000 asukasta
5   100000 asukasta tai enemmän
```

Kysymysten vastausvaihtoehtojen taitekohdat tulisi määritellä yksikäsitteisesti siten, että tasaluvut eivät kuulu kahteen eri luokkaan. Esimerkissä 10a on lisäksi ilmeistä, että mainitun asian mittaaminen päivinä viikossa ei välttämättä ole tarkoituksenmukaisin mittayksikkö, ja että kysymyksen ja vastausvaihtoehtojen välillä on muutakin ristiriitaa. Mikä voisi olla soveltuvampi tapa mitata kotimaisten viihdeohjelmien katselumäärää?

Sanallisesti määriä kuvailevissa luokituksissa ovat usein ongelmana myös taitekohtien katvealueet, ei pelkästään vaihtoehtojen päällekkäisyys. Esimerkissä 10c esiintyy kuitenkin molempia, joskaan kummassakaan tapauksessa puutteet eivät ole suuria. Oletetaan, että tutkija haluaisi selvittää tietyn radiokanavan uutislähetysten keskimääräistä kuuntelemisaktiivisuutta viimeksi kuluneiden 12 kuukauden ajalta, ja että hän laatisi mittaria varten seuraavan vastausskaalan:

Esimerkki 10c:

```
1   Useita kertoja päivässä
2   Päivittäin
3   Muutaman kerran viikossa mutta en päivittäin
4   Kerran pari viikossa
5   Kerran pari kuukaudessa
6   Harvemmin tai en koskaan
```

Vaihtoehtojen 1 ja 2 välillä on toki koko vastausskaalaa ajatellen selvä eroavuus: luultavasti useimmat vastaajat ymmärtäisivät äärivaihtoehdon 'useita kertoja päivässä' siten, että se kuvaa vähintään kahdesti päivässä tapahtuvaa asiaa. 'Päivittäin' -vaihtoehto ei sitä vastoin rajaa asiaa määrällisesti yhtä selvästi, joskin luultavasti useimmat vastaajat ymmärtäisivät sen vaihtoehtojen 1 ja 3 väliin sijoittuvaksi, vain kerran päivässä tapahtuvaksi asiaksi. Pilkuntarkasti tarkasteltuna vaihtoehdoilla 1 ja 2 on näin muotoiltuina päällekkäisyysriski, joten 'päivittäin' voitaisiin korjata esimerkiksi vaihtoehdolla 'kerran päivässä'.

Mittaamisen katvealueet jäävät tässä vastausskaalassa selvimmin vaihtoehtojen 4 ja 5 väliin. Esimerkiksi kolme kertaa kuukaudessa on vähemmän kuin kerran viikossa mutta joidenkin mielestä enemmän kuin pari kertaa kuukaudessa. Pitäisikö vastausvaihtoehdot muuttaa kokonaan numeerisiksi tai yhdistää vahvemmin numeerista ja sanallista kuvailua?

Skaalan muuttaminen kokonaan numeeriseksi voisi olla hankalaa, koska ihmiset ovat tottuneet puhumaan ja ajattelemaan pitkän ajan kuluessa tapahtuvaa toimintaa päiviin, viikkoihin ja kuukausiin väljemmin liittyvillä termeillä. Muistinvaraisia asioita ei pidä myöskään kysyä liian tarkasti, koska silloin vastaaminen käy hankalaksi, ja se voi puolestaan vaikuttaa vastausprosenttiin. Kompromissiratkaisun voisi tarjota seuraava vaihtoehtolista. Tosin siihen ei liene yhtä helppoa vastata kuin edelliseen esimerkkiin.

Esimerkki 10d:

```
1   Ainakin kahdesti päivässä
2   Noin kerran päivässä
3   Noin 4-6 kertaa viikossa
4   Noin 1-3 kertaa viikossa
6   Noin 1-3 kertaa kuukaudessa
7   Harvemmin tai en koskaan
```

Eksaktin kvantitatiivisesti määriteltyjen vastausvaihtoehtojen mittareissa on yleensä verraten yksinkertaista muodostaa toisensa poissulkevia vastausvaihtoehtoja. Sanallisia vaihtoehtoja käytettäessä sisällöllisesti tarkoituksenmukaisten ja teknisesti toimivien skaalojen rakentaminen on huomattavasti työläämpää. Seuraava esimerkki ilmentää useita asiaan liittyviä tyypillisiä ongelmia silloin, kun lyhyt kysymys ryöstäytyy vaihtoehtojensa kautta aiottua monialaisemmaksi:

Esimerkki 10e:

```
Autojen pysäköintitilaa keskustassa
1   On nykyisellään sopivasti
2   On vähennettävä, jotta saadaan tilaa muille toiminnoille
3   On rakennettava maan päälle
4   Tulee rakentaa maan alle vähentämättä maanpäällisiä
    pysäköintipaikkoja
5   Tulee siirtää maan alle siten, että samalla vähennetään
    maanpäällisiä pysäköintipaikkoja
```

Oheisen kysymyksen ongelma on siinä, että yhteen ja samaan kysymykseen on ahdettu liian monen kysymyksen eväät. Tällaisessa tapauksessa vaihtoehdot eivät juurikaan voi muodostua toisensa poissulkeviksi, eikä kysymyksellä muutoinkaan saada tasapainoista kuvaa sen sisältämiin asioihin liittyvistä mielipiteistä. Kysymyksessähän ollaan kiinnostuneita ainakin siitä, mitä vastaajat ajattelevat

- keskustan pysäköintitilojen nykyisestä määrästä yleensä (vaihtoehto 1),
- tarpeesta saada tilaa muille toiminnoille keskustassa (vaihtoehto 2),
- keskustan pysäköintitilojen rakentamistarpeesta ja tarkemmasta sijainnista (vaihtoehdot 3, 4 ja 5)
- keskustan maanpäällisten tilojen määrästä (vaihtoehdot 4 ja 5),
- keskustan pysäköintitilojen rakentamistarpeesta maan alle (vaihtoehdot 4 ja 5) sekä
- tarpeesta käsitellä keskustan pysäköintitilojen rakentamista kokonaisuutena, jossa uusien paikkojen rakentaminen voisi vaikuttaa olemassa olevien paikkojen määrään (vaihtoehdot 4 ja 5).

Kysymys tulisikin jakaa osiin ja muodostaa tarvittaessa aihealueita koskevia erillisiä kysymyksiä toimivine vastausvaihtoehtoineen. Tämä neuvo pätee hyvin usein niihin tilanteisiin, joissa kysymyksestä tuntuu olevan vaikeata muodostaa tasapainoista ja vaihtoehdoiltaan toisensa poissulkevaa skaalaa.

11. Johdattelevat ja epätasapainoiset kysymykset

Kyselytutkimuksen tulokset ovat hyvin konkreettisesti sidoksissa kysymyksenasetteluun. Saadakseen tieteellisesti päteviä tuloksia tutkijan tulee suunnitella lomakkeen kysymykset siten, että kysymyslauseet ja vastausvaihtoehdot muodostuvat riittävän kattaviksi ja tasapainoisiksi, unohtamatta kuitenkaan riittävää tiiviyttä ja sanavalintojen yksinkertaisuutta.

Seuraavassa muutama esimerkki tyypillisesti epätasapainoisista kysymyksistä ja niiden vastausvaihtoehdoista:

11a:

```
Ovatko talvet tulleet lämpimämmiksi viime vuosina?

1 Hyvin paljon lämpimämmiksi
2 Melko paljon lämpimämmiksi
3 Jonkin verran lämpimämmiksi
4 Eivät ole tulleet lämpimämmiksi

Jos vastasit "eivät ole tulleet lämpimämmiksi", perustele kantasi:
_____________________________________________________________________
_____________________________________________________________________
```

11b:

```
Pitäisikö ministerin N.N. erota tehtävästään ..... vuoksi?

1 Ehdottomasti pitäisi
2 Ei pitäisi
3 En osaa sanoa
```

11c:

```
Kuinka riittävästi järjestön X toiminnassa on yleensä ottaen otettu
huomioon jäsenkunnan mielipiteet?

1 Täysin riittävästi
2 Melko riittävästi
3 Ei kovinkaan riittävästi
4 Ei lainkaan riittävästi
```

Sarjan kaikkia kolmea kysymystä yhdistää se, että vastauksen suunta ikään kuin oletetaan ja annetaan jo kysymyslauseessa. Tasapainoinen kysymyksenasettelu keskittyisi käsityksiin talvien keskimääräisestä lämpötilasta, ministerin toimien arvioinnista eroamismahdollisuuden ja työssä jatkamisen näkökulmasta, sekä siihen, missä määrin jäsenkunnan mielipiteet ovat tulleet huomioon otetuiksi järjestön toiminnassa. Vastausvaihtoehtojen tulisi tällöin vaihdella keskimäärää kylmemmästä keskimäärää lämpimämpään, työstä eroamisesta työssä jatkamiseen ja täysin riittävästä huomioon ottamisesta täysin riittämättömään huomioon ottamiseen.

Käytännössäkin valitettavan yleinen johdattelukeino on jättää kysymättä esimerkiksi kyselyn teettäjälle epäedullisia asioita tai painottaa ja täsmentää kysymysten sisältöalueita tarkoitushakuisesti. Paljon yleisempää lienee kuitenkin johdattelevien kysymysten laatiminen tiedostamatta niiden epätasapainoisuutta.

***Nyrkkisääntö 1*:** Asenteita mittaavissa kysymyksissä tasapainoisen vastausasteikon tulee yleensä sisältää kaksi toisilleen vastakohtaista ääripäätä sekä neutraali vastausvaihtoehto.

***Nyrkkisääntö 2*:** Kysymyslauseen tulee yleensä sisältää viittauksia joko molempiin vastausskaalan ääripäihin tai ei kumpaakaan niistä.

Esimerkiksi:

```
Mitä mieltä olette seuraavista väitteistä?
```

tai

Missä määrin olette samaa tai eri mieltä seuraavista väitteistä?

Täysin samaa mieltä
Jokseenkin samaa mieltä
En samaa enkä eri mieltä
Jokseenkin eri mieltä
Täysin eri mieltä
En osaa sanoa

Muuttujien muunnokset

Tutkimusaineiston analyysin yhteydessä tulee usein vastaan tilanne, jossa olemassa olevia muuttujia on tarpeellista jotenkin muuttaa tai niiden pohjalta on tarve luoda kokonaan uusia muuttujia. Esimerkiksi kyselyaineiston yksi muuttuja voi kuvata vastaajan syntymävuotta. Tutkimusraportissa on kuitenkin havainnollisempaa käyttää muuttujana vastaajan ikää vastaushetkellä kuin hänen syntymävuottaan. Näin syntymävuosimuuttuja on muunnettava niin, että se kuvaa vastaajan ikää (ks. verkosta harjoitus 1). Toinen tyypillinen esimerkki tarpeesta muuttaa alkuperäisen muuttujan koodausta on tilanne, jossa muuttujaa pitää luokitella ennen ristiintaulukointia. Tällöin ikämuuttuja (tai syntymävuosimuuttuja) on muunnettava valittuja ikäryhmiä kuvaavaksi muuttujaksi. Tällaista toimenpidettä kutsutaan muuttujan uudelleenkoodaukseksi (*recode*).

Uusien muuttujien luominen

Uuden muuttujan luomisessa otetaan lähtökohdaksi yksi tai useampi olemassa oleva muuttuja. Uuden muuttuja luontitavat vaihtelevat hiukan eri tilasto-ohjelmistoilla. Yleinen periaate on kuitenkin, että luotava muuttuja esitetään matemaattisen kaavan muodossa. Kaava voi sisältää erilaisia matemaattisia operaatiota, kuten yhteen-, vähennys-, kerto- tai jakolaskuja. Myös monimutkaisemmat matemaattiset operaatiot kuten logaritmin tai neliöjuuren ottaminen muuttujista ovat mahdollisia.

Oletetaan, että kunta-aineistoa käyttävä tutkia haluaa lisätä analyysiinsa kunnan asukastiheyttä kuvaavan muuttujan. Aineistossa ei kuitenkaan ole tällaista muuttujaa, mutta tutkijan onneksi siitä löytyvät kunnan väkilukua ja pinta-alaa kuvaavat muuttujat. Uusi kunnan asukastiheyttä kuvaava muuttuja voidaan luoda yksinkertaisesti suorittamalla seuraavanlainen laskuoperaatio: TIHEYS = ASUKASLUKU / PINTA-ALA

Yllä olevassa kaavassa 'TIHEYS' on uusi kunnan asukastiheyttä kuvaava muuttuja ja se luodaan jakamalla 'ASUKASLUKU' -muuttujan arvo 'PINTA-ALA' -muuttujan arvolla.

Summamuuttujia luodessa täytyy myös ymmärtää käytännön tasolla, miten uusia muuttujia luodaan. Summamuuttuja luodaan yhdistämällä useita samaa ilmiötä eri tavoin mittaavia muuttujia. Kuten nimikin antaa ymmärtää, yleisin tapa luoda summamuuttuja on laskea sen pohjana olevat muuttujien arvot yhteen. Esimerkiksi, jos tutkija haluaa muodostaa summamuuttujan viiteen eri kysymykseen saaduista vastauksista (KYS1, KYS2, KYS3, KYS4, KYS5), tapahtuu se seuraavasti: SUMMA = KYS1 + KYS2 + KYS3 + KYS4 + KYS5

Kaavassa 'SUMMA' on uuden summamuuttujan nimi. Tässä esimerkissä uusi summamuuttuja ei ole samalla asteikolla kuin alkuperäiset kysymykset. Joskus voi olla havainnollisempaa, että luotu summamuuttuja vaihtelisi samalla välillä kuin ne alkuperäiset osiot, josta summamuuttuja muodostettiin. Jos oletetaan, että esimerkin kysymykset KYS1-5 saavat jokainen arvoja välillä 1-5, on summamuuttujan mahdollinen vaihteluväli 5-25. Vaihtoehtoinen tapa muodostaa summamuuttuja on SUMMA = (KYS1 + KYS2 + KYS3 + KYS4 + KYS5)/5

Näin muodostettu summamuuttuja sisältää periaatteessa saman informaation kuin aikaisempi summamuuttuja, mutta sen vaihteluväli on sama kuin alkuperäisillä kysymyksillä. Tämä helpottaa summamuuttujan arvojen tulkintaa, koska nämä arvot voi suoraan suhteuttaa alkuperäisten kysymysten vastausvaihtoehtoihin.

Kolmas esimerkki tilanteesta, jossa uuden muuttujan luominen on tarpeen on jo edellä mainittu vastaajan syntymävuoden muunnos vastaajan vastaushetken iäksi. 1990-luvun puolenvälin kansainvälisen World Values Surveyn kyselyn Suomen osa-aineisto kerättiin vuonna 1996 (ks. verkosta aineiston FSD0153 kuvaus). Kyselyssä kysyttiin vastaajan syntymävuotta

(muuttuja V215) tai tarkalleen ottaen syntymävuoden kahta viimeistä numeroa (oletuksena oli, että kukaan vastaajista ei ole syntynyt 1800-luvulla). Vastaajan ikä saadaan selville luomalla uusi muuttuja seuraavalla tavalla: IKÄ = 96 - SYNTYMÄVUOSI

Jos vastaaja on ilmoittanut syntymävuodekseen vuoden 70 saa oheisen kaavan mukaan ikämuuttujan arvoksi 26 jne. Oheisella laskukaavalla tulee vastaajien ikään tietysti pieniä virheitä riippuen siitä, mihin vuodenaikaan kysely tehtiin. Suurimmillaankin nämä virheet ovat alle vuoden, joten niillä tuskin on suurta vaikutusta tulosten kannalta.

Lisäksi uusien muuttujien luomista tarvitaan tilanteissa, joissa alkuperäisen muuttujan jakauma on sellainen, että muuttujan käyttö sellaisenaan ei ole järkevää empiirisessä analyysissa. Tällaisessa tapauksessa muuttujalle voidaan tehdä muunnos, jonka jälkeen sen jakauma noudattaa lähemmin normaalijakaumaa. Usein käytettyjä muunnoksia tällaisessa yhteydessä ovat esimerkiksi logaritmin tai neliöjuuren ottaminen alkuperäisestä muuttujasta.

Muuttujien uudelleenkoodaus

Muuttujien uudelleenkoodaus tarkoittaa sitä, että alkuperäisen muuttujan arvot vaihdetaan uusiin arvoihin. Esimerkiksi aineistossa voi vastaajan sukupuoli olla koodattu niin, että mies saa arvon yksi ja nainen arvon kaksi. Joissain tapauksissa (esimerkiksi regressioanalyysin yhteydessä) on kuitenkin järkevää muuttaa muuttujan koodausta niin, että toinen sukupuoli saa arvon nolla ja toinen arvon yksi. Tällaista muutosta kutsutaan uudelleenkoodaukseksi.

Uudelleenkoodaus on mahdollista tehdä tilasto-ohjelmistojen avulla kahdella eri tavalla. Ensimmäinen vaihtoehto on, että alkuperäisen muuttujan koodaus muutetaan uudeksi (*recode into same variable*). Tällöin kuitenkin menetetään muuttujan alkuperäiset arvon. Toinen vaihtoehto on muodostaa uusi muuttuja, joka sisältää uudet muuttujan arvot (*recode into different variable*). Käytännössä jälkimmäinen menettely on turvallisempi, koska virheen sattuessa alkuperäinen muuttuja on vielä tallessa, ja virhe voidaan korjata.

Uudelleenkoodausta tarvitaan esimerkiksi silloin, kun halutaan muuttaa alkuperäisen muuttujan "suuntaa" (eli pieneksi arvoksi koodatut vastausvaihtoehdot halutaan muuttaa suuriksi arvoiksi ja päinvastoin). Tämä on erityisen tärkeää summamuuttujien luomisen yhteydessä. Summamuuttujaa tehtäessä täytyy kaikki käytettävät muuttujat koodata siten, että suuret muuttujan arvot kuvaavat jokaisen muuttujan osalta samansuuntaisesti mitattavaa asiaa. Muutoin summamuuttuja on virheellinen.

Joskus muuttujan 'suunta' kannattaa muuttaa jo pelkästään havainnollisuuden vuoksi. Esimerkiksi yhdessä World Values -kyselyn osassa tiedustellaan vastaajan terveydentilaa (ks. osaWVS-aineiston frekvenssit FSD0153, muuttuja V11). Vastaajat saavat kuvailla omaa terveyttään seuraavin vaihtoehdoin: "erittäin hyvä", "melko hyvä", "kohtalainen", "melko huono" ja "erittäin huono". Vastaukset on koodattu niin, että ne jotka pitävät terveyttään erittäin hyvänä saavat arvon 1, melko hyvänä arvon 2, kohtalaisena arvon 3, melko huonona arvon 4 ja erittäin huonona arvon 5. Muuttujaa voisi kuvata nimellä 'terveysmuuttuja', mutta tämä nimi olisi harhaanjohtava, koska muuttujan suuret arvot kuvaavat itse asiassa huonoa terveydentilaa. Uudelleenkodaus tekisi muuttujasta havainnollisemman. Tällöin suuret arvot kuvastaisivat hyvää terveydentilaa. Tämä tapahtuu niin, että tilasto-ohjelmiston avulla luodaan uusi 'terveys' -muuttuja, jossa alkuperäisen muuttujan arvo 1 korvataan arvolla 5, arvo 2 korvataan arvolla 4 jne.

Uudelleenkoodauksen käyttö on myös erittäin yleistä silloin, kun välimatka- tai suhdeasteikolla mitattu muuttuja (katso aiemmasta luvusta muuttujien mittaustaso) halutaan muuttaa luokitelluksi järjestelyasteikolliseksi muuttujaksi. Esimerkiksi luokittelematonta ikämuuttujaa ei useinkaan voi käyttää ristiintaulukoinnissa käytännön syistä. Ikämuuttuja voidaan kuitenkin uudelleenkoodauksen avulla muuntaa ikäluokkamuuttujaksi, jonka arvot kuvastavat vastaajan kuulumista tiettyyn ikäryhmään. Esimerkiksi vastaajan ikä voidaan uudelleenkoodata kolmeen luokkaan seuraavalla tavalla: kaikki alle 35-vuotiaat vastaajat saavat

arvon 1, 35-59-vuotiaat saavat arvon 2 ja kaikki yli 59-vuotiaat arvon 3. Tätä uudelleenluokiteltua muuttujaa voidaan käyttää ristiintaulukoinnissa (ks. esimerkki ristiintaulukon elaboroinnista).

Summamuuttuja

Summamuuttujaksi nimitetään muuttujaa, jonka arvot on saatu **laskemalla yhteen** useiden erillisten, mutta samaa ilmiötä mittaavien muuttujien arvot.

Kyselytutkimuksessa summamuuttujia käytetään usein asenneväittämiin saatujen vastausten yhdistämisessä. Asenneväittämillä tutkitaan vastaajien mielipidettä tietystä asiasta. Väittämistä saadaan tilastolliset muuttujat, joilla laskutoimitukset ovat mahdollisia. Tiivistetty kuva asenteista saadaan summamuuttujan avulla.

Samaan aihealueeseen asennoitumista voidaan mitata yksittäisessä tutkimuksessa jopa useilla kymmenillä väittämillä. Samalla kysymyspatteristolla voidaan selvittää hyvin moneen eri asiaan liittyviä mielipiteitä tai faktatietoja. Tällöin voidaan muodostaa useita, eri asioita indikoivia summamuuttujia. Yhteenlaskettavat muuttujat ovat mitta-asteikoltaan yleensä järjestystasoisia. Muuttujien yhteenlaskemista käytetään mm. Likert-asteikollisilla, kuten seuraavassa esimerkissä, Guttman-asteikollisilla eli kasautuvilla tai 0-1-muuttujilla.

Esimerkiksi tilastotieteeseen liittyvillä asenneväittämillä voidaan mitata motivoituneisuutta tilastotieteen opintoihin. Mitä motivoituneempia tilastotieteen kurssille osallistuvat henkilöt ovat, sen helpommin he todennäköisesti omaksuvat opetettavat asiat. Motivoituneisuus operationalisoidaan sitä ilmentäviksi väittämiksi. Seuraavassa taulukossa on muutamia väittämiä, joilla voidaan kartoittaa vastaajien asennetta tilastotieteeseen. Kysely on suunnattu opiskelijoille, jotka tuntevat käsitteet kvantitatiivinen ja kvalitatiivinen tutkimus.

Taulukko 1. Tilastotieteen opiskelumotivaatiota mittaavia asenneväittämiä.

	1 täysin eri mieltä	2 joks. eri mieltä	3 joks. samaa mieltä	4 täysin samaa mieltä
Tilastotiede on hyödyllistä				
Jokaisen yliopistotutkinnon suorittaneen tulee tietää tilastotieteen perusasiat				
Tilasto-ohjelmisto on erinomainen apuväline kvantitatiivisessa tutkimuksessa				
Olisin valmis hyödyntämään tilastotiedettä jopa kvalitatiivisessa tutkimuksessa				

Asenneväittämillä saatu tieto opiskelumotivaatiosta tiivistyy summamuuttujassa, joka muodostetaan laskemalla yhteen muuttujien numeeriset koodit. Esimerkkimme vastausten koodaus: 1=täysin eri mieltä, 2=jokseenkin eri mieltä, 3=jokseenkin samaa mieltä, 4=täysin samaa mieltä. Henkilö, joka on kaikista neljästä väittämästä eri mieltä (1), saa motivaatiopistemääräkseen 4 ja joka on kaikkien väittämien kanssa samaa mieltä (4), saa motivaatiopistemääräkseen 16. Motivaatio-muuttujan arvot voivat siis vaihdella välillä 4-16: pienin arvo 4 kuvaa heikointa ja suurin arvo 16 voimakkainta motivoituneisuutta. Muut arvot voidaan tulkita suhteessa minimiin ja maksimiin sekä alkuperäisten muuttujien koodeihin. Seuraavassa taulukossa on konkreettinen esimerkki viiden opiskelijan antamista vastauksista ja summamuuttujan arvojen laskemisesta.

Taulukko 2. Alkuperäisten muuttujien ja summamuuttujan arvot.

	Tilastotiede on hyödyllistä	Tutkinnon suorittaneen tulee tietää tilasto-tieteen perusasiat	Tilasto-ohjelmisto on erin-omainen apuväline kvantitat. tutk:ssa	Olen valmis hyödyntä-mään tilastotie-dettä jopa kvalitat. tutk:ssani		summa-muuttuja asteikolla [4,16]		summa-muuttuja alkuperäiselle asteikolle muutettuna
opiskelija 1	1 =täysin eri mieltä	1 =täysin eri mieltä	1 =täysin eri mieltä	1 =täysin eri mieltä		4		1= ei motivoitunut
opiskelija 2	2 =jokseenkin eri mieltä	3 =jokseenkin samaa mieltä	3 =jokseenkin samaa mieltä	1 =täysin eri mieltä	⇒	9	⇒	2=ei juurikaan motivoitunut
opiskelija 3	4 =täysin samaa mieltä	4 =täysin samaa mieltä	4 =täysin samaa mieltä	4 =täysin samaa mieltä		16		4=erittäin hyvin motivoitunut
opiskelija 4	3 =jokseenkin samaa mieltä	3 =jokseenkin samaa mieltä	3 =jokseenkin samaa mieltä	2 =jokseenkin eri mieltä		11		3=jonkin verran motivoitunut
opiskelija 5	3 =jokseenkin samaa mieltä	3 =jokseenkin samaa mieltä	4 =täysin samaa mieltä	5 , =puuttuva tieto		, =puuttuva tieto		3= jonkin verran motivoitunut

Summamuuttujan saamat arvot voidaan palauttaa samalle vaihteluvälille (1-4), kuin alkuperäisten yhteenlaskettavien muuttujien. Tällöin muuttujan numeerisille koodeille voidaan haluttaessa keksiä sanalliset ilmaisut: 1=ei motivoitunut, 2=ei juurikaan motivoitunut, 3=jonkin verran motivoitunut, 4=erittäin hyvin motivoitunut. Alkuperäiselle asteikolle joudutaan palaamaan silloin, kun aineistossa on paljon puuttuvaa tietoa. Muutoin eri opiskelijoiden saamat summamuuttujan arvot eivät ole keskenään vertailukelpoisia. Kun puuttuva tieto korvataan nollalla, tilasto-ohjelma pystyy laskemaan summamuuttujan arvon myös niille henkilöille, jotka eivät ole ilmaisseet mielipidettään kaikkiin summattaviin väittämiin. Tästä on harjoitusesimerkki tietovarannon SPSS-osiossa. (Ks. verkkoversiosta SPSS-harjoitus 1c.)

Asenteita tutkittaessa on kiinnitettävä huomiota väittämien ja vastausvaihtoehtojen "suuntaan". 'Tilastotiede on hyödyllistä' on myönteinen väite. Asennetta voidaan mitata myös kielteisillä väittämillä, kuten 'Tilastotieteen opiskelu on ajanhukkaa'. Summamuuttujaa muodostettaessa on huolehdittava siitä, että yhdistettävien muuttujien koodaus on yhteensopiva.

Jos em. väittämä 'Tilastotieteen opiskelu on ajanhukkaa', halutaan liittää po. summamuuttujaan, joka kuvaa motivoituneisuutta opiskeluun, silloin väittämän vastausvaihtoehdot tulee koodata uudelleen käänteisesti: 1=täysin samaa mieltä, ..., 4=täysin eri mieltä. Tässä esimerkissä koodimuutoksia halutaan tehdä mahdollisimman vähän, joten kielteisen väittämän koodaus käännetään myönteisiä väittämiä vastaavaksi: 1→4, 2→3, 3→2, 4→1 (ks. taulukko 3). Tällöin kielteisen väittämän ja myönteisten väittämien vastausten koodit ovat sopusoinnussa: Pieni luku tarkoittaa kaikissa väittämissä kielteistä asennoitumista ja huonoa motivaatiota, suuri luku hyvää motivaatiota. Yhteenlaskettavien väittämien koodeilla on näin ollen looginen tulkinta ja tällöin myös summamuuttujaa voidaan tulkita. Jos viisi väittämää

lasketaan yhteen, on pienin mahdollinen summa 5 ja suurin 20. Pieni luku tarkoittaa huonoa motivaatiota, suuri hyvää. Yhteenlaskettavien muuttujien ja valmiin summamuuttujan koodausta voidaan muuttaa hyvinkin eri tavoilla, kunhan logiikka on selkeä eikä alkuperäinen informaatio muutu; esimerkiksi siten, että summamuuttujan asteikko saadaan alkamaan nollasta. (Ks. summamuuttujan muodostamiseen liittyvät harjoitukset tietovarannon SPSS-osiosta.)

Taulukko 3. Samansuuntainen koodaus.

Tilastotiede on hyödyllistä		Tilastotieteen opiskelu on ajanhukkaa **Alkuperäinen koodaus**		Tilastotieteen opiskelu on ajanhukkaa **Käännetty koodaus**
1 =täysin eri mieltä, KIELTEINEN		1 =täysin eri mieltä, MYÖNTEINEN		4 =täysin eri mieltä, MYÖNTEINEN
2 =jokseenkin eri mieltä		2 =jokseenkin eri mieltä	⇒	3 =jokseenkin eri mieltä
3 =jokseenkin samaa mieltä		3 =jokseenkin samaa mieltä		2 =jokseenkin samaa mieltä
4 =täysin samaa mieltä, MYÖNTEINEN		4 =täysin samaa mieltä, KIELTEINEN		1 =täysin samaa mieltä, KIELTEINEN

Saman ilmiöalueen mittaamisen useilla erilaisilla kysymyksillä voidaan nähdä parantavan mittarin reliabiliteettia. Tällöin satunnaisvirheen vaikutus pienenee. On kuitenkin mahdotonta määritellä yleispätevästi, kuinka monta muuttujaa summamuuttujaan tarvitaan. Muuttujien valinta summamuuttujaan voi olla täysin sisällöllinen, mutta sitä voidaan perustella muuttujien välisillä keskinäisillä korrelaatioilla. Faktorianalyysiä voidaan hyödyntää valittaessa yhteenlaskettavia muuttujia summamuuttujaan. Olennaista on, että yhdistettävät muuttujat ovat sisällöllisesti järkeviä. On syytä kiinnittää huomiota myös siihen, etteivät ne ole sisällöllisesti päällekkäisiä.

Summamuuttujan jakaumaa voidaan tarkastella frekvenssitaulukolla. Tiivistetysti sitä voidaan kuvata myös yksiulotteisen jakauman tunnusluvuilla ja graafisesti mm. histogrammina tai pylväsdiagrammina. Kun summamuuttuja on riippuvuustarkasteluissa selitettävänä muuttujana, sille voidaan laskea ja esittää graafisesti tunnuslukuja (esim. mediaanipylväät) selittävän muuttujan (esim. sukupuoli) ryhmissä. Laajalla vaihteluvälillä arvoja saavan summamuuttujan graafiseen esittämiseen soveltuu luontevasti laatikko-jana -esitys. Kun summamuuttuja on luokiteltu tai palautettu alkuperäiselle asteikolle, riippuvuustarkasteluissa voidaan käyttää ristiintaulukointia ja 100% pylväskuvioita.

-- HARJOITUSTEHTÄVÄ --

Tehtävä 1. World Value Survey -aineistoon liittyvä harjoitus: Tasa-arvo
Valitse aineistosta tasa-arvoajatusta mittaavat muuttujat.

- Mieti yksittäisiä muuttujia sisällöllisesti, mitä ne mittaavat?
- Mieti, kuinka laajasti kyseiset muuttujat mielestäsi mittaavat tasa-arvon kannattamista. Ovatko kyseiset väittämät mielestäsi hyviä mittaamaan tasa-arvoajatusta?
- Mitä näistä kolmesta muuttujasta muodostettava summamuuttuja kertoo. Mikä on sen "idea"?
- Valitse tapa, jolla muodostat SPSS-ohjelmistossa summamuuttujan. Kerro siitä: Miten käsittelet esim. "en osaa sanoa" vaihtoehdot ja teetkö koodimuutoksia ym.

Puuttuvat havainnot

Lähes kaikissa määrällisissä aineistoissa on havaintoyksikköjä, joista ei syystä tai toisesta ole pystytty mittaamaan kaikkien muuttujien arvoja. Tällaisia tapauksia kutsutaan puuttuviksi havainnoiksi. Niillä voi olla suuri merkitys aineiston analyysin kannalta. Jos puuttuvat havainnot poistetaan analyysista, pienenee havaintoyksikköjen määrä ja saatujen tulosten tarkkuus voi kärsiä. Vielä suurempi ongelma on silloin, jos puuttuvat havainnot eivät ole jakautuneet satunnaisesti havaintoyksikköjen kesken, vaan joissakin ryhmissä niitä on huomattavasti enemmän kuin toisissa. Tilanne saattaa pahimmassa tapauksessa vääristää analyysin tuloksia merkittävästi. Näiden syiden vuoksi puuttuvien havaintojen käsittelyyn kannattaa perehtyä ennen varsinaisen analyysin aloittamista. Seuraavassa asiaa käsitellään erityisesti kyselytutkimusten näkökulmasta.

Havaintojen puuttumiselle voi olla useita eri syitä. Usein kyselytutkimuksissa kaikkien vastaajien ei ole edes tarkoitus vastata kaikkiin kysymyksiin. Esimerkiksi jos vastaaja ilmoittaa, ettei hänellä ole lapsia, ei hänen tarvitse vastata kysymyksiin, joissa tiedustellaan lasten ikää. Tällaiset puuttuvat havainnot ovat jo lomakkeen suunnitteluvaiheessa tiedossa, eivätkä ne aiheuta suuria ongelmia aineiston analyysissa. Sen sijaan muut mahdolliset puuttuvien havaintojen syyt päänvaivaa tutkijalle. Vastaamatta voidaan jättää epähuomiossa tai viitseliäisyyden puutteessa. Joskus vastaajat kieltäytyvät vastaamasta johonkin tiettyyn kysymykseen. Joskus kysymys voi taas käsitellä niin arkaluonteisia asioita, että kaikki vastaajat eivät halua ilmoittaa mielipidettään. Toisinaan vastaus voi olla niin epämääräinen, ettei siitä yksikäsitteisesti selviä, mitä vastaaja on tarkoittanut (esimerkiksi kirjoitetusta numerosta ei saa selvää). Eikä puuttuvan havainnon syy ole aina edes tiedossa. Se voi johtua myös haastattelijan tai aineiston koodaajan virheestä.

Kyselytutkimuksissa vaihtoehdot 'en osaa sanoa', 'en halua sanoa' tai 'en tiedä' aiheuttavat joskus ongelmia aineiston jatkoanalyysille. Usein näitä vastausvaihtoehtoja käsitellään puuttuvina tietoina. Tämä ratkaisu ei välttämättä ole perusteltu, jos tällaisia vastauksia on paljon. Analyysin tulokset voivat muuttua, jos puuttuvat vastaukset eivät ole jakautuneet sattumanvaraisesti vastaajien kesken, vaan niiden yleisyys vaihtelee tarkasteltavien ryhmien mukaan. Lisäksi vastausten 'en osaa sanoa' tai 'en tiedä' analyysi voi olla mielenkiintoinen tutkimusongelman kannalta. Jos tietyn tyyppisillä vastaajilla ei ole mielipidettä jostain yhteiskunnallisesta ilmiöstä, voi tämä tieto olla itsessään arvokas tulkittaessa vastaajien suhtautumista tutkittavaan ilmiöön.

Puuttuvien havaintojen käsitteleminen

Koska puuttuvat havainnot voivat myös vääristää analyysin tuloksia, täytyy niiden käsittelyyn kiinnittää erityistä huomiota. Ongelman korjaamiseksi tai ainakin lievittämiseksi on esitetty useita erilaisia menetelmiä. Yleispätevää toimintasääntöä ei ole, vaan soveltuva ratkaisu täytyy valita tapauskohtaisesti. Seuraava toimenpidejaottelu perustuu Hertelin (1976) asiaa käsittelevään artikkeliin.

Puuttuvien havaintojen poistaminen

Yksinkertaisin lähestymistapa puuttuvien havaintojen ongelmaan on poistaa analyysista kaikki havaintoyksiköt, joista on puuttuvia tietoja yhdessäkin analyysiin sisälletyissä muuttujissa. Englanninkielisissä tilasto-ohjelmissa tätä toimenpidettä kutsutaan nimellä listwise deletion. Ongelmana tässä lähestymistavassa on, että se voi pienentää otoksen kokoa huomattavasti. Tämä tulee erityisen selvästi esille monimuuttujamenetelmiä sovellettaessa, jolloin analyysissa voi olla mukana useita, joskus jopa kymmeniä, muuttujia. Analyysin ulkopuolelle jäävät kaikki havaintoyksiköt, joista puuttuu yksikin arvo jostakin analyysissa

mukana olevasta muuttujasta. Jos puuttuvat havainnot keskittyvät kuitenkin vain pieneen osaan havaintoyksiköistä, voi näiden poistaminen analyysista olla järkevää. Ennen tätä toimenpidettä kannattaa tarkistaa (esimerkiksi ristiintaulukoinnin avulla) ovatko puuttuvat havainnot jakautuneet satunnaisesti tutkimusongelman kannalta mielenkiintoisten ryhmien välillä, vai keskittyvätkö ne joihinkin erityisiin ryhmiin. Jälkimmäisessä tapauksessa puuttuvien havaintojen poistaminen analyysista voi vääristää lopputuloksia.

Muuttujien poistaminen

Jos jostakin muuttujasta puuttuu huomattava määrä havaintoja, kannattaa pohtia koko muuttujan pudottamista pois analyysista. Tämä on suositeltavaa ainakin silloin, kun aineistossa on muita muuttujia, jotka mittaavat samaa asiaa. Hyvänä puolena tässä ratkaisussa on, että havaintoyksikköjen määrä ei toimenpiteen seurauksena vähene. Ratkaisua ei tietenkään voi suositella silloin, kun muuttuja on tutkimuskysymyksen kannalta tärkeä ja sen poisjättäminen vaikeuttaa tutkimusongelman ratkaisua.

Puuttuvien havaintojen parittainen poistaminen

Useat monimuuttujamenetelmät perustuvat muuttujien kovarianssi- tai korrelaatiomatriisin analysoinnille (esimerkiksi faktorianalyysi ja regressioanalyysi). Tällaisessa tapauksessa puuttuvia havaintoja voidaan poistaa analyysista ns. parittaisesti (pairwise deletion). Tämä tarkoittaa sitä, että korrelaatiomatriisia laskettaessa otetaan huomioon kaikki ne havaintoyksiköt, joista on tiedot niillä kahdella muuttujalla, joista korrelaatio lasketaan. Näin saadussa korrelaatiomatriisissa jokainen korrelaatioarvo voi perustua erilaiseen havaintoyksikköjen määrään. Tämänkin menetelmän seurauksena aineisto pienenee, mutta ei läheskään yhtä paljon verrattuna tilanteeseen, jossa kaikki puuttuvia tietoja sisältävät havaintoyksiköt poistettaisiin analyysista.

Keskiarvon käyttö

Jos puuttuvia havaintoja ei voida poistaa, yksi vaihtoehto on koodata puuttuvien muuttujan arvojen tilalle jokin ennalta päätetty arvo ja sisällyttää siten kaikki havaintoyksiköt analyysiin. Yleensä puuttuvien havaintojen tilalle koodataan muuttujan keskiarvo. Keskiarvon käyttöä perustellaan sillä, että jos tutkijalla ei ole etukäteen mitään tietoa puuttuvan havainnon arvosta, paras "arvaus" täksi arvoksi on juuri koko aineiston keskiarvo. Ilmeinen etu tämän menetelmän käytössä on, että se ei pienennä aineiston kokoa. Huono puoli on, että keskiarvojen käyttö johtaa muuttujien hajonnan pienenemiseen. Jos puuttuvia havaintoja on paljon, voi tällä olla suuri merkitys jatkoanalyysin kannalta. Käytännössä muuttujien hajonnan pienenemisestä seuraa, että niiden välinen korrelaatio pienenee. Näin keskiarvojen käyttö puuttuvien havaintojen tilalla tekee monimuuttujamenetelmien tuloksista "konservatiivisempia", eli havaitut yhteydet muuttujien välillä eivät ole niin vahvoja, kuin jos puuttuvia havaintoja olisi aineistossa vähemmän.

Ryhmäkeskiarvojen käyttö

Puuttuvat muuttujan arvot voidaan korvata koko muuttujan keskiarvon sijasta myös ryhmäkeskiarvoilla. Jos esimerkiksi vastaajien koulutustaso on mitattu kolmiluokkaisella mittarilla, jaetaan aineisto näihin kolmeen ryhmään ja lasketaan jokaiselle ryhmälle oma keskiarvo kiinnostuksen kohteena olevasta muuttujasta. Tämän jälkeen puuttuvat havainnot korvataan näillä ryhmäkeskiarvoilla. Jos vastaaja kuuluu akateemisen tutkinnon suorittaneiden ryhmään ja hänellä on jossain kysymyksessä puuttuva havainto, koodataan puuttuvan havainnon tilalle akateemisten tällä muuttujalla saama keskiarvo jne. Tämän menetelmän ongelma on, että se korostaa ryhmien sisäistä samankaltaisuutta ja ryhmien välisiä eroja. Seuraukset ovat

päinvastaiset kuin koko muuttujan keskiarvojen käytössä puuttuvien havaintojen tilalla. Ryhmäkeskiarvojen käyttö voi vääristää tuloksia kasvattamalla muuttujien välisiä korrelaatioita.

Muita tapoja

Edellä esiteltiin yleisimpiä tapoja käsitellä puuttuvia havaintoja. Niiden lisäksi on muitakin mahdollisuuksia. Yksi tapa on jakaa aineisto ryhmiin (esimerkiksi miehiin ja naisiin) ja koodata puuttuvan arvon kohdalle havaintomatriisissa edellisen havainnon arvo. Tämä tarkoittaa, että puuttuvien arvojen tilalle koodataan useita eri arvoja, ei ainoastaan keskiarvoja. Menetelmän etu on, että se ei vähennä muuttujien hajontaa niin kuin pelkkien keskiarvojen käyttö. Myös regressioanalyysia voidaan käyttää puuttuvien havaintojen "oikeiden" arvojen löytämiseksi. Tämä menetelmä on monimutkaisempi kuin edellä esitellyt vaihtoehdot.

Puuttuvien havaintojen koodaaminen

Puuttuvien havaintojen muodostamien ongelmien ratkaisemiseen ei ole helppo antaa yleispäteviä toimintaohjeita. Jos puuttuvia havaintoja ei poisteta analyysista, ne on koodattava havaintomatriisiin siten, että niiden erityisluonne tulee selvästi esille. Samoin jos puuttuvan havainnon syy on selvillä, kannattaa eri syistä johtuvat puuttuvat havainnot koodata eri koodeilla.

Periaatteessa puuttuvan havainnon voi koodata millä koodilla tahansa. Valinta riippuu kuitenkin siitä, millainen on muuttujan arvojen alkuperäinen vaihteluväli. Puuttuvien havaintojen koodi kannattaa joka tapauksessa valita niin, että se eroaa selkeästi muuttuja saamista "oikeista" arvoista. Usein puuttuvan havainnon koodina käytetään arvoja 9, 99 tai 999 edellyttäen, että ne eivät ole muuttujan valideja arvoja. Myös nollaa käytetään usein puuttuvan tiedon arvona, mutta tällöinkin tulee kiinnittää erityistä huomiota siihen, ettei '0' ole sisällöllisesti hyväksyttävä tieto (esimerkiksi vastaaja ei ole osallistunut kertaakaan kysyttyyn toimintaan).

Ennen varsinaisen tilastoanalyysin aloittamista tulee ehdottomasti tarkistaa muuttujien puuttuvien havaintojen ja tietojen koodaus ja onko tilasto-ohjelmassa määritelty puuttuvien havaintojen koodi niin, että niitä ei oteta automaattisesti mukaan analyysiin. Jos esimerkiksi perheen lapsien määrää mittaavassa muuttujassa puuttuva havainto on koodattu arvolla 999 ja näitä havaintoja ei ole muistettu poistaa analyysista, voi perheiden keskimääräinen lapsiluku olla yllättävän suuri.

Lähteet

- Hertel, Bradley R. (1976): *Minimizing Error Variance Introduced by Missing Data Routines in Survey Analysis*. Sociological Methods & Research 4: 459-474.

Kyselyaineiston havaintojen painottaminen

Kyselyaineiston havaintoja voi olla tarkoituksenmukaista painottaa, mikäli otoksen edustavuutta tutkiva katoanalyysi osoittaa aineistosta systemaattisia vinoutumia. Katoanalyysissa vertaillaan otoksen ja perusjoukon vastaavuutta niiden keskeisten rakennetekijöiden osalta, joista tiedot ovat saatavilla. Henkilöaineistoissa tällaisia ovat yleensä vastaajien sukupuoli, ikä, asuinpaikka-/alue sekä koulutustaso tai ammatti. Lisäksi katoanalyysiin on tarpeen ja mahdollisuuksien mukaan sisällytettävä vertailuja sellaisten tekijöiden suhteen, joilla tiedetään olevanmerkittävää vaikutusta tutkittaviin aihealueisiin.

Otosaineiston sosiodemografinen poikkeavuus perusjoukosta ei silti välttämättä merkitse sitä, että tutkittavia aiheita koskevat tulokset eivät olisi yleistettävissä perusjoukkoon. Saattaa olla, että otoksessa yli- tai aliedustetut ryhmät eivät poikkea merkittävästi keskimääräisistä tuloksista kiinnostuksen kohteena olevien ilmiöiden osalta.

Usein joidenkin ryhmien ali- tai yliedustus otoksessa kuitenkin on ongelma tulosten yleistettävyydelle. Asian ratkaisemiseksi voidaan käyttää ns. jälkiositusta (*post-stratification*), joka on on yleinen painottamistekniikka kyselytutkimuksissa. Siinä tutkimuksen muuttujia painotetaan populaation jakaumalla.

Milloin painottaa aineistoa?

Esimerkiksi suuri vastauskato, väärä otosasetelma tai puuttuvien havaintojen määrä voivat aiheuttaa kyselytutkimuksissa vinoutumia otosaineiston jakaumaan. Systemaattiseten vinoutumien vuoksi otos ei enää ole satunnainen. Tällöin aineistoa voidaan painottaa, jotta otosaineisto kuvaisi paremmin otospopulaatiota. Kyselytutkimus on saatettu tehdä esimerkiksi niin, että otospopulaatio on jaettu ryhmiin ja näistä ryhmistä on valittu sama määrä haastateltaviksi. Jos tiedetään tutkittavan populaation jakauma ennalta (esim. ikä-, sukupuoli- ja ammattirakenne), on suotavaa että jo aineiston keräämisvaiheessa otetaan huomioon otospopulaation rakenne (ks. otantamenetelmät) ja valitaan haastateltavien määrät populaation rakenteen mukaisesti. Jos aineisto on jo kerätty, muuttujien jakaumavirhettä voi korjata painottamalla aineistoa niin, että se kuvaa mahdollisimman tarkasti ennakoitua jakaumaa. Aineiston painottaminen ja painojen käyttäminen analyyseissä estää tiettyjen ryhmien yli- tai aliedustukset.

Oletetaan, että suomalainen valtakunnallinen henkilöaineisto on kerätty vuonna 2000 haastattelemalla 15-64 vuotiaita miehiä ja naisia eri ikäryhmissä (15-19, 20-29, 30-49, 50-64). Jokaisen ryhmän koko on 50 henkilöä (15-19 vuotiaita miehiä on 50, 15-19 naisia on 50, ..) eli koko aineistossa on yhteensä 400 havaintoa. Aineisto ei kuvaa suomalaisen väestön oikeaa ikä- ja sukupuolirakennetta, koska ko. ikä- ja sukupuoliryhmien osuudet suomalaisessa 15-64 vuotiaassa väestössä eivät ole yhtä suuria. Suomalaisten ikä- ja sukupuolirakenne vuonna 2000 saadaan laskettua esimerkiksi Tilastokeskuksen Suomi lukuina: väestö -taulukon avulla (http://www.stat.fi/tk/tp/tasku/taskus_vaesto.html).

Taulukko 1. Suomalaisten ikä- ja sukupuolirakenne vuonna 2000.

Ikäryhmä	Mies	Nainen	Yhteensä
15-19	4.9%	4.7%	9.6%
20-29	9.3%	8.9%	18.3%
30-49	22.1%	21.4%	43.5%
50-64	14.2%	14.5%	28.7%
Yhteensä	50.5%	49.5%	100.0%

Kuinka painot lasketaan

Jos aineistoa painotetaan vain yhden muuttujan perusteella, lasketaan ensin aineistosta ko. muuttujan frekvenssijakauma. Lisäksi täytyy tietää koko aineiston havaintojen lukumäärä ja luonnollisesti myös tutkittavan populaation jakauma. Esimerkkiaineiston perusteella sukupuolijakauma on siis 50% miehiä (n=200) ja 50% (n=200) naisia. Tilastokeskuksen mukaan sukupuolijakauma tutkittavalle populaatiolle olisi 50.5% miehiä ja 49.5% naisia. Painot w_1 saadaan laskettua kaavalla:

$$w_1 = \frac{NK_1}{n_1}$$

missä
N on havaintojen lukumäärä
K_i on toivottu jakauma ryhmässä i
n_i on havaittu jakauma ryhmässä i

Seuraavasta ilmenenee laskenta esimerkkitapaukselle.

Painotettava ryhmä	Aineiston koko (N)	Toivottu jakauma (K_i)	N * K_i	Havaittu jakauma (n_i)	Paino (w_i)
Mies	400	0.505	202	200	1.01
Nainen	400	0.495	198	200	0.99

Sukupuolijakauma aineistossa on lähellä oikeaa, joten myös sille lasketut painokertoimet ovat lähellä arvoa 1.

Useamman muuttujan tapauksessa painotettaville ryhmille lasketaan jakaumat aineistosta ristiintaulukoinnin avulla. Esimerkiksi painotus sukupuolen ja ikäryhmien mukaan olisi seuraava:

Painotettava ryhmä	Aineiston koko (N)	Toivottu jakauma (K_i)	N * K_i	Havaittu jakauma (n_i)	Paino (w_i)
Mies 15 - 19	400	0.049	19.6	50	0.392
Mies 20 - 34	400	0.093	37.2	50	0.744
Mies 35 - 49	400	0.221	88.4	50	1.768
Mies 50 - 64	400	0.142	56.8	50	1.136
Nainen 15 - 19	400	0.047	18.8	50	0.376
Nainen 20 - 34	400	0.089	35.6	50	0.712
Nainen 35 - 49	400	0.212	84.8	50	1.696
Nainen 50 - 64	400	0.145	58.0	50	1.160

Painokertoimien käyttö

Kuvatulla menetelmällä painokertoimet voi laskea vain sellaisille havainnoille, joissa painotettavia ryhmiä kuvaavat muuttujat eivät saa puuttuvia arvoja. Havainnot, joiden painokerroin on puuttuva, poistetaan analyyseista. Mikäli haluat kiertää tämän rajoitteen, tutustu kehittyneempiin painotusmenetelmiin lisätiedoissa mainituissa artikkeleissa.

Painokertoimia voi käyttää kaikissa aineistoon liittyvissä analyyseissä. Kun painokertoimet on laskettu painomuuttujiin käytössä olevalla tilasto-ohjelmalla, voidaan aineiston painotus ottaa käyttöön. Tämä tapahtuu eri tavalla eri tilasto-ohjelmissa. Tutustu verkossa SPSS harjoitukseen 1, jossa painokertoimet otetaan käyttöön.

Jos havainto saa lukua yksi suuremman painokertoimen (w > 1), on ryhmä, jota tämä havainto edustaa, aliedustettu aineistossa. Vastaavasti jos painokerroin on lukua yksi pienempi (w < 1), on havainnon edustama ryhmä yliedustettu.

Tarkastelussa on laskettu ns. analyysipainokertoimet, joiden summa aineistossa on havaintojen lukumäärä. Kertomalla analyysipainokerroin sopivalla luvulla (populaation koko jaettuna havaintojen lukumäärällä) saadaan ns. korottava paino. Tällöin korottavien painokertoimien summa on perusjoukon koko.

-- HARJOITUSTEHTÄVÄ --

Laske ISSP 2000 aineistoon uudet painomuuttujat, joissa painot on laskettu seuraavilla Suomen vuoden 2000 väestöjakaumilla. Käytä apuna MOTV:n verkkoversiossa olevia Tilastokeskuksen ikä- ja sukupuolijakaumataulukkoa sekä maakuntataulukkoa (Taulukot ovat Excel-muodossa. Jos selaimesi ei osaa avata niitä oikein, tallenna ne ensin kiintolevylle ja avaa MS Excel-yhteensopivalla taulukkolaskentaohjelmalla.)

a. sukupuoli
b. ikäluokka (15-19, 20-34, 35-49, 50-64, 65-74), sukupuoli
c. ikäluokka (15-19, 20-34, 35-49, 50-64, 65-74), sukupuoli, maakunta

ISSP 2000 aineiston alkuperäiset painokertoimet (weight, weight_2) on laskettu kalibrointimenetelmällä käyttäen apuna seuraavia väestöjakaumia:

1. sukupuoli,
2. ikäluokka (15-19, 20-24, ..., 64-69, 70-74),
3. kunta ja
4. kuntatyyppi (kaupunki - maaseutu).

Menetelmien tyyppejä ja soveltuvan menetelmän valinta

Määrällisten menetelmien kirjo on erittäin laaja ja niitä voi luokitella eri tavoin. Yksi luokittelutapa liittyy siihen, onko menetelmän kohteena yksittäinen muuttuja vai useita muuttujia. Jos kiinnostuksen kohteena on yksi muuttuja ja sen arvojen jakauma, voidaan puhua **yhden muuttujan menetelmistä** (*univariate methods*). Jos tarkastelun kohteena on yhtä aikaa useita muuttujia voidaan taas puhua **kahden muuttujan menetelmistä** (*bivariate methods*) tai, jos muuttujia on useampia kuin kaksi, **monimuuttujamenetelmistä** (*multivariate methods*). Lisäksi soveltuvan tutkimusmenetelmän valita riippuu muuttujien mittaustasosta.

Kun tarkastelun kohteena on vain yksi muuttuja, kiinnitetään yleensä huomiota muuttujan arvojen jakaumaan. Jakauman kuvailuun sopivat esimerkiksi graafinen tarkastelu, keskiluvut ja hajontaluvut. Soveltuvien keski- ja hajontalukujen valinta riippuu muuttujan mittaustasosta.

Jos kyseessä on kahden tai useamman muuttujan yhtäaikainen tarkastelu, voidaan menetelmiä luokitella sen mukaan sisältyykö niihin (joko eksplisiittinen tai implisiittinen) kausaalioletus. Esimerkiksi muuttujien välisiä korrelaatiokertoimia voidaan käyttää muuttujien yhteisvaihtelun tarkasteluun tekemättä etukäteen vahvoja oletuksia muuttujien kausaalisuhteista. Samoin eksploratiivinen faktorianalyysi ja ryhmittelyanalyysi ovat menetelmiä, jotka eivät varsinaisesti edellytä oletuksia muuttujien välisistä kausaalisuhteista.

Soveltuvan menetelmän valinta

Taulukossa 1 on ryhmitelty soveltuvia monimuuttujamenetelmiä siinä tapauksessa, että tutkijalla on etukäteen tehty kausaalioletus eli hän on valinnut selitettävän muuttujan ja yhden tai useamman muuttujan, joita käytetään selittävinä muuttujina. Tällaisessa tapauksessa muuttujien mittaustaso vaikuttaa soveltuvan menetelmän valintaan.

Taulukko 1. Soveltuvan monimuuttujamenetelmän valinta.

		Selitettävä muuttuja	
		Luokittelu- tai järjestysasteikko	Välimatka- tai suhdeasteikko
Selittävä muuttuja	Luokittelu- tai järjestysasteikko	• Ristiintaulukointi • Log-lineaariset mallit	• Varianssianalyysi
	Välimatka- tai suhdeasteikko	• Logistinen regressio • Multinomiaalinen regressio	• Regressioanalyysi

Jos sekä selitettävä että selittävä muuttuja ovat luokittelu- tai järjestysasteikollisia muuttujia, analyysimenetelmäksi käyvät esimerkiksi ristiintaulukointi tai log-lineaariset mallit. Ristiintaulukointi sopii tilanteeseen, jossa selittäviä muuttujia on vain yksi tai enintään muutama. Jos selittäviä muuttujia on useita, tulee ristiintaulukoiden tulkinta usein ongelmalliseksi. Tällaisessa tapauksessa saattaa log-lineaaristen mallien käyttö olla parempi vaihtoehto.

Kun selitettävä muuttuja on mitattu vähintään välimatka-asteikolla ja selittävä muuttuja on luokittelu- tai järjestysasteikollinen, tilanteeseen soveltuva menetelmä on varianssianalyysi. Sen avulla voidaan tutkia esimerkiksi sitä, kuinka paljon sukupuoli selittää naisten ja miesten palkkaeroja.

Regressioanalyysia voidaan käyttää silloin kun sekä selittävä että selitettävä muuttuja ovat mittaustasoltaan vähintään välimatka-asteikon muuttujia. Sen avulla voidaan esimerkiksi tutkia, mikä on työntekijän iän vaikutus hänen palkkaansa.

Silloin kun selitettävä muuttuja on enintään järjestysasteikollinen ja selittävä muuttuja on mitattu välimatka- tai suhdeasteikolla, on tarjolla kaksi vaihtoehtoa riippuen selitettävän muuttujan luonteesta. Jos selitettävä muuttuja on dikotomia (eli sillä on vain kaksi mahdollista arvoa), tarkoituksenmukainen analyysimenetelmä on logistinen regressioanalyysi. Sen avulla voi tutkia esimerkiksi sitä, miten ikä vaikuttaa siihen käyvätkö ihmiset äänestämässä vai eivät. Jos selitettävässä muuttujassa on enemmän kuin kaksi vaihtoehtoa, voidaan puolestaan käyttää multinomiaalista regressioanalyysia. Tutkija voi esimerkiksi analysoida sitä, miten ikä vaikuttaa siihen, äänestääkö vastaaja hallituspuolueiden ehdokkaita, oppositiopuolueiden ehdokkaita vai jättääkö hän äänestämättä kokonaan.

Lisäksi kannattaa muistaa, että sellaisessakin tilanteessa, jossa selittävinä muuttujina on eri mittaustason muuttujia, voidaan käyttää useampia edellä mainittuja menetelmiä. Esimerkiksi regressioanalyysissa (samoin kuin sen logistisessa tai multinomiaalisessa versiossa) voidaan käyttää selittäjinä myös luokittelu- tai järjestysasteikon muuttujia tekemällä niistä ns. dummy-muuttujia. Samoin varianssianalyysiin voi tarvitessa lisätä luokittelu-tai järjestysasteikollisten selittäjien joukkoon välimatka- tai suhdeasteikon muuttuja eli ns. kovariaatin.

Tilastollinen päättely

Määrällisen aineiston analyysissä tehdään usein ero kuvailevan tilasto-analyysin ja tilastollisen päättelyn välillä. **Kuvaileva tilastoanalyysi** *(descriptive statistics)* pyrkii nimensä mukaan kuvailemaan ja tiivistämään jonkin määrällisen muuttujan jakaumaa tai useamman määrällisen muuttujan yhteisvaihtelua pyrkimättä kuitenkaan tekemään tulosten pohjalta yleistyksiä mihinkään laajempaan perusjoukkoon. Jos kohteena on vain yksi muuttuja voidaan kuvailuun käyttää esimerkiksi muuttujien keskilukuja tai hajontalukuja. Useamman muuttujan tapuaksessa voidaan käyttää esimerkiksi korrelaatiokertoimia kuvaamaan niiden yhteisvaihtelua.

Otantaan perustuvissa yhteiskuntatieteellisissä tutkimuksissa ei kuitenkaan olla varsinaisesti kiinnostu-neita otoksesta vaan sen perusjoukon ominaisuuk-sista. Tällöin tarvitaan **tilastollista päättelyä** *(inferential statistics)*. Tilastollisen päättelyn avulla voidaan arvioida kuinka hyvin otoksesta saadut tulokset pitävät paikkansa perusjoukossa. Kyse on siis siitä, kuinka todennäköisesti otoksen avulla saadut tulokset voidaan yleistää koko perusjoukkoa koskeviksi tuloksiksi.

Kuvitellaan esimerkiksi tilanne, jossa kyselytutkimuksen avulla pyritään kartoittamaan suomalaisten mielipiteitä siitä, pitäisikö Suomen liittyä Natoon. Otoksessa 40% naisista ja 50% miehistä vastasi myöntä-västi kysymykseen Suomen Nato-jäsenyydestä. Varsinainen tutkimuksen mielenkiinto ei kuitenkaan ole otoksessa, vaan pyrkimys on selvittää mahdollisimman luotettavasti, kuinka suuri osuus perusjoukon (eli kaikki täysikäiset suomalaiset) naisista ja miehistä kannattaa jäsenyyttä. Tällöin keskeiseksi kysymykseksi nousee, mitä näiden otostulosten avulla voidaan päätellä yleensä Suomen naisista ja miehistä. Eroavatko miehet ja naiset perusjoukossa todella mielipiteiltään vai onko kyse vain satunnaisista otannan mukanaan tuomasta eroista? Tilastollinen päättely vastaa tällaisiin kysymyksiin.

Luottamusväli ja luottamustaso

Tilastollisen päättelyn kaksi keskeistä käsitettä ovat luottamusväli ja luottamustaso. **Luottamusväli** *(confidence interval)* kertoo millä välillä todellinen perusjoukon tunnusluvun arvo on tietyllä todennäköisyydellä. Käyttäen edelleen Nato-kyselyä esimerkkinä, voidaan kuvitella, että otoksessa 45% kaikista vastaajista ilmoitti kannattavansa Suomen Nato-jäsenyyttä. Koska tähän lukuun vaikuttavat monet satunnaiset tekijät, emme voi suoraan päätellä, että myös perusjoukossa (kaikki täysi-ikäiset suomalaiset) vastaava osuus on täysin sama. On kuitenkin to-dennäköistä, että perusjoukon mielipidettä kuvaava arvo on lähellä otoksesta saatua arvoa. Voimme esimerkiksi päätellä, että 95 %:n todennäköisyydellä Nato-jäsenyyttä kannattavien ihmisten osuus perusjoukossa on välillä 40-50 %. Tätä väliä kutsutaan luottamusväliksi.

Luottamustaso *(confidence level)* kertoo, millä todennäköisyydellä perusjoukkoa kuvaava tunnusluku on jollain tietyllä luottamusvälillä. Esimerkiksi 95 %:n todennäköisyydellä 40-50 % suomalaisista ha-luaa Suomen liittyvän Natoon. Luottamustaso on tällöin 95 % todennäköisyys.

Luottamustaso ja luottamusväli ovat siis täysin toisiinsa sitoutuneita käsitteitä. Tieto luottamusvälistä ei ole mielekäs, jos ei ole tietoa luottamustasosta ja päinvastoin. Olennaista on, että luottamustason kasvaessa laajenee myös luottamusväli. Toisin sanoen tämä tarkoittaa siis sitä, että mitä suuremmalla varmuudella haluamme tietää, millä välillä jokin perusjoukon tunnusluku sijaitsee, sitä suurempi on luottamusväli. Jos esimerkiksi haluaisimme tietää, millä välillä suomalaisten Nato-jäsenyyden kannatus on 99 % luottamustasolla, luottamusväli olisi suurempi kuin 95 % prosentin luottamustasolla (esimerkiksi 30-60 %). Jos olisimme valmiita tyytymään esimerkiksi 90 % luottamustasoon, väli voisi olla 43-47 %.

Otantajakauma

Luottamusvälin ja luottamustason ymmärtämiseksi ja laskemiseksi tarvitaan otantajakauman *(sampling distribution)* käsitettä. Otantajakauma on helpointa kuvailla esimerkin avulla. Kuvitellaan, että edellä esimerkkinä käytetty Nato-kysely on tehty käyttäen 1000 hengen satunnaistotosta (katso otantamenetelmät). Tämän otoksen vastaajista 45 % kannattaa Suomen Nato-jäsenyyttä. Koska tiedetään, että otokseen valintaan vaikuttavat satunnaiset tekijät, on luultavaa, että jos sama tutkimus tehtäisiin uudelleen käyttäen jälleen 1000 hengen otosta, Nato-jäsenyyden kannatus ei olisi tässä uudessa otoksessa täsmälleen sama kuin ensimmäisessä otoksessa. Oletetaan, että tässä toisessa otoksessa Nato-jäsenyyden kannatus olisi 42 %. Jos tutkimus toistettaisiin vielä kerran saman kokoisella satunnaisotoksella, jäsenyyden kannatus voisi olla 46 %. Tätä prosessia voitaisiin edelleen toistaa useita kertoja ja jokaisen uuden otoksen perusteella saataisiin uusi Nato-jäsenyyden kannatusta kuvaava prosenttiluku. Näistä luvuista voidaan muodostaa uusi muuttuja, jonka jakaumaa voidaan kutsua Nato-jäsenyyden kannatuksen otantajakaumaksi.

Määritelmän muukaan otantajakauma viittaa sellaiseen tunnusluvun jakaumaan, joka saadaan ottamalla kaikki mahdolliset saman kokoiset otokset perusjoukosta. Jos kiinnostuksen kohteena oleva muuttuja on Nato-jäsenyyttä kannattavien suomalaisten osuus kaikista suomalaisista ja otoksen koko on 1000 vastaajaa, saadaan Nato-kannattajien osuuden otantajakauma ottamalla kaikki mahdolliset 1000 hengen otokset suomalaisista ja kirjaamalla ylös saatu Nato-kannattajien osuus. Näiden kirjattujen kannattajalukujen jakauma on Nato-jäsenyyden kannatusta kuvaavan muuttujan otantajakauma. Viidestä miljoonasta suomalaisesta voidaan ottaa kuitenkin valtava määrä 1000 hengen otoksia. Niinpä otantajakauma on usein itse asiassa vain teoreettinen jakauma, jota ei empiirisesti yleensä pystytä määrittämään. Yleinen idea kuitenkin on, että käyttämällä tilastotieteen menetelmiä otantajakauman keskeiset piirteet pystytään saamaan selville.

Luottamusvälin laskeminen

Kuvitellaan, että aiemmin esitetyssä Nato-kysymyksessä on vain kaksi vaihtoehtoa eli vastaajat ovat joko jäsenyyden kannalla tai sitä vastaan. Vastaajista 45 % kannatti ja 55 % vastusti jäsenyyttä. Nyt tehtävänä on selvittää, millä välillä perusjoukon Nato-kannatus on tietyllä varmuudella. Kun vaihtoehtoja on vain kaksi, saadaan tulos käyttämällä seuraavaa kaavaa:

$$S = \sqrt{\frac{pq}{n}}$$

Kaavassa S tarkoittaa mielenkiinnon kohteena olevan tunnusluvun keskivirhettä (eli sen otantajakauman keskihajontaa), p on 'kyllä' vastanneiden prosenttiosuus, q on 'ei' vastanneiden prosenttiosuus ja n on otoksen koko. Sijoittamalla luvut (p=45, q=55, n=1000) kaavaan saadaan keskivirheen arvoksi noin 1,57. Tätä lukua voidaan käyttää hyväksi määriteltäessä Nato-kannatuksen luottamusväli perusjoukossa.

Nato-kannatuksen 95 % luottamusväli saadaan kaavasta p ± 1,96*S eli 45 ± 1.96*1,57. Tämä väli on 41,9 %-48,1 %. Eli johtopäätöksenä tutkija voisi todeta, että suomalaisten Nato-kannatus on 95 % prosentin todennäköisyydellä 41,9 % ja 48,1 % välillä. Käytännössä tämä tarkoittaa sitä, että jos suomalaisista otettaisiin hyvin suuri määrä 1000 hengen otoksia, 95 % näistä otoksista Nato-kannatus olisi edellä mainitulla välillä. Jos luottamustasoksi valitaan 99 %, kasvaa myös luottamusväli. Tällöin väli saadaan kaavasta p ± 2,58*S eli se olisi 40,9 %-49,1 %. Edelliset kertoimet (1,96 ja 2,58) saadaan normaalijakaumasta. Se, miten ne on johdettu, selitetään tilastotieteen oppikirjoissa, joten tässä yhteydessä siihen ei paneuduta syvemmin. Hyvä muistisääntö on, että 95 % luottamusväli saadaan noin ±2*keskivirhe, ja 99 % prosentin luottamustasolla vastaava kerroin on noin 2,5.

Jos kiinnostuksen kohteena on jonkin muuttujan keskiarvo, saadaan sen keskivirhe *(standard error of the mean)* kaavasta:

$$S = \frac{s}{\sqrt{n}}$$

Kaavassa S on keskiarvon keskivirhe, s on otoksesta laskettu muuttujan keskihajonta ja n on otoskoko. Keskiarvon keskivirhettä käytetään samalla tavalla kuin edellisessä esimerkissä.

Esimerkkinä keskiarvon keskivirheen käytöstä voidaan käyttää vuoden 1996 World Values -kyselyn Suomen osa-aineiston (ks. verkosta osaWVS-aineiston frekvenssit) kysymystä v123, jossa vastaajia pyydettiin arvioimaan itseään vasemmisto-oikeisto -mittarilla. Tässä mittarissa oli arvoja yhdestä kymmeneen, ja pienet luvut kuvastivat vasemmistolaisuutta ja suuret luvut oikeistolaisuutta. Etukäteen voidaan arvioida, että suomalaisten keskiarvo mittarilla on jossain sen keskivaiheilla, eli arvon 5,5 lähettyvillä. Seuraavaksi tutkitaan, eroaako suomalaisten keskiarvo tilastollisesti merkitsevästi tästä luvusta.

Kyselyn vastaajista 856 suostui sijoittamaan itsensä vasemmisto-oikeisto -ulottuvuudelle. Keskiarvo oli 5,61 eli keskimäärin suomalaiset vaikuttaisivat olevan hiukan keskipisteen "oikeammalla" puolella. Otoksesta laskettu muuttujan keskihajonta oli 1,92. Käyttämällä edellä esiteltyä keskiarvon keskivirheen kaavaa, saadaan keskivirheen arvoksi 0,19 (=1,92/v856). Samoin kuin edellisessä esimerkissä voidaan 95% luottamusväli suomalaisten keskimääräiselle sijoittumiselle oikeisto-vasemmisto -ulottuvuudella laskea kaavasta 5,61 ± 1,96*0,19 eli se on 5,24 - 5,98. Koska luku 5,5 sijoittuu tämän luottamusvälin sisään, johtopäätös on, että suomalaisten keskimääräisen poliittisen sijoittumisen ei voida sanoa eroavan 95% varmuudella laskennallisesta keskipisteestä. Lukijan tulkintojen varaan jääköön se, mitä tämä kertoo ulottuvuuden kyvystä kuvata suomalaista puoluejärjestelmää.

Otoksen ja perusjoukon suuruuden merkitys

Edellä esitettyä keskiarvon keskivirheen kaavaa voidaan käyttää hyväksi tarkasteltaessa otoskoon merkitystä tilastollisessa päättelyssä. Kaavassa on jakajana otoskoon neliöjuuri. Tämä tarkoittaa sitä, että otoskoon kasvaessa keskivirhe pienenee ja valitun luottamustason luottamusvälit kapenevat. Toisin sanoen tämä vahvistaa sinänsä intuitiivisestikin selvän havainnon, että otoskoon kasvaessa pystytään tekemään tarkempia arvioita kiinnostuksen kohteena olevista ilmiöistä. Koska kaavassa on jakajana otoskoon neliöjuuri, ei otoskoon kasvulla ja tarkentuneilla perusjoukon estimaateilla ole kuitenkaan lineaarista yhteyttä. Neliöjuuren takia täytyy otoskoko nelinkertaistaa, jotta luottamusväli pystyttäisiin pienentämään puoleen.

Toinen (ja vaikeammin intuitiivisesti ymmärrettävä) havainto on se, että perusjoukon koolla ei ole vaikutusta tilastollisten yleistysten tarkkuuteen. Edellä esitellyssä keskiarvon keskivirheen kaavassa ei ole perusjoukon koko mukana. Tämä tarkoittaa karkeasti ottaen sitä, että samankokoisilla otoksilla voidaan arvioida samoja ilmiöitä väestömäärältään erikokoisissa valtioissa jokseenkin samalla tarkkuudella. Tämä huomioiden ei ole yllättävää, että esimerkiksi presidenttiehdokkaiden kannatusmittaukset tehdään sekä Suomessa että Yhdysvalloissa suurin piirtein samanlaisilla otoskoilla (1000-2000 vastaajaa). Koska molemmissa maissa kyselyiden tilaajat ovat valmiita hyväksymään saman tarkkuustason valtakunnallisissa tuloksissa, ei Yhdysvalloissa olisi järkevää lähteä tekemään tutkimuksia paljon suuremmilla otoksilla kuin Suomessa.

Keskiluvut

Yhden muuttujan analyysissa mielenkiinto useimmiten kohdistuu muuttujan jakaumaan eli siihen, miten ja mille vaihteluvälille muuttujan arvon ovat jakautuneet. Yksi tapa tarkastella jakaumaa on käyttää graafisia kuvioita. Joskus on tarpeellista tiivistää jakaumaa kuvaava informaatio yhden tai useamman tunnusluvun avulla. Tällöin voidaan käyttää ns. keski- ja hajontalukuja. Keskiluvut kuvaavat muuttujien arvojen keskimääräistä suuruutta ja hajontaluvut sitä, kuinka paljon muuttujan arvot vaihtelevat.

Soveltuvan keskiluvun valintaan vaikuttaa muuttujan mittaustaso. Taulukossa 1 on esitetty sopivat keskiluvut tarkastelun kohteena olevan muuttujan mittaustason mukaan.

Taulukko 1. Soveltuvan keskiluvun valinta muuttujan mittaustason mukaan.

x = voi käyttää - = ei voi käyttää		**Muuttujan mittaustaso**			
		Luokittelu-asteikko	Järjestys-asteikko	Välimatka-asteikko	Suhde-asteikko
Keskiluku	Moodi	X	X	X	X
	Mediaani	-	X	X	X
	Artimeettinen keskiarvo	-	-	X	X
	Geometrinen ja harmoninen keskiarvo	-	-	-	X

Taulukossa 2 on esitetty kuvitteellinen esimerkkiaineisto työpaikan työntekijöistä ja kolmesta heiltä mitatusta muuttujasta. Sukupuolimuuttuja on mitattu luokitteluasteikolla, koska siinä on kaksi vaihtoehtoa, joita ei voi asettaa suuruusjärjestykseen. Koulutusmuuttuja on järjestysasteikon muuttuja. Siinä on kolme vaihtoehtoa, jotka voidaan järjestää koulutuksen laajuuden mukaan. Lapsien lukumäärä on suhdeasteikolla mitattu muuttuja. Taulukon muuttujien tietoja voidaan tiivistää tunnusluvuiksi käyttämällä soveltuvia keskilukuja.

Taulukko 2. Kuvitteellinen aineisto työpaikan kymmenestä työntekijästä.

Työntekijän havaintonumero	**Sukupuoli**	**Koulutus**	**Lapsien määrä**
Työntekijä 1	Mies	Peruskoulu	0
Työntekijä 2	Nainen	Keskiaste	4
Työntekijä 3	Nainen	Keskiaste	1
Työntekijä 4	Mies	Korkeakoulu	1
Työntekijä 5	Nainen	Keskiaste	2
Työntekijä 6	Nainen	Korkeakoulu	1
Työntekijä 7	Nainen	Korkeakoulu	1
Työntekijä 8	Mies	Peruskoulu	0
Työntekijä 9	Mies	Korkeakoulu	0
Työntekijä 10	Nainen	Keskiaste	2

Moodi

Moodi *(mode)* eli tyyppiarvo on kaikkein joustavin keskiluku siinä mielessä, että sitä voidaan käyttää kaikissa tilanteissa muuttujan mittaustasosta huolimatta. Jos muuttujan mittaustaso on luokitteluasteikko, on moodi ainoa mahdollinen keskiluku. Moodi on yksinkertaisesti se muuttujan arvo, jonka frekvenssi aineistossa on suurin.

Esimerkiksi taulukon 2 aineistossa on neljä miestä ja kuusi naista. Näin sukupuolimuuttujan moodi on "nainen". Yleisin lapsiluku on yksi, eli taulukossa esitetyn lapsilukumuuttujan moodi on 1.

Muuttujalla voi olla myös useita moodeja. Näin käy silloin kun kahden tai useamman muuttujan arvon frekvenssi ovat yhtä suuria ja samalla suurimmat koko aineistossa. Koulutuksen osalta yleisimmät arvot ovat "keskiasteen" ja "korkeakouluasteen koulutus" eli koulutusmuuttujalla on kaksi moodia.

Mediaani

Mediaani *(median)* on keskiluku, jota voidaan käyttää järjestysasteikolla, välimatka- tai suhdeasteikolla mitatun muuttujan yhteydessä. Mediaani on suuruusjärjestykseen asetetuista muuttujan arvoista keskimmäinen. Jos havaintoja on parillinen määrä riippuu mediaanin arvo siitä, onko muuttuja mitattu järjestysasteikolla vai välimatka- tai suhdeasteikolla. Jos mittaustaso on järjestysasteikko, on mediaani tässä tapauksessa kumpikin keskimmäisistä arvoista. Jos mittaustasona on välimatka- tai suhdeasteikko, on mediaani kahden keskimmäisen arvon keskiarvo.

Esimerkiksi taulukon 2 lapsien määrää koskevat havainnot voidaan asettaa suuruusjärjestykseen seuraavalla tavalla: 0 0 0 1 1 1 1 2 2 4

Koska taulukossa on parillinen määrä havaintoja, täytyy mediaanin määrittelemiseksi löytää kaksi keskimmäistä arvoa. Nämä ovat 1 ja 1. Koska muuttuja mittaustaso on suhdeasteikko, on aineiston mediaani näiden kahden havainnon keskiarvo eli 1.

Koulutusmuuttujan osalta havaintojen suuruusjärjestys on (P=peruskoulu, KE=keskiaste, KO=korkeakoulu): P P KE KE KE KE KO KO KO KO

Mediaani on tämän muuttujan osalta "keskiasteen koulutus".

Koska sukupuolimuuttuja on luokitteluasteikon muuttuja, siitä ei voi laskea mediaania.

Mediaanin erityinen hyöty keskilukuna on, että siihen eivät vaikuta muista muuttujan arvoista huomattavasti poikkeavat suuret tai pienet arvot. Jos havaintojen määrä on pieni, voi tällaiset äärimmäisen poikkeavat arvot vaikuttaa suuresti aritmeettisen keskiarvon suuruuteen. Tämän vuoksi esimerkiksi palkkatietoja raportoitaessa käytetään yleensä keskilukuna mediaania keskiarvon sijasta. Tällöin joidenkin henkilöiden erittäin suuret palkat eivät vaikuta "vääristävästi" tuloksiin, kun keskustellaan keskimääräisestä palkasta.

Soveltuvan keskiluvun valinta riippuu myös siitä, mitä muuttujan ominaisuutta halutaan korostaa. Joissakin tapauksissa palkkojen aritmeettinen keskiarvo voi olla parempi keskiluvun mittari kuin mediaani.

Aritmeettinen keskiarvo

Aritmeettinen keskiarvo *(mean)* on kaikkein yleisin muuttujan "keskimääräisyyttä" kuvaava keskiluku. Sitä käytetään välimatka- tai suhdeasteikolla mitattuihin muuttujiin. Aritmeettinen keskiarvo saadaan laskemalla kaikki havaintojen arvot yhteen ja jakamalla saatu summa havaintojen määrällä.

Eli tutun kaavan mukaan: *keskiarvo = havaintojen summa / havaintojen määrä*. Esimerkiksi taulukon 2 lapsimäärien keskiarvo on (0+4+1+1+2+1+1+0+0+2)/10=1,2.

Aritmeettinen keskiarvo on intuitiivisesti helppo ymmärtää ja siksi erittäin suosittu keskiluku. Silti kannattaa muistaa, että poikkeavat muuttujan arvot voivat vaikuttaa suuresti

aritmeettisen keskiarvon suuruuteen etenkin pienissä aineistoissa. Esimerkiksi lukusarjan (1,1,1,1,100) aritmeettinen keskiarvo on 20,8 ja saman lukusarjan mediaani 1.

Geometrinen ja harmoninen keskiarvo

Geometrinen ja harmoninen keskiarvo ovat suhdeasteikon muuttujille sopivia keskilukuja. Koska aritmeettinen keskiarvo sopii hyvin myös suhdeasteikon muuttujille, on geometrisen ja harmonisen keskiarvon käyttö harvinaista. Näitä lukuja käytetään lähinnä kasvuilmiöihin ja indeksilaskentaan liittyvissä erikoistapauksissa.

Geometrinen keskiarvo (G) voidaan laskea seuraavasta kaavasta:

$$G = \sqrt[n]{x_1 \times x_2 \times ... \times x_n}$$

Tässä kaavassa n viittaa havaintojen määrään ja x_1 ensimmäisen havainnon arvoon, x_2 toisen havainnon arvoon jne. Geometrisessa keskiarvossa siis kaikki havaintojen arvot kerrotaan keskenään ja saadusta tuloksesta otetaan n:s juuri. Geometrista keskiarvoa voidaan käyttää hyväksi esimerkiksi laskettaessa hintaindeksistä keskimääräistä vuotuista hintatason nousua.

Harmoninen keskiarvo (H) lasketaan kaavasta:

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n}}$$

Kaavassa n viittaa jälleen havaintojen määrään. Samoin kuin geometrisella keskiarvolla, harmonisella keskiarvolla on sovelluksia indeksilaskennassa.

Hajontaluvut

Keskilukujen lisäksi hajontaluvut ovat erittäin yleisiä muuttujan jakaumaa kuvaavia mittalukuja. Hajontaluvut kertovat, kuinka muuttujan arvot vaihtelevat käytetyn keskiluvun "ympärillä". Kahdella muuttujalla voi olla sama keskiluku (esimerkiksi keskiarvo), mutta niiden hajonta voi olla täysin erilainen. Siksi muuttujan jakaumaa kuvatessa on tapana esittää sekä sopiva keskiluku että hajontaluku.

Samoin kuin keskiluvuissa muuttujan mittaustaso vaikuttaa soveltuvan hajontaluvun valintaan. Taulukossa 1 on esitetty, mitkä hajontaluvut sopivat millekin muuttujan mittaustasolle.

Taulukko 1. Sopivan hajontaluvun valinta muuttujan mittaustason mukaan.

x = voi käyttää - = ei voi käyttää		**Muuttujan mittaustaso**			
		Luokittelu-asteikko	Järjestys-asteikko	Välimatka-asteikko	Suhde-asteikko
Hajontaluku	Variaatiosuhde	X	X	X	X
	Vaihteluväli	-	X	X	X
	Vaihteluvälin pituus	-	-	X	X
	Keskihajonta	-	-	X	X
	Variaatiokerroin	-	-	X	X

Taulukossa 2 on esitetty kuvitteellinen aineisto, jota käytetään erilaisten hajontalukujen esittelemiseksi. Siinä on kolme eri mittaustason muuttujaa.

Työntekijän havaintonumero	**Sukupuoli**	**Koulutus**	**Lapsien määrä**
Työntekijä 1	Mies	Peruskoulu	0
Työntekijä 2	Nainen	Keskiaste	4
Työntekijä 3	Nainen	Keskiaste	1
Työntekijä 4	Mies	Korkeakoulu	1
Työntekijä 5	Nainen	Keskiaste	2
Työntekijä 6	Nainen	Korkeakoulu	1
Työntekijä 7	Nainen	Korkeakoulu	1
Työntekijä 8	Mies	Peruskoulu	0
Työntekijä 9	Mies	Korkeakoulu	0
Työntekijä 10	Nainen	Keskiaste	2

Variaatiosuhde

Variaatiosuhde *(variation ratio)* on hajontaluku, jota voidaan käyttää luokitteluasteikollisen muuttujan yhteydessä. Se on helppo laskea ja ymmärtää. Variaatiosuhde kertoo, kuinka suuri osuus havainnoista on muuttujan moodiluokassa. Variaatiosuhde (v) lasketaan kaavasta: *v = 1 - (havaintojen määrä moodiluokassa / havaintojen määrä)*.

Variaatiosuhde vaihtelee nollan ja yhden välillä. Se saa arvon nolla, jos kaikki muuttujan arvot ovat moodiluokassa. Tässä tapauksessa muuttujan arvot eivät tietenkään vaihtele

ollenkaan, joten on luontevaa, että hajontaluku saa arvon nolla. Mitä lähempänä yhtä variaatiosuhde on, sitä enemmän aineistossa on hajontaa.

Taulukon 2 aineistossa lapsiluvun yleisin arvon on yksi (eli se on muuttujan moodiluokka) ja aineistossa on neljä työntekijää, joilla on yksi lapsi perheessään. Näin aineiston variaatiosuhde on 1-(4/10)=0,6.

Sukupuolimuuttujan yleisin arvo on "nainen", joita on aineistossa kuusi. Näin variaatiosuhde on tämän muuttujan osalta 1-(6/10)=0,4.

Koulutusmuuttujan osalta aineistossa on kaksi moodia ("keskiaste" ja "korkeakoulu"). Variaatiosuhde lasketaan siitä muuttujan luokasta, jossa on eniten havaintoja. Koska tässä tapauksessa on kaksi tällaista luokkaa, ei ole väliä kummasta variaatiosuhde lasketaan. Se on koulutuksen osalta 1-(4/10)=0,6.

Variaatiosuhdetta käytetään yleensä vain luokitteluasteikollisten muuttujien yhteydessä. Muuttujan ollessa välimatka- tai suhdeasteikollinen sen käyttö ei useimmiten ole järkevää, vaan on luonnollista valita jokin hajontaluku, joka sopii paremmin tähän tarkoitukseen. Välimatka- tai suhdeasteikon muuttuja voi saada suuren määrän erilaisia arvoja jollain tietyllä välillä. Tällaisessa tapauksessa variaatiosuhteen käyttäminen hajontalukuna ei ole mielekästä, koska on epätodennäköistä, että moodiluokassa olisi kovinkaan monta havaintoa. Jos tutkitaan esimerkiksi nettopalkkoja jollain tietyllä teollisuuden alalla, on epätodennäköistä, että löytyisi suuri ryhmä työntekijöitä, joilla on täsmälleen sama palkka.

Vaihteluväli

Vaihteluväli *(range)* on järjestys-, välimatka- ja suhdeasteikon muuttujille sopiva hajontaluku. Se ilmoittaa yksinkertaisesti pienimmän ja suurimman muuttujan arvon välin. Määritelmän mukaan vaihteluväli on $W=[x_1, x_n]$ silloin kun havaintojen arvot on sijoitettu suuruusjärjestykseen aloittaen pienimmästä muuttujan arvosta. x_1 viittaa edellisessä kaavassa aineiston pienimpään arvoon ja x_n sen suurimpaan arvoon.

Esimerkiksi taulukon 2 aineistossa lapsien lukumäärän pienin arvo on nolla ja suurin arvo neljä. Näin vaihteluväli muuttujan arvoille on W=[0, 4]. Koulutuksen osalta vaihteluväli on W=["peruskoulu", "korkeakoulu"]. Koska sukupuolimuuttuja on luokitteluasteikollinen, siitä ei ole järkevää tarkastella vaihteluväliä.

Vaihteluvälin pituus

Kun muuttuja on mitattu välimatka- tai suhdeasteikolla, voidaan puhua vaihteluvälin pituudesta. Se on yksinkertaisesti muuttujan suurimman ja pienimmän arvon erotus. Kaavana se voidaan ilmaista seuraavasti: $w = x_n - x_1$

Kaavassa w on vaihteluvälin pituus, x_n muuttujan suurin arvo ja x_1 muuttujan pienin arvo.

Esimerkiksi taulukon 2 aineistossa vaihteluvälin pituus on 4-0 = 4.

Keskihajonta

Keskihajonta *(standard deviation)* on hajontaluku välimatka- tai suhdeasteikon muuttujille. Keskihajonta on ehkä kaikkein yleisimmin käytetty hajontaluku. Keskihajonta kuvaa sitä, kuinka kaukana yksittäiset muuttujan arvot ovat keskimäärin muuttujan aritmeettisesta keskiarvosta. Keskihajonta (s) lasketaan kaavasta

$$s = \sqrt{\frac{1}{n}\sum_{i-1}^{n}(x_i - \bar{x})^2}$$

Kaavassa x_i viittaa i:nnen havainnon arvoon ja $\bar{x}$ tarkoittaa aineiston aritmeettista keskiarvoa. Sigma -merkki (S) tarkoittaa yhteenlaskua. Esitetyssä kaavassa lasketaan jokaisen havainnon arvon erotus koko aineiston keskiarvosta. Tämän jälkeen erotus korotetaan neliöön.

Tämän jälkeen kaikki saadut arvot lasketaan yhteen. Tämä saatu summa jaetaan havaintojen määrällä (n) ja saadusta tuloksesta otetaan vielä neliöjuuri keskihajonnan saamiseksi. Mitä suurempi saatu arvo on, sitä enemmän muuttujan arvoissa on hajontaa ja päinvastoin.

Edellä mainittu keskihajonnan kaava on tarkoitettu tilanteisiin, jossa on tarkasteltavana koko perusjoukko. Jos kyse on otoksesta käytetään usein termiä otoskeskihajonta ja silloin täytyy käyttää hieman erilaista kaavaa. Tällöin kaava on

$$s = \sqrt{\frac{1}{n-1}\sum_{i-1}^{n}(x_i - \bar{x})^2}$$

Tässä kaavassa jakajana on havaintojen määrä vähennettynä yhdellä. Etenkin suurissa aineistoissa ero näiden kahden kaavan antamilla keskihajontaluvuilla on pieni.

Jos oletetaan, että taulukon 2 aineisto on koko perusjoukko (eli kaikki työpaikan työntekijät) saadaan lapsiluvun keskihajonnaksi (ensimmäisen kaavan mukaan) 1,17. Jos taas oletetaan, että kyseessä on iso työpaikka ja aineisto on vain kymmenen hengen otos koko perusjoukosta, saadaan keskihajonnaksi (jälkimmäisen kaavan mukaan) 1,23. Ero on pieni, vaikka aineisto koostuikin vain kymmenestä havainnosta.

Keskihajonnan käsitteeseen liittyy usein myös varianssin käsite. Varianssilla tarkoitetaan keskihajonnan neliötä (s^2). Varianssia käytetään monessa tilastolliseen päättelyyn liittyvässä yhteydessä.

Variaatiokerroin

Kahden eri otoksen keskihajontojen keskinäinen vertailu on joskus ongelmallista, koska keskihajonta vaihtelee aineiston keskiarvon myötä. Variaatiokerroin *(coefficient of variation)* on hajontaluku, joka suhteuttaa keskihajonnan aineiston keskiarvoon. Se lasketaan kaavasta

$$V = s / \bar{x}$$

Kaavassa s on muuttujan keskihajonta ja $\bar{x}$ on muuttujan keskiarvo. Käytännössä siis keskihajonta suhteutetaan muuttujan keskiarvoon. Näin kahden ryhmän hajonnan vertailu on mielekkäämpää.

Ristiintaulukointi

Ristiintaulukoinnilla tutkitaan muuttujien jakautumista ja niiden välisiä riippuvuuksia. Riippuvuus- tai riippumattomuustarkastelussa tutkitaan, onko tarkastelun kohteena olevan selitettävän muuttujan jakauma erilainen selittävän muuttujan eri luokissa.

Tutkimuskysymyksenä voi olla esimerkiksi se, eroavatko naiset ja miehet siinä, kuinka hyvänä tai huonona asiana he pitävät Suomen EU-jäsenyyttä. Ristiintaulukointi kertoo eroavatko nais- ja miesvastaajien vastausjakaumat toisistaan. Jos vastausskaala on dikotominen kyllä/ei, lasketaan vaihtoehtojen osuudet sukupuolimuuttujan kahdessa eri luokassa ja verrataan niiden suuruuksia. Tässä esimerkissä sekä selittävässä että selitettävässä muuttujassa on vain kaksi luokkaa, mutta niissä voisi olla myös useampia luokkia. Ristiitaulukoinnissa voidaan käyttää myös välimatka- tai suhdeasteikolla mitattuja muuttujia, mutta ne on sitä ennen uudelleenkoodattava luokitelluiksi muuttujiksi.

Ristiintaulukon muodostaminen

Suraavassa esimerkissä tutkitaan miesten ja naisten välisiä eroja politiikasta keskustelemisen aktiivisuudessa. Esimerkkiaineistossa (osaWVS-aineiston kuvaus verkossa) on pyydetty vastausta seuraavaan kysymykseen: "Kun olette tekemisissä ystävienne kanssa, niin keskusteletteko heidän kanssaan poliittisista asioista usein, silloin tällöin, vai ei koskaan?" (kysymys V37). Ristiintaulukoinnin avulla pystytään vastaamaan siihen, kuinka aktiivisia naiset ja miehet ovat ja onko aktiivisuus yhtä suuri verrattaessa sukupuolia toisiinsa.

Taulukon 1 kuudessa solussa on esitetty ristiintaulukoinnin tuottamat vastaajien lukumäärät.

Taulukko 1. Aktiivisuus keskustella poliittisista asioita ystävien kanssa sukupuolen mukaan (absoluuttiset luvut).

	Mies	Nainen
Usein	43	29
Silloin tällöin	323	298
En koskaan	108	174

Taulukko 1 osoittaa, miten vastaajat ovat jakautuneet sarake- (sukupuoli) ja rivimuuttujan (keskustelun aktiivisuus) eri vaihtoehtoihin. Esimerkiksi 43 miesvastaajaa ilmoitti keskustelevansa ystäviensä kanssa politiikasta usein. Naisvastaajissa heitä oli 29. Taulukosta on kuitenkin vaikea havaita suoraan, eroavatko sukupuolet politiikasta keskustelun aktiviteetin suhteen toisistaan. Luvuthan eivät ole suoraan vertailukelpoisia, koska nais- ja miesvastaajien määrät otoksessa eroavat toisistaan. Tämän vuoksi on syytä laskea uuteen ristiintaulukkoon prosenttijakaumat selitettävälle muuttujalle. Tämä on tehty taulukossa 2.

Taulukko 2. Aktiivisuus keskustella poliittisista asioita ystävien kanssa sukupuolen mukaan (%).

	Mies	Nainen
Usein	9	6
Silloin tällöin	68	60
En koskaan	23	35
Yhteensä (n) c^2 =18,4; vapausasteita=2; p<0,01	100 474	100 501

Ristiintaulukoinnissa tarkastellaan siis ehdollisia jakaumia. Tämä tarkoittaa sitä, että mielenkiinnon kohteena olevan selitettävän muuttujan jakaumaa tarkastellaan selittävän muuttujan eri luokissa. Koska selitettävän muuttujan arvot jakautuvat vain harvoin tasaisesti selittävän muuttujan luokkiin, on analyysissa selkeyden vuoksi tarpeellista käyttää suhteellista jakaumaa eli laskea prosenttiosuudet.

Taulukon 2 esimerkki selventää asiaa. Myös nyt sarakkeilla ovat selittävän muuttujan (sukupuoli) luokat ja riveillä selitettävän muuttujan luokat. Taulukon prosenttijakaumat osoittavat selkeästi naisten ja miesten erot keskusteluaktiviteetissa. Naisista 35% ei keskustele koskaan poliittisista asioista ystäviensä kanssa, kun taas miesten osalta vastaava luku on 23%. Usein poliittisista asioista keskustelee miehistä 9% ja naisista 6%. Voidaan tehdä johtopäätös, että otoksen perusteella miehet puhuvat politiikasta ystäviensä kanssa useammin kuin naiset.

Ristiintaulukon alimmalla prosenttirivillä on laskettu prosenttiosuudet yhteen. Pyöristys voi joskus aiheuttaa pienen poikkeaman sadasta prosentista, mutta yleensä yhteenlaskettu prosenttiosuus ilmoitetaan silti tasalukuna (100%). Yhteenlaskettu prosentti on syytä lisätä taulukkoon, koska se kertoo lukijalle heti mihin suuntaan taulukon prosenttijakaumat on laskettu. Lisäksi on tapana ilmoittaa absoluuttiset määrät (n), joiden perustalta prosenttiluvut on laskettu. Näin lukija pystyy arvioimaan myös tulosten luotettavuutta. Lisäksi taulukossa 2 on esitetty merkitsevyystestin tulokset. Näiden tulosten tulkinta ja niiden laskeminen käydään läpi kohta omassa osiossaan. Lisäksi taulukkojen raportointia ja ulkoasua käsitellään toisaalla tarkemmin.

Ristiintaulukoitaessa on tarkkaan mietittävä mihin suuntaan prosenttijakaumat tulee laskea. Tämän ratkaisee tutkimusongelma. Jos taulukossa 2 prosentit olisikin laskettu vaakasuoraan, tulokset eivät olisi vastanneet esitettyyn kysymykseen siitä, eroavatko miehet ja naiset keskusteluaktiviteettinsa suhteen. Prosentit olisivat kertoneet esimerkiksi "usein" keskustelevan ryhmän sukupuolirakenteen eli sen, kuinka suuri osuus heistä on miehiä tai naisia.

Jos otoksessa olisi ollut jostakin syystä huomattavasti enemmän naisia kuin miehiä, olisi naisten prosenttiosuus ollut luultavasti kaikissa keskusteluaktiviteetin ryhmissä suurempi kuin miesten prosenttiosuus. Tämä tulos ei kuitenkaan olisi kertonut mitään siitä, ovatko naiset enemmän tai vähemmän aktiivisia keskustelemaan politiikasta ystäviensä kanssa kun heitä verrataan miehiin.

Selittävän ja selitettävän muuttujan sijainnille ristiintaulukoinnissa ei ole olemassa yhtä yleispätevää sääntöä. Jos ristiintaulukkoon sisältyy selkeä kausaalinen asetelma, on tavanomaista asettaa selittävä muuttuja taulukon yläreunaan eli sarakkeille ja selitettävä muuttuja taulukon sivulle eri riveille. Tällöin prosentit lasketaan sarakkeiden sisällä siten, että yhteenlasketut prosenttiluvut ja lukumäärät sijoittuvat taulukon alalaitaan. Joskus selittävässä muuttujassa voi kuitenkin olla niin monta luokkaa, että käytännön syistä ne kannattaa sijoittaa riveille ja selitettävän muuttujan luokat sarakkeille. Tässä tapauksessa prosenttijakauma on tietenkin laskettava riveittäin.

Erityistapauksissa voi olla tarpeellista laskea prosenttiosuudet koko aineistosta, eikä ainoastaan selittävän muuttujan luokkien sisällä. Tutkija voi esimerkiksi haluta tietää, kuinka suuri osuus koko aineistossa on tietyn ikäisiä naisia. Tämän tuloksen hän saa ristiintaulukoimalla

iän sukupuolen mukaan ja laskemalla solujen lukumäärien prosenttiosuudet kaikkien havaintoyksikköjen määrästä.

Ristiintaulukon merkitsevyyden testaus

Kuten tilastollisen päättelyn osiossa todetaan, otoksiin perustuvissa tutkimuksissa mielenkiinnon kohteena on se, voidaanko otoksessa havaittujen erojen pätevän myös perusjoukossa (eli tässä esimerkissä kaikki täysi-ikäiset suomalaiset). Taulukon 2 prosenttiluvut osoittavat miesten ja naisten erot otoksessa, mutta tärkeä kysymys on, voidaanko näistä tuloksista päätellä tarpeeksi varmasti, että sukupuolten välinen ero säilyy myös tarkasteltaessa koko perusjoukkoa. Tällaiset kysymykset kuuluvat tilastollisen päättelyn alaan. Ristiintaulukoille soveltuva tilastollisen merkitsevyyden testausmenetelmä on ns. c^2-testi ("khii-toiseen testi"; c-merkki on yksi kreikkalaisista aakkosista).

c^2-testi on ns. riippumattomuustesti. Sen lähtökohtaisena oletuksena eli nollahypoteesina on muuttujien välinen riippumattomuus. Esimerkissämme tämä edellyttää, että miehet ja naiset eivät eroa keskusteluaktiviteetissaan toisistaan. Toisin sanoen sukupuoli ja politiikasta keskusteleminen olisivat siis toisistaan riippumattomia muuttujia.

Testin perustana on havaittujen frekvenssien ja odotettujen frekvenssien erotusten suuruus. Odotetuilla frekvensseillä tarkoitetaan sitä havaintojen jakaumaa, joka syntyisi, jos miehet ja naiset keskustelisivat politiikasta yhtä aktiivisesti. Esimerkiksi taulukossa 2 tämä tarkoittaisi sitä, että miesten ja naisten kohdalla prosenttiluvut olisivat täysin samat.

c^2-testissä tarkastellaan sitä, kuinka paljon havaitut ja odotetut frekvenssit eroavat toisistaan. Jos erot ovat tarpeeksi suuria, voidaan todeta, että havaitut erot eivät todennäköisesti johdu ainoastaan sattumasta, vaan ne ovat löydettävissä myös perusjoukossa.

Käytännössä testin tulokset tiivistyvät p-lukuun. Se kertoo virhepäätelmän todennäköisyyden silloin kun oletetaan, että otoksessa havaitut erot löytyvät myös perusjoukosta. P:n arvon ollessa alle 0,05 todetaan, että erot ovat tilastollisesti merkitseviä. Taulukossa 2 c^2-testin tulos on $p<0,01$ eli päätelmänä on, että suomalaiset naiset ja miehet eroavat toisistaan tavoissaan keskustella politiikasta ystäviensä kanssa (miehet keskustelevat enemmän). Tämä päätelmä voi olla virheellinen, mutta virheen todennäköisyys on alle yhden prosentin (eli $p<0,01$). Jos testin osoittama p:n arvo olisi ollut suurempi kuin 0,05, olisi päätelmä ollut, että miehet ja naiset eivät eroa tilastollisesti merkitsevästi toisistaan sen suhteen, kuinka usein he keskustelevat politiikasta ystäviensä kanssa.

c^2-testin periaatteet ja laskutapa on esitelty tarkemmin omassa alaluvussaan: *Ristiintaulukon riippumattomuustesti*.

Ristiintaulukon tilastollisen merkitsevyyden testaamisessa kannattaa huomioida, että testaus ei kerro mitään ristiintaulukon sisältämien erojen sisällöllisestä merkitsevyydestä. Testi kertoo vain kuinka todennäköistä on, että otoksessa havaitut erot ovat olemassa myös perusjoukossa. Jos otoskoko on hyvin suuri, on todennäköistä, että pienikin riippuvuus muuttujien välillä antaa tilastollisesti merkitsevän c^2-testituloksen. Siksi on tärkeää muistaa, että tilastollisen merkitsevyyden lisäksi täytyy aina pohtia myös erovaisuuksien suuruuden sisällöllistä merkitystä. Vastuu johtopäätöksistä on loppujen lopuksi aina tutkijalla.

Ristiintaulukon elaboraatio

Elaboraatiolla tarkoitetaan prosessia, jossa jo löytynyttä kausaalisuhdetta yritetään tarkentaa tuomalla analyysiin mukaan asiaan vaikuttavia lisätekijöitä. Seuraavassa esimerkissä tarkastellaan ristiintaulukoinnin avulla sitä, miten löytynyt yhteys sukupuolen ja keskusteluaktiviteetin välillä muuttuu, jos sitä tarkastellaan eri ikäryhmissä.

Kuten taulukko 2 osoitti, miehillä ja naisilla vaikuttaisi olevan eroavaisuuksia heidän aktiivisuudessaan keskustella poliittisista asioista ystäviensä kanssa. Seuraavassa esimerkissä tarkastellaan, miten näkemys sukupuolien välisestä erosta muuttuu, jos asiaa tarkastellaan eri ikäryhmissä. Tätä varten aineiston ikämuuttuja on luokiteltu kolmeen eri luokkaan (alle 35 vuotta, 35-59 vuotta ja 60 vuotta täyttäneet; ks. uusien muuttujien luominen). Ristiintaulukointi tehdään nyt kaikille kolmelle ryhmälle erikseen. Tulokset ovat taulukossa 3.

Taulukko 3. Aktiivisuus keskustella poliittisista asioista ystävien kanssa sukupuolen mukaan ikäryhmittäin (%).

	Alle 35 v.		35-59 v.		60 v. täyttäneet	
	Mies	Nainen	Mies	Nainen	Mies	Nainen
Usein	4	2	10	8	17	9
Silloin tällöin	64	58	71	62	72	59
En koskaan	33	40	19	31	11	32
Yhteensä	100	100	100	100	100	100
N	183	184	194	199	96	118
	c^2=2,8; vapausast.=2; p=0,24		c^2=7,8; vapausast.=2; p=0,02		c^2=14,4; vapausast.=2; p<0,01	

Aiemmin havaittu näkemys sukupuolen ja poliittisen keskusteluaktiviteetin välisestä suhteesta tarkentuu, kun sitä tarkastellaan vastaajien ikäryhmän suhteen. Nuorimmat naiset keskustelevat politiikasta ystävien kesken kaikkien vähiten. Lisäksi taulukoiden merkitsevyystestien tulkinta tarkentaa kuvaa sukupuolien välisestä erosta. Alle 35-vuotiaiden osalta c^2-riippumattomuustestin p-arvo on selkeästi 0,05 suurempi. Tämä tarkoittaa, että näiden tulosten nojalla ei voida sanoa, että tässä ikäryhmässä miesten ja naisten keskusteluaktiivisuus politiikasta olisi erilainen. Yleispäätelmänä voisi olla, että nuorimmassa ikäryhmässä naiset keskustelevat politiikasta ystäviensä kanssa yhtä usein kuin miehet, mutta tätä vanhemmissa ikäryhmissä miehet ovat aktiivisempia politiikasta keskustelijoita kuin naiset.

Elaborointia voi suorittaa ristiintaulukoimalla monia muuttujia keskenään. Tällöin tulee kuitenkin kiinnittää huomiota siihen, että tarkasteltavissa osaryhmissä havaintoyksikköjen määrä ei laske niin pieneksi, että se estää pätevien yleistysten tekemisen. Lisäksi kannattaa ottaa huomioon, että monimutkaisista taulukoista tulee hyvin nopeasti hankalasti hahmotettavia. Käytännössä ristiintaulukointi sopii erityisesti kahden tai enintään kolmen yksittäisen muuttujan välisten yhteyksien tarkasteluun. Jos selittäviä muuttujia on useita ja niissä on kaikissa useita luokkia on syytä harkita muiden välineiden, kuten monimuuttujamenetelmien käyttöä. Käyttämässämme esimerkissä voitaisiin harkita ns. loglineaarisien-mallien käyttöä.

Ristiintaulukon riippumattomuustesti

Kuten aiemmin todettiin täytyy otokseen perustuvat ristiintaulukot alistaa ns. riippumattomuustestille, joka kertoo kuinka todennäköistä on, että riippuvuus on syntynyt ainoastaan otantasattuman vaikutuksesta niin, että muuttujat ovat perusjoukossa toisistaan riippumattomia. Yleisimmin käytetty testi on ns. Pearsonin c^2-testi, joka perustuu havaittujen ja odotettujen frekvenssien vertailuun.

Taulukko 1 on jo aiemmin käytetty esimerkkiristiintaulukko sukupuolen ja poliittisista asioista keskustelemisen aktiivisuuden suhteesta. Taulukkoon on kuitenkin lisätty muutamia lukuja, joita tarvitaan c^2-testin laskemiseksi. Normaalisti näitä tietoja ei tietenkään tarvitse raportoida, koska kaikki tilasto-ohjelmistot laskevat testin automaattisesti tai ainakin pyydettäessä. Taulukkoon on ensinnäkin lisätty "yhteensä" sarake sen oikeaan laitaan. Näitä

lukuja tarvitaan odotettujen frekvenssien laskemiseen. Luvut osoittavat, että 72 vastaajaa ilmoitti keskustelevansa usein politiikasta ystäviensä kanssa, 621 silloin tällöin jne.

Taulukko 1. Aktiivisuus keskustella poliittisista asioita ystävien kanssa sukupuolen mukaan (havaitut frekvenssit, odotetut frekvenssit ja prosenttijakauma).

	Mies	Nainen	Yhteensä
Usein	43 35,0 9 %	29 37,0 6 %	72
Silloin tällöin	323 301,9 68 %	298 319,1 60 %	621
Ei koskaan	108 137,1 23 %	174 144,9 35 %	282
Yhteensä n	100 % 474	100 % 501	 975

Jokaiseen ristiintaulukon soluun on myös lisätty prosenttiluvun yläpuolelle kaksi lukua. Ensimmäinen näistä on havaittu solufrekvenssi (*observed frequency*). Esimerkiksi 43 naista ja 29 miestä vastasi keskustelevansa usein ystäviensä kanssa politiikasta. Tämän luvun alla on solun odotettu frekvenssi (*expected frequency*), eli luku, joka osoittaa kuinka monta vastaajaa solussa todennäköisesti olisi, jos miehet ja naiset eivät eroaisi toisistaan keskustelutapojensa suhteen. Jos keskusteluaktiivisuus todellakin olisi täysin riippumaton sukupuolesta, olisi todennäköisintä, että silloin 35,0 miestä ja 37,0 naista vastaisivat keskustelevansa politiikasta usein. Syy sille, että naisten odotettu frekvenssi on hiukan miesten vastaavaa suurempi, johtuu siitä, että otoksessa on hiukan enemmän naisia kuin miehiä.

Odotetut frekvenssit voidaan laskea seuraavasta kaavasta:

$$E_{ij} = \frac{O_i \times O_j}{N}$$

jossa

E_{ij} = i:nnen rivin ja j:nnen sarakkeen odotettu (Expected) frekvenssi

O_i = i:nnen rivin reunajakauma (eli rivin valinneiden vastaajien kokonaissumma)

O_j = j:nnen sarakkeen reunajakauma (eli sarakkeen valinneiden vastaajien kokonaissumma)

N = havaintojen määrä taulukossa

Esimerkiksi "En koskaan" vastaavien miesten odotettu frekvenssi saadaan laskemalla (282*474)/975 = 137,1. Vastaava luku naisten osalta on 144,9 (=(282*501)/975).

Kuten jo aiemmin todettiin, c^2-testi perustuu havaittujen ja odotettujen frekvenssien eroille. Käytännössä testissä lasketaan ns. c^2-luku, joka kuvastaa sitä, kuinka paljon havaitut ja odotetut frekvenssit eroavat toisistaan. Kun c^2-luku on suuri, eroavat nämä frekvenssit paljon toisistaan ja kun se on pieni, ovat erot havaittujen ja odotettujen frekvenssien välillä pienet. c^2-luku laskentaan seuraavan kaavan avulla:

$$c^2 = \sum_{i-1}^{R} \sum_{j-1}^{C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

jossa

E_{ij} = i:nnen rivin ja j:nnen sarakkeen odotettu frekvenssi

O_{ij} = i:nnen rivin ja j:nnen sarakkeen havaittu frekvenssi

R = Rivien määrä

C = Sarakkeiden määrä

Käytännössä em. kaava tarkoittaa sitä, että jokainen ristiintaulukon solu käydään läpi ja jokaisessa niissä lasketaan ensin odotetun ja havaitun frekvenssin erotus, joka sen jälkeen korotetaan neliöön. Tämän jälkeen saatu tulos jaetaan odotetun frekvenssin arvolla. Lopuksi nämä solukohtaiset arvot lasketaan kaikki yhteen ja lopputuloksena on koko ristiintaulukon c^2-luku.

Esimerkiksi yllä olevan taulukon c^2-luku saadaan kaavasta
(43-35,0)^2/35,0+(29-37,0)^2/37,0+(323-301,9)^2/301,9+ (298-319,1)^2/319,1+(108-137,1)^2/137,1+(174-144,9)^2/144,9=18,4

Lopuksi tarvittava p:n arvo saadaan c^2-jakaumasta, joka löytyy taulukkona esimerkiksi useimpien metodioppaiden liitteenä. Käytännössä tietenkin tilasto-ohjelmistot antavat tarvittavan p-arvon suoraan. Oikean p-arvon saamiseen tarvitaan vielä vapausasteiden määrä. Ristiintaulukossa vapausasteiden määrä saadaan kaavasta (rivien määrä - 1) * (sarakkeiden määrä - 1). Koska esimerkkitaulukossa rivejä on kolme ja sarakkeita on kaksi, on vapausasteiden määrä 2*1 eli 2. Näiden tietojen avulla oikea p-arvo voidaan hakea taulukoista. Tässä tapauksessa se on selvästi pienempi kuin 0,01 eli c^2-testin tulos on erittäin merkitsevä ja näin ollen nollahypoteesi (eli olettamus siitä, että otoksessa havaittu miesten ja naisten ero johtuu pelkästään sattumasta) voidaan hylätä. Miehet todellakin keskustelevat ystäviensä kanssa politiikasta enemmän kuin naiset.

Korrelaatio ja riippuvuusluvut

Korrelaatio

Kahden muuttujan välisen riippuvuuden astetta voidaan nimittää yleisessä merkityksessä korrelaatioksi. Jos korrelaatio on voimakasta, voidaan toisen muuttujan arvoista päätellä toisen muuttujan arvot melko täsmällisesti. Jos korrelaatio on heikko, ei muuttujien välillä ole yhteisvaihtelua. Korrelaatiolla voidaan joskus viitata myös tavallisimmin käytettyyn Pearsonin tulomomenttikorrelaatioon, jota selvitetään seuraavassa. Tämä riippuvuuslukuja koskeva tietovarannon osuus esittelee erilaiset tunnusluvut hyvin tiiviisti keskittyen kuvaamaan esimerkein niiden laskentaperiaatteita.

Pearsonin korrelaatiokerroin, r

Yleisin käytetty korrelaatiota kuvaava tunnusluku on Pearsonin tulomomenttikorrelaatiokerroin (r). Se on vähintään kahden intervalliasteikollisen muuttujan keskinäisen lineaarisen riippuvuuden voimakkuutta kuvaava tilastollinen tunnusluku. Korrelaatiokerroin lasketaan kaavalla

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n s_x s_y}$$

missä

n on lukuparien x_i, y_i lukumäärä

s_x, s_y ovat muuttujien x ja y keskihajonnat ja

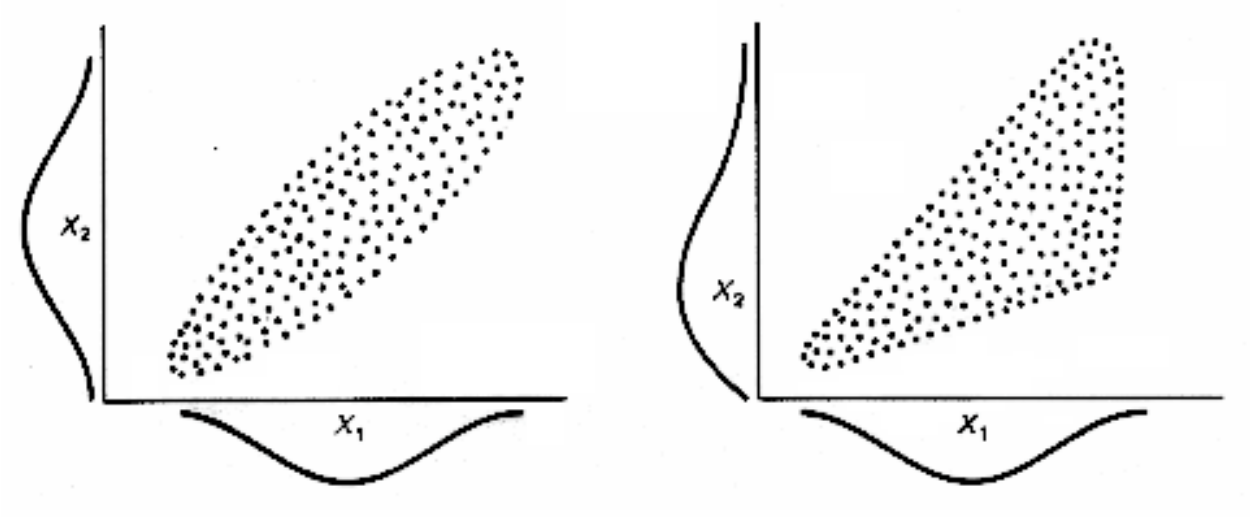
$\bar{x}, \bar{y}$ ovat muuttujien x ja y keskiarvot.

Tulomomenttikorrelaatiokertoimen arvo vaihtelee välillä -1 ... +1. Korrelaatiokertoimen ollessa 0, ei muuttujien välillä ole lineaarista riippuvuutta. Vastaavasti arvoilla (+/-) 1 muuttujien välillä on täydellinen positiivinen / negatiivinen lineaarinen riippuvuus. Täydellisen lineaarisen riippuvuuden tapauksessa muuttujien kaikki arvot sijoittuvat hajontakuviossa samalle suoralle viivalle. Yleensä muuttujien välinen korrelaatiokerroin poikkeaa nollasta. Tämä voi johtua myös sattumasta. Korrelaatiokertoimen merkitsevyystason avulla voidaan arvioida kertoimen tilastollista merkitsevyyttä. Usein raportoidaan myös Pearsonin korrelaatiokertoimen neliö (r^2). Esimerkiksi jos $r^2 = 0.32$, sanotaan, että selittävä muuttuja selittää 32% selitettävän muuttujan varianssista.

Myös korrelaatiokertoimen käyttöön liittyy useita yleisiä tilastoanalyysin sudenkuoppia:

- Korrelaatiokerroin ei automaattisesti anna informaatiota siitä vallitseeko, muuttujien välillä kausaalinen suhde.
- Jos myös muut muuttujat kuin selittävä muuttuja vaikuttavat tarkasteltavaan muuttujaan, silloin kaikki yhteinen kovarianssi, jota niillä on selittävän muuttujan kanssa, luetaan ainoalle selittävälle muuttujalle.
- Jos muuttujien välillä on epälineaarista riippuvuutta, sen määrä tulee huomattavasti aliarvoiduksi.
- Yksittäiset poikkeavat havaintoarvot voivat vaikuttaa suuresti korrelaatiokertoimen arvoon, minkä vuoksi on suositeltavaa aina tulostaa tutkittavien muuttujien hajontakuvio.
- Korrelaatiokerroin voi olla harhaanjohtava, esimerkiksi silloin, jos tarkasteltavat muuttujat eivät ole homoskedastisia.

Muuttujien välinen homoskedastisuus tarkoittaa, että toisella muuttujalla on suunnilleen sama varianssi kaikissa toisen muuttujan luokissa.
Vasemmanpuoleisessa kuvassa oletus on voimassa. Oikenpuoleisessa kuvassa muuttujan x_2 varianssi kasvaa muuttujan x_1 kasvaessa.
Tilannetta voi korjata muuttujamuunnoksilla (kuten logaritmi, neliöjuuri, ...)



Tilannetta voi korjata muuttujamuunnoksilla (kuten logaritmi, neliöjuuri, ...)

Riippuvuusluvut

Riippuvuusluvut luokitteluasteikollisille muuttujille

Kontingenssikerroin (*Contingency Coefficient*)

Kontingenssikerroin C kuvaa kahden luokitteluasteikollisen muuttujan välistä riippuvuutta ja sen määrittelee kaava:

$$C = \sqrt{\frac{c^2}{N + c^2}}$$

missä N on havaintojen määrä

c^2:n testisuureen arvon laskeminen ("khii toiseen testi") on selitetty ristiintaulukoinnin yhteydessä. Korrelaatiokertoimen arvot vaihtelevat välillä 0 ... 1. Kontingenssikertoimen tilastollista merkitsevyyttä testataan c^2-testisuureen avulla, joka on c^2-jakautunut vapausastein (1–1) (m–1), jossa l ja m ovat muuttujien luokkien lukumäärät.

Tarkastellaan sitten esimerkkiä, jossa tutkitaan kahden luokitteluasteikollisen muuttujan Y (vastaajan asuinmaakunta) ja X (vastaajan äidinkieli) riippuvuutta henkilöaineistossa. Näiden muuttujien kaksiulotteinen yhteisfrekvenssijakauma on seuraava:

	äidinkieli (X)		
asuinmaakunta (Y)		suomi	ruotsi
	uusimaa	(a) 76	(b) 13
	muu maakunta	(c) 229	(d) 5

Taulukon perusteella c^2 arvoksi saadaan 19.053, joten kontingenssikertoimeksi saadaan (sijoittamalla edellä esitettyyn kaavaan):

$$C = \sqrt{\frac{19{,}053}{323 + 19{,}053}} \approx 0{,}236$$

Asuinmaakunnalla ja äidinkielellä olisi siis verrattain pieni riippuvuus.

Kontingenssikertoimen käyttökelpoisuus empiirisenä riippuvuuslukuna perustuu ensisijaisesti siihen, että muuttujilta ei vaadita kuin luokitteluasteikollinen mittaustarkkuus. Myöskään jakaumaoletuksia ei ole. Kontingenssikertoimella on muutamia heikkouksia:

- Kontingenssikerroin ei voi saada negatiivisia arvoja, joten sen avulla ei voi päätellä riippuvuuden suuntaa
- Kontingenssikertoimien keskinäinen vertailu ei ole mielekästä, mikäli ne perustuvat eri kokoisiin taulukoihin.
- Suurin arvo, jonka kontingenssikerroin voi saavuttaa, on aina pienempi kuin 1. Lisäksi taulukoille, joiden rivi- ja sarakemäärät ovat yhtä suuret, suurin arvo on *sqrt((r-1)/r)*. Esimerkiksi taulukolle, jossa on kaksi saraketta ja riviä suurin saavutettavissa oleva kontingenssikertoimen arvo on siis v((2-1)/2) = 0.71.
- Kontingenssikerroin ei ole vertailukelpoinen järjestyskorrelaatiokertoimien eikä Pearsonin korrelaatiokertoimen kanssa

Yulen Q

Käytetään kahden luokitteluasteikollisen dikotomisen muuttujan riippuvuuden tarkastelussa. Yulen Q määritellään 2x2 yhteisfrekvenssijakaumataulukon diagonaalisolujen tulojen erotuksen ja summan osamääränä. Siis:

$$Q = \frac{(ad - bc)}{(ad + bc)}$$

Esimerkkitaulukon tapauksessa `Q`:n arvoksi saadaan:

$$Q = \frac{(76 \times 5 - 13 \times 229)}{(76 \times 5 + 13 \times 229)} \approx -0{,}774$$

Phi kerroin, j

Tätäkin kerroita käytetään kahden luokitteluasteikollisen dikotomisen muuttujan riippuvuuden tarkastelussa. Jakaumaoletuksena vaaditaan, että muuttujat olisivat luonnollisesti dikotomisia. Phi -kerroin lasketaan kaavalla:

$$j = \frac{(ad - bc)}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

Esimerkkitaulukon tapauksessa `Phi`:n arvoksi saadaan:

$$j = \frac{(76 \times 5 - 13 \times 229)}{\sqrt{(76+13)(229+5)(76+229)(13+5)}} \approx -0{,}243$$

Tuloksesta huomataan, että tarkasteltavien muuttujien riippuvuus on suunnilleen samaa luokkaa kuin kontingenssikertoimella laskettuna. Lisäksi saadaan selville riippuvuuden suunta, joka on negatiivinen.

Riski (Relative Risk, RR)

Tämä suhdeluku sopii niin ikään kahden luokitteluasteikollisen dikotomisen muuttujan riippuvuuden tarkasteluun. Tunnusluku on yleinen terveystieteissä, mutta sopii myös sosiaalitieteiden tilanteisiin, joissa toinen muuttuja on käsittely/syy ja toinen vaikutus/seuraus. Riski ja ristitulosuhde (odds ratio) lasketaan kaavoilla:

$$RR = \frac{\frac{a}{a+c}}{\frac{b}{b+d}} = \frac{\frac{76}{76+229}}{\frac{13}{13+5}} \approx 0{,}345$$

$$OR = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{\frac{76}{229}}{\frac{13}{5}} \approx 0{,}128$$

Cramerin V

Cramerin V on suosittu c^2-perustainen riippuvuusluku, jota käytetään kahden luokitteluasteikollisen muuttujan riippuvuuden tarkastelussa. Se lasketaan kaavalla:

$$V = \sqrt{\frac{c^2}{NM}}$$

missä

N on havaintojen lukumäärä ja

M on minimi sarakkeiden ja rivien lukumäärästä –1.

V vaihtelee välillä 0 ... 1, riippumatta yhteisjakaumataulukon koosta. Koska V:n otosjakauma tunnetaan, sen keskivirhe ja merkitsevyys voidaan laskea. Esimerkkitaulukossa sen arvoksi saadaan:

$$V = \sqrt{\frac{19{,}053}{323 \times (2-1)}} \approx 0{,}243$$

Lambda, l

Lambdaa käytetään kahden luokitteluasteikollisen muuttujan riippuvuuden tarkastelussa ja sen symmetrinen arvo lasketaan kaavalla:

$$l = \frac{\sum f_r + \sum f_c - (F_r + F_c)}{2N - (F_r + F_c)}$$

missä

f_r on suurin frekvenssi rivimuuttujan luokassa r

f_c on suurin frekvenssi sarakemuuttujan luokassa c

F_r on suurin frekvenssi rivimuuttujan reunajakaumassa

F_c on suurin frekvenssi sarakemuuttujan reunajakaumassa ja

N on havaintojen määrä.

Lambda vaihtelee välillä 0 ... 1. Se kertoo, kuinka tarkasti voidaan ennustaa toisen muuttujan arvo, kun toisen muuttujan arvo tiedetään. Koska lambdalla on tunnettu otosjakauma, voidaan sen keskivirhe ja merkitsevyys laskea. Tilastolliset ohjelmistot, kuten esimerkiksi SPSS, laskevat asymptoottisen keskivirheen (*ASE, Asymptotic Standart Error*).

Esimerkkitaulukossa lambdan arvoksi saadaan:

$$l = \frac{(76+229)+(229+13)-(234+305)}{2\times 323-(234+305)} \approx 0{,}075$$

Lambdasta on myös asymmetrinen versio, jossa täytyy määritellä kumpi muuttuja on selittäjä ja kumpi selitettävä. Kaavaksi muodostuu tällöin:

$$l = \frac{\sum f_i - F_d}{N - F_d}$$

missä

f_i on suurin frekvenssi selittävän muuttujan luokassa i

F_d on suurin frekvenssi selitettävän muuttujan reunajakaumassa d ja

N on havaintojen lukumäärä.

Jos halutaan selittää asuinmaakuntaa (selitettävä) äidinkielellä (selittäjä), tulokseksi saadaan:

$$l = \frac{229+13-234}{323-234} \approx 0{,}090$$

Epävarmuuskerroin (*Uncertainty, Entropy Coefficient*)

Epävarmuuskerroin on lambdaa vastaava tunnusluku, joka vaihtelee välillä 0 ... 1. Sen keskivirhe ja merkitsevyys voidaan laskea. Tulkintana on, lambdaa vastaavasti, ennuste toisen muuttujan arvosta, jos tiedetään toisen muuttujan arvo. Epävarmuuskerroin on asymmetrinen riippuvuusluku. Kertoimen arvo riippuu siis siitä, kumpi muuttuja on selittävä/selitettävä. Useat tilastolliset ohjelmistot laskevat myös symmetrisen epävarmuuskertoimen, joka on keskiarvo kahdesta asymmetrisestä kertoimesta. Merkinnässä UC(R|C) rivimuuttuja (Y) on selitettävä ja sarakemuuttuja (X) selittäjä.

$$UC_{R|C} = \frac{h(x)+h(y)-h(xy)}{h(y)}$$

$$UC_{C|R} = \frac{h(y)+h(x)-h(xy)}{h(x)}$$

$$UC_{sym} = 2\frac{h(x)+h(y)-h(xy)}{h(x)+h(y)}$$

missä

x on sarakemuuttuja

y on rivimuuttuja

n on otoskoko

r_j on rivisumma riveille 1…R

c_k on sarakesumma sarakkeille 1…C

n_{jk} on solufrekvenssi rivillä j, sarakkeella k

ln on luonnollinen logaritmi

$$h(x) = -\sum_{j=1}^{R} \frac{r_j}{n} \times \ln\left(\frac{r_j}{n}\right)$$

$$h(y) = -\sum_{k=1}^{C} \frac{C_k}{n} \times \ln\left(\frac{c_k}{n}\right)$$

$$h(xy) = -\sum_{j=1}^{R}\sum_{k=1}^{C} \frac{n_{jk}}{n} \times \ln\left(\frac{n_{jk}}{n}\right)$$

Riippuvuusluvut järjestysasteikollisille muuttujille

Tarkastellaan kahta järjestysasteikollista muuttujaa X (vastaajan koulutus) ja Y (vastaajan bruttotulot/kk) henkilöaineistossa. Näiden muuttujien kaksiulotteinen yhteisfrekvenssijakauma on seuraava:

		alle 7000 mk	7000-12000 mk	yli 12000 mk
koulutus (Y)	perusaste	(a) 33	(b) 25	(c) 3
	väh. keskiaste	(d) 54	(e) 102	(f) 56

Parien käsite

Ylläolevan taulukon perusteella voidaan määritellä:

Parin tyyppi	**Parien lukumäärä**	**Symboli**
samansuuntainen	a(e+f) + b(f)	P
vastakkaissuuntainen	c(d+e) + b(d)	Q
sidos muuttujassa X	ad + be + cf	X_o
sidos muuttujassa Y	a(b+c) + bc + d(e+f) + ef	Y_o

Spearmanin rho, r

Spearmanin r on useimmin käytetty järjestyskorrelaatiokerroin vähintään järjestysasteikollisten muuttujien välillä. Kertoimen laskenta aloitetaan järjestämällä aineisto suuruusjärjestykseen toisen muuttujan suhteen. Tämän jälkeen annetaan muuttujille järjestysluvut (*rank*) `(1, 2, ..., N)` muuttujan arvojen mukaan ja lasketaan havaintopareittain järjestyslukujen erotus `D`. Itse kerroin saadaan tällöin kaavasta:

$$r = 1 - \frac{6\sum_{i=1}^{N} D_i^2}{N(N^2-1)}$$

Voidaan osoittaa, että Spearmanin r on järjestysluvuista laskettu Pearsonin korrelaatiokerroin. Laskettaessa r:ta edellytetään, että muuttujien järjestysluvuissa ei esiinny tasatuloksia eli sidoksia. Pieni sidosmäärä voidaan käsitellä käyttämällä tasatuloksista järjestyslukujen keskiarvoja. Esimerkkiaineistolla r:n arvoksi tulee 0.345.

Kendallin tau-b ja tau-c, t

Kendallin tau-b on riippuvuusluku, jonka laskenta perustuu saman- ja vastakkaissuuntaisten parien erotukseen jaettuna X- ja Y-muuttujien ei sidottujen parien lukumäärien geometrisellä keskiarvolla eli:

$$t_b = \frac{(P-Q)}{\sqrt{(P+Q+Y_0)(P+Q+X_0)}} = \frac{(6614-1818)}{\sqrt{(6614+1818+15243)(6614+1818+4500)}} \approx 0{,}274$$

Tau-b:tä käytetään usein 2x2 jakaumataulun tilanteessa, mutta se sopii myös useampiluokkaisiin muuttujiin. Useampiluokkaisille muuttujille on kehitetty variaationa tau-c, joka lasketaan kaavalla

$$t_c = (P+Q) \times \frac{2M}{N^2(M-1)} = (6614+1818) \times \frac{2 \times 2}{273^2 \times (2-1)} \approx 0{,}257$$

missä

M on minimi sarakkeiden ja rivien lukumäärästä ja N on havaintojen lukumäärä.

Goodmanin ja Kruskalin gamma, g

Gamma on symmetrinen riippuvuuluku, joka vaihtelee välillä -1 ... +1. Se perustuu saman- ja vastakkaissuuntaisten parien väliseen eroon, joka lasketaan kaavalla:

$$g = \frac{P-Q}{P+Q} = \frac{6614-1818}{6614+1818} \approx 0{,}569$$

Koska gammalla on tunnettu otosjakauma, joka lähenee suurilla otoksilla normaalijakaumaa, voidaan sen keskivirhe ja merkitsevyys laskea.

Osittaiskorrelaatio

Osittaiskorrelaatio on kahden muuttujan välinen korrelaatio, kun yhden tai useamman muuttujan vaikutus on poistettu (vakioitu). Tämä voidaan tehdä myös laskemalla muuttujien korrelaatio kolmannen tekijän osajoukoissa. Esimerkiksi jäätelon kulutus ja hukkumiskuolemien määrä korreloivat voimakkaasti. Muuttujien välinen korrelaatio johtuu siitä, että molemmat korreloivat lämpötilan kanssa. Sisällöllisesti mielekäs korrelaatio saadaan laskemalla osittaiskorrelaatio jäätelön kulutuksen ja hukkumiskuolemien määrän välillä, kun lämpötilan vaikutus on poistettu. Osittaiskorrelaatiosta ei kuitenkaan näy, onko alkuperäinen kahden muuttujan yhteys samanlainen vai erilainen vakioitavan muuttujan eri arvoilla. Tulkinnan kannalta on tärkeää tietää muuttujien aikajärjestys. Osittaiskorrelaatiota merkitään usein luvulla $r_{xy.z}$ niin, että vakioitava muuttuja erotetaan pisteellä alkuperäisen korrelaation muuttujista. Laskentakaava osoittaiskorrelaatiolle on:

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1-r_{xz})^2(1-r_{yz})^2}}$$

Myös osittaiskorrelaatio kuvaa muuttujien lineaarista yhteyttä, joka vaihtelee –1 ... +1 välillä. Osittaiskorrelaatio voidaan yleistää useamman muuttujan samanaikaiseen vakiointiin lisäämällä vakioitavia muuttujia ja soveltamalla kaavaa useita kertoja.

-- HARJOITUSTEHTÄVIÄ --

1. Kaksi professoria asettaa tutkijan virkaan hakijat seuraavaan paremmuusjärjestykseen.

hakija	A	B	C	D	E	F	G
prof. A	3	1	7	2	4	5	6
prof. B	1	3	2	4	7	6	5

Määritä Spearmanin rho. Ovatko professorien mielipiteet samansuuntaiset?

2. Laske Pearsonin korrelaatiokerroin oheisesta aineistosta. Jos aineistoon lisätään havainto (20,100), niin miten korrelaatiokerroin muuttuu?

ikä vuosina:															
5	8	9	10	10	11	11	12	12	13	14	14	14	15	18	18
testipistemäärä:															
70	148	250	238	245	162	215	341	303	325	270	346	227	302	378	395

Hypoteesien testaus

Useimmat määrällisen analyysin menetelmät perustuvat tilastollisten hypoteesien testaukseen. Aina tätä ei ole kuitenkaan helppo huomata, koska hypoteesit ovat menetelmien käytäntöön "sisäänrakennettuja" ja niitä ei tavallisesti tuoda eksplisiittisesti esille, vaan tutkijan oletetaan ymmärtävän hypoteesien testauksen periaatteet. Tämän vuoksi hypoteesien testauksen perusperiaatteiden ja niihin liittyvien ongelmien ymmärtäminen on olennaista, jotta menetelmien antamien tulosten mielekäs tulkinta olisi mahdollinen. Tässä osiossa esitetään ensin hypoteesien testauksen ns. "oppikirjamalli". Lopussa käydään vielä lyhyesti läpi tilastolliseen hypoteesien testaukseen liittyviä ongelmia.

Hypoteesien testaus etenee seuraavien viiden vaiheen kautta. Nämä ovat 1) hypoteesien valinta, 2) sopivan tilastollisen testin valinta, 3) merkitsevyystason valinta, 4) testin suorittaminen ja 5) lopullisen päätöksen tekeminen.

Hypoteesien valinta

Tutkimusta tehdessä tutkijalla on joitakin oletuksia siitä, minkälaisia eroja tai samankaltaisuuksia perusjoukosta mitattujen muuttujien välillä löytyy. Nämä oletukset perustuvat useimmiten teoreettiselle keskustelulle tai aikaisemmassa tutkimuksessa löydetyille havainnoille. Kuvitellaan esimerkiksi, että tutkija olisi kerännyt satunnaisotoksen avulla aineiston suomalaisten ansaitsemista palkoista jollain tietyllä talouden sektorilla. Tutkija on kiinnostunut naisten ja miesten välisistä palkkaeroista ja niihin vaikuttavista syistä. Ensimmäisenä hänen kannattaa tutkia, eroavatko miesten ja naisten keskipalkat toisistaan tällä sektorilla. Merkitään, että μ_n tarkoittaa naisten keskipalkkaa ja μ_m miesten keskipalkkaa tutkimuksen perusjoukossa.

Tutkijan lähtöoletuksena on, että sukupuolet eroavat palkkatasoltaan toisistaan. Hypoteesin testauksen yleisenä ideana on, että tutkija muotoilee hypoteesin, joka on vastoin hänen alkuperäistä oletustaan ja sen jälkeen tutkii, voidaanko tämä hypoteesi kumota empiirisen aineiston perusteella. Tätä alkuperäisen oletuksen vastaista hypoteesia kutsutaan nimellä **nollahypoteesi** (*null hypothesis*). Nollahypoteesia on tapana merkitä H_0.

Nollahypoteesin lisäksi tutkija tarvitsee **vastahypoteesin** (*alternative hypothesis*), joka hyväksytään, jos nollahypoteesi pystytään kumoamaan. Tätä hypoteesia merkitään H_1.

Nyt hypoteesit voidaan merkitä formaalisti seuraavalla tavalla:

$$H_0 : \mu_n = \mu_m$$

$$H_1 : \mu_n \neq \mu_m$$

Toisin sanoen tutkijan nollahypoteesi on, että miesten ja naisten keskipalkat ovat samansuuruiset perusjoukossa ja vaihtoehtoisen hypoteesin mukaan ne eroavat toisistaan. Kyseessä on ns. **kaksisuuntainen** (*two-tailed*) hypoteesin testaus, koska tutkija ei tee oletusta siitä, onko miesten vai naisten keskipalkat suurempia, vaan olettaa ainoastaan, että ne eroavat toisistaan.

Vaihtoehtoinen mahdollisuus olisi ottaa lähtökohdaksi olettamus naisten keskipalkan pienemmyydestä. Tällöin olisi kyse **yksisuuntaisesta** (*one-tailed*) hypoteesin testauksesta. Tämä tilanne voidaan esittää seuraavasti:

$$H_0 : \mu_n = \mu_m$$

$$H_1 : \mu_n < \mu_m$$

Edellä esitetyt hypoteesit ovat vain esimerkkejä mahdollisista nolla- ja vastahypoteeseista. Sopivien hypoteesien valinta perustuu aina tutkimusongelmaan, teoriaan ja aikaisempaan tutkimukseen. Niiden ei tarvitse liittyä tutkittavien arvojen saman- tai erisuuruisuuteen, vaan

testauksen periaate toimii myös muiden tilastollisten suureiden kanssa. Esimerkiksi tutkittaessa koulutuksen ja ansiotason suhdetta voi nollahypoteesi olla, että näiden tekijöiden välillä ei ole perusjoukossa korrelaatiota, ja vaihtoehtoinen hypoteesi, että niiden välillä on positiivinen korrelaatio.

Tilastollisen testin valinta

Sopivan tilastollisen testin valinta riippuu tutkimusongelmasta, muuttujien mittaustasosta, toisiinsa verrattavien ryhmien määrästä ja monesta muusta asiasta. Erilaisia testejä on olemassa suuri määrä ja niiden tarkempi käsittely ei tässä yhteydessä ole tarkoituksenmukaista, vaan pyrkimyksenä on välittää kuva testauksen yleisistä periaatteista. Kannattaa huomata, että useimmat määrällisen analyysin menetelmät sisältävät "automaattisesti" hypoteesien testausta. Esimerkiksi monen muuttujan regressioanalyysin yhteydessä testataan jokaisen selittävän muuttujan osalta, eroaako niiden regressiokerroin tilastollisesti merkitsevästi nollasta. Tätä varten tutkijan ei kuitenkaan tarvitse muotoilla jokaisen muuttujan osalta omia hypoteesejaan, vaan analyysiohjelmistot tekevät testauksen automaattisesti ja tutkijan tehtävä on kiinnittää huomiota testien antamien tulosten oikeaan tulkintaan. Vastaava tilanne on ristiintaulukoinnin tilastollisen testauksen yhteydessä. Tässä yhteydessä käytetyn ns. c^2-testin nollahypoteesi on, että ristiintaulukon kaksi muuttujaa ovat toisistaan riippumattomia, ja jos testitulos on suotuisa, voidaan nollahypoteesi hylätä ja todeta, että muuttujien välillä on yhteys perusjoukossa.

Merkitsevyystason valinta

Merkitsevyystason valinta määrittää todennäköisyyden sille, että tutkija hylkää nollahypoteesin, vaikka se on todellisuudessa pätevä. Kyse on siis virheellisen valinnan riskistä. Tämän takia merkitsevyystasoa kutsutaan joskus myös riskitasoksi. Tilastollisen päättelyn avulla ei voida koskaan sanoa varmuudella, että jokin hypoteesi on tosi tai epätosi, vaan kyse on aina siitä, millä todennäköisyydellä tutkija on valmis hylkäämään hypoteesin.

Yleisesti tieteellisessä tutkimuksessa käytetään 0,05 (eli 5%) tai 0.01 (eli 1%) riskitasoa. Jos kriteerinä käytetään 5% riskitasoa, tarkoittaa tämä, että tulos on tutkimuksen perusjoukossa 95% varmuudella pätevä, mutta samalla, että virheen todennäköisyys on 5%. Tämä tarkoittaa toisin sanoen sitä, että jos perusjoukosta poimittaisiin 100 samankokoista satunnaisotosta, näissä 95:ssä nollahypoteesi hylättäisiin ja 5:ssä se jäisi voimaan. Jos riskitasona käytettäisiin 1%-tasoa, vain yhdessä sadasta otoksesta nollahypoteesi jäisi voimaan.

Jokaisen tilastollisen testin tuloksena saadaan ns. p-arvo, joka ilmoittaa virheellisen päätelmän todennäköisyyden. Jos p-arvo on alle 0,05 on tapana puhua tuloksesta tilastollisesti "melkein merkitsevänä", jos se on alle 0,01 tilastollisesti "merkitsevänä" ja jos se on alle 0,001 tilastollisesti "erittäin merkitsevänä". Taulukoissa on tapana merkitä "melkein merkitsevät" tulokset yhdellä tähdellä (*), "merkitsevät" tulokset kahdella (**) ja "erittäin merkitsevät" tulokset kolmella tähdellä (***).

Kannattaa muistaa, että usein käytetyt 5%- ja 1%-riskitasot ovat täysin sopimuksenvaraisia. Periaatteessa rajat voisivat olla esimerkiksi 6% ja 2%. Tilastotieteen teoriasta ei löydy 5%- ja 1%-luottamustasoille mitään erityistä perustetta, vaan ne ovat vain vuosien saatossa muodostuneet käytännöiksi. Tämän vuoksi on tärkeää, että tutkija kiinnittää testitulosten lisäksi aina huomiota myös tulosten sisällölliseen merkityksellisyyteen.

Testin suorittaminen

Kuten jo aiemmin todettiin, tilastolliseen hypoteesien testaukseen on valtava määrä erilaisia testejä tutkimusongelman ja muuttujien mittaustason luonteesta riippuen. Yksinkertaiset testit voidaan tehdä laskukoneen ja tilastollisten taulukoiden tai taulukkolaskentaohjelman

avulla. Käytännössä on kuitenkin paras käyttää tähän tarkoitukseen tehtyjä tilastollisia tietokoneohjelmistoja, joihin kaikki tärkeimmät tilastomenetelmät on valmiiksi ohjelmoitu. Näin laskuvirheiden mahdollisuus pienenee, ja tutkija voi keskittyä tulosten oikeaan tulkintaan.

Päätös nollahypoteesin hylkäämisestä tai hyväksymisestä

Kun tilastolliset testisuureet ja niiden todennäköisyydet on laskettu, on tutkijan tehtävä päätös siitä hylätäänkö nollahypoteesi vai ei. Puhtaasti tilastolliselta kannalta katsottuna tämä tehtävä on helppo. Jos tilastollisen testin antama tulos on pienempi kuin valittu riskitaso, hylätään nollahypoteesi ja todetaan, että vastahypoteesi sai tukea. Muutoin todetaan, että nollahypoteesia ei voitu kumota. Pelkkä tilastotieteellinen tarkastelu ei kuitenkaan sellaisenaan riitä, vaan tutkijan on lähestyttävä asiaa myös sisällöllisesti tutkimusongelman kannalta. Nollahypoteesin hylkäämiseen tai hyväksymiseen liittyy myös muita tekijöitä. Kuvitellaan esimerkki, jossa tutkija on kehittänyt johonkin vakavaan sairauteen uuden lääkkeen. Hänen tutkimustuloksensa osoittavat, että lääkettä käyttäneet testiryhmän jäsenet selviävät hengissä sairaudesta suuremmalla todennäköisyydellä kuin plaseboa nauttineet kontrolliryhmän jäsenet. Tilastollinen testi kuitenkin osoittaa, että ero on merkitsevä vain 6% riskitasolla. Tässä tapauksessa tutkijan tuskin kannattaa luopua lääkkeen jatkotutkimuksista ainoastaan sen takia, että testitulosten p-arvot eivät olleet alle 5%.

Nollahypoteesin hylkäyksen tai hyväksymisen yhteydessä on itse asiassa mahdollisuus kahteen eri virheeseen. Niin sanottu **hylkäämisvirhe** (*type I error*) tapahtuu silloin, kun nollahypoteesi hylätään, vaikka se itse asiassa on tosi. Tämän hylkäämisvirheen todennäköisyys on, kuten jo aiemmin todettiin, valittu riskitaso. Toista virhettä voidaan kutsua **hyväksymisvirheeksi** (*type II error*). Hyväksymisvirhe tapahtuu silloin, kun nollahypoteesi hyväksytään, silloin kun se on epätosi. On tärkeää ymmärtää, että hyväksymisvirheen ja hylkäämisvirheen todennäköisyydet ovat toisistaan riippuvaisia. Jos tutkija asettaa hylkäämisvirheelle erittäin matalan rajan (esimerkiksi 0,1%) kasvaa hyväksymisvirheen riski ja päinvastoin.

Tilastollisten testien kritiikki

Yhteiskuntatieteissä hypoteesien testauksen lähestymistapaa kohtaan on usein esitetty kritiikkiä. Kritiikki voidaan jakaa ainakin kolmeen ryhmään: 1)teknisluontoiseen, 2)tieteenfilosofiseen ja 3)tutkimuksen käytäntöihin liittyvään kritiikkiin (ks. esim. Henkel 1976). Tekninen kritiikki lähtee siitä, että tilastollisten testien taustalla on tiettyjä olettamuksia, joiden pätevyys käytännön tutkimuksessa voidaan usein asettaa kyseenalaiseksi. Näistä oletuksista tärkein liittyy otoksen luonteeseen. Tilastollisen päättelyn menetelmä pätee silloin kun, aineistona käytetään aitoa satunnaisotantaa jostain perusjoukosta (ks. otantamenetelmät). Monet yhteiskuntatieteellisen tutkimuksen aineistot eivät kuitenkaan ole tiukassa merkityksessä satunnaisotoksia. Esimerkiksi postikyselyssä vastaamatta jättäneiden osuus voi olla useita kymmeniä prosentteja lähetettyjen lomakkeiden määrästä. Näin suuri kato aiheuttaa väistämättä ongelmia otoksen satunnaisuuden kannalta.

Tieteenfilosofiselta kannalta hypoteesien tilastollista testausta voi kritisoida monelta kannalta. Ehkä tärkein kritiikki kohdistuu nollahypoteesin luonteeseen. Usein hypoteesin testauksen yhteydessä oletetaan nollahypoteesina, että jonkin parametrin arvo on nolla. Tämä oletus on kuitenkin triviaali. On hyvin epätodennäköistä, että minkään parametrin arvo on todellisuudessa tasan nolla, joten nollahypoteesin hylkääminen ei käytännössä lisää tutkimusongelman kannalta tietoamme paljoakaan. Samoin aiemmin esitettyä nollahypoteesia, jonka mukaan miesten ja naisten keskipalkat ovat yhtä suuret, voidaan kritisoida samalla tavalla. Ei tutkija todellisuudessa oleta, että keskipalkat ovat pennilleen (sentilleen!) täysin samat.

Sisällöllisesti mielekäs kysymys on, kuinka paljon keskipalkat eroavat toisistaan ja mihin suuntaan.

Yhteiskuntatieteellisen tutkimuksen käytännöistä lähtevä kritiikki voidaan jakaa kahteen luokkaan. Ensimmäinen liittyy tilastollisten testien väärään käyttötapaan. Usein tutkijat eivät täysin ymmärrä testauksen luonnetta, osaa valita oikeaa testiä tai osaa tulkita niiden tuloksia. Tähän kritiikkiin on tietysti ratkaisuna tutkijoiden parempi koulutus ja tutkimusten kriittinen arviointi menetelmien käytön osalta. Toinen tutkimuksen käytäntöihin liittyvä kritiikki liittyy tilastollisten testien "ritualistiseen" käyttöön. Tämä tarkoittaa liiallista keskittymistä pelkästään tilastollisten testien tulokseen sisällöllisten tulkintojen kustannuksella. Tilastollisten testien pitäisi olla lähtökohta tulosten tulkinnalle, ei päätepiste. Useimmiten tutkijaa ei kiinnosta pelkästään se, eroavatko kaksi ryhmää toisistaan jonkun tekijän suhteen, vaan se kuinka paljon ryhmät eroavat ja mikä merkitys tällä eron suuruudella on tutkimusongelman kannalta.

Tämä tilastollisten testien ja hypoteesin testaamisen kritiikki tärkeää ottaa huomioon. Vasta-argumentti esitetylle kritiikille korostaa pragmaattista lähestymistapaa. Sen mukaan tilastolliseen testaukseen liittyy ongelmia, mutta se ei tarkoita, että testauksesta täytyisi luopua kokonaan, vaan että ongelmat on otettava huomioon tutkimusprosessissa. Pragmaattisen kannan mukaan tilastollisilla testeillä saadaan niihin liittyvistä ongelmista huolimatta informaatiota, joka on otettava huomioon tulosten tulkinnassa. Niiden avulla pystytään esimerkiksi arvioimaan (ainakin suunnilleen), kuinka todennäköisesti otoksessa havaitut erovaisuudet ovat sattumasta johtuvia, eivätkä näin ollen kuvaa mitään systemaattista yhteyttä tutkittavien tekijöiden välillä. Tämän vuoksi tutkimuksissa on hyvä raportoida testien tulokset, koska silloin ainakin lukija voi käyttää tuloksia hyväkseen arvioidessaan tulosten merkityksellisyyttä.

Lähteet

- Henkel, Ramon E. (1976): *Tests of Significance*. Sage, Beverly Hills.

Varianssianalyysi

Varianssianalyysia (*analysis of variance* tai *ANOVA*) käytetään tutkittaessa eroavatko kahden tai useamman ryhmän keskiarvot tilastollisesti merkitsevästi toisistaan. Varianssianalyysilla voidaan esimerkiksi tutkia eroavatko naisten ja miesten keskipalkat toisistaan jossakin yrityksessä tai ovatko eri maahanmuuttajaryhmiin kuuluvien koululaisten todistusten arvosanat keskiarvoiltaan toisistaan poikkeavia. Varianssianalyysia on perinteisesti pidetty kokeellisen analyysin perusmenetelmänä ja sen käyttö onkin ollut yleistä esimerkiksi lääketieteessä. Sillä on kuitenkin useita sovellusmahdollisuuksia myös yhteiskuntatieteiden aloilla.

Varianssianalyysin käyttöön liittyy useita laajennusmahdollisuuksia. Tässä yhteydessä keskitytään ns. yksisuuntaiseen varianssianalyysiin, joka on vaihtoehdoista yksinkertaisin. Lopussa esitellään lyhyesti myös kaksisuuntainen varianssianalyysi, kovarianssianalyysi ja monen muuttujan varianssianalyysi (*MANOVA*).

Yksisuuntainen varianssianalyysi

Yksisuuntainen varianssianalyysi (*one-way analysis of variance*) on varianssianalyysin muodoista yksinkertaisin. Koska varianssianalyysissa tarkastellaan selitettävien muuttujien on ryhmäkeskiarvoja, täytyy selitettävän muuttujan olla sellainen, että siitä on järkevää laskea aritmeettinen keskiarvo (eli käytännössä välimatka- tai suhdelukuasteikon muuttuja, ks. muuttujien mittaustaso ja keskiluvut). Yksisuuntaisessa varianssianalyysissa on vain yksi selittävä muuttuja. Koska tämä muuttuja kuvaa havaintoyksikköjen jakautumista luokkiin, on sen mittaustaso oltava joko luokittelu- tai järjestysasteikko.

Varianssianalyysin avulla tutkitaan sitä, ovatko selitettävän muuttujan keskiarvot tilastollisesti merkitsevästi erisuuruisia selittävän muuttujan eri luokissa. Analyysin lähtöoletuksena eli nollahypoteesina (ks. hypoteesien testaus) on, että kiinnostuksen kohteena olevien luokkien keskiarvot ovat yhtä suuret. Jos varianssianalyysin tuloksena nollahypoteesi voidaan hylätä, selitettävän muuttujan keskiarvojen välillä on eroja selittävän muuttujan eri luokissa.

Käytännössä varianssianalyysi perustuu siihen, että selitettävän muuttujan varianssi (ks. hajontaluvut) jaetaan kahteen osaan. Näistä ensimmäinen mittaa luokkien sisäistä hajontaa ja toinen luokkakeskiarvojen välistä hajontaa. Jos nämä kaksi varianssia eivät eroa kovinkaan paljon toisistaan, on todennäköistä, että eri luokkien saamat keskiarvot ovat peräisin samankaltaisesta jakaumasta. Tällöin niiden välillä ei ole tilastollisesti merkitsevää eroa. Jos taas nämä kaksi varianssia eroavat toisistaan tarpeeksi nollahypoteesi voidaan hylätä. Tilastollisena testinä varianssianalyysissa käytetään ns. F-testiä, joka kertoo millä todennäköisyydellä nollahypoteesi ryhmäkeskiarvojen yhtäläisyydestä voidaan hylätä.

Esimerkki yksisuuntaisesta varianssianalyysista

Seuraavassa esimerkissä tutkitaan suomalaisten suhtautumista tuloerojen pienentämiseen tai niiden kasvattamiseen. Aineistona käytetään vuoden 1996 World Values Survey -tutkimuksen Suomen osa-aineistoa (ks. verkosta aineistonkuvaus FSD0153). Kyselyssä pyydettiin vastaajia kertomaan mielipiteensä jatkumolla 1-10, jossa pienet arvot kuvastivat vastaajan halua tasata tuloeroja pienemmäksi ja suuret arvot vastaajan halua lisätä tuloeroja (kysymys V125). Asteikon ääripäitä kuvaavat tekstit olivat "tulotaso pitäisi maassamme saada tasaisemmaksi" ja "tarvitsemme suurempia tuloeroja palkitaksemme enemmän kansalaisten yritteliäisyyttä".

Skaalan keskimmäiset vaihtoehdot olivat 5 ja 6, jolloin kaikkien vastaajien keskiarvo 4,16 oli tuloerojen voimakkaampaa tasaamista kannattavalla puolella.

Selittävänä muuttujana esimerkissä on vastaajien subjektiivinen luokka-asema eli tarkemmin ilmaistuna heidän oma näkemyksensä siitä, mihin yhteiskuntaluokkaan he kuuluvat (V226). Kysymyksessä annettiin vaihtoehdoksi viisi erilaista yhteiskuntaluokkaa: "yläluokka", "ylempi keskiluokka", "alempi keskiluokka", "ylempi työväenluokka" ja "alempi työväenluokka". Koska vastaajista vain neljä määritteli itsensä yläluokkaan kuuluvaksi, on seuraavassa analyysissa vaihtoehdot "yläluokka" ja "ylempi keskiluokka" yhdistetty (ks. muuttujien uudelleenkoodaus).

Varianssianalyysin tulokset on esitetty taulukossa 1. Taulukon yläosa kuvaa tuloeromuuttujan keskiarvoja selittävän muuttujan luokissa. Itsensä yläluokkaan tai ylempään keskiluokkaan kuuluvaksi määrittelevät vastaajat suhtautuvat tuloerojen kasvattamiseen suopeimmin (keskiarvo 5,33). Eniten tuloerojen pienentämisen kannalla ovat alempaan työväenluokkaan kuuluvat vastaajat (keskiarvo 3,26).

Vastaajan yhteiskuntaluokka	**Suhtautuminen tuloeroihin (ryhmäkeskiarvo)**
Yläluokka tai ylempi keskiluokka	5,33
Alempi keskiluokka	4,19
Ylempi työväenluokka	3,96
Alempi työväenluokka	3,26
F-testi	122,6
p-arvo	p<0,001
eta^2	0,08

Taulukko 1. Eri yhteiskuntaluokkiin itsensä sijoittaneiden vastaajien suhtautuminen tuloeroihin. Varianssianalyysin tulokset.

Taulukon alaosan F-testiluku ja siihen liittyvä p-arvo kuvaavat ryhmien välisten erojen tilastollista merkitsevyyttä. Koska p-arvo on selvästi pienempi kuin yleisesti raja-arvona pidetty 0,05, voidaan nollahypoteesi ryhmäkeskiarvojen samansuuruisuudesta hylätä. Toisin sanoen eri yhteiskuntaluokkiin subjektiivisesti kuuluvien välillä on eroja suhtautumisessa tuloeroihin. Korkeimpiin yhteiskuntaluokkiin itsensä sijoittavat suomalaiset ovat valmiimpia hyväksymään suuret tuloerot ja yritteliäisyyden palkitsemisen kuin alempiin yhteiskuntaluokkiin kuuluvat.

Taulukon 1 alalaidassa esitetty ns. etan neliö kuvaa sitä, kuinka paljon selitettävän muuttuja vaihtelusta pystytään selittämään selittävän muuttujan avulla. Eta^2 on tunnuslukuna verrattavissa regressioanalyysin yhteydessä käytettävään R^2-lukuun. Se voi saada arvoja nollan ja yhden väliltä ja suuret arvot kuvastavat selittävän muuttujan parempaan selitysvoimaa. Taulukon 1 esimerkissä eta^2–luku saa arvon 0,08, joka on suhteellisen pieni luku. Luku voidaan tulkita niin, että yhteiskuntaluokkiin sijoittumista kuvaavan muuttujan avulla voidaan selittää 8% vastaajien suhtautumisen vaihtelusta tuloerojen kasvattamiseen tai niiden pienentämiseen. Selitysosuus ja muut tulokset ovat tietenkin sidoksissa aineistoon ja siinä käytettyihin operationalisointeihin.

Varianssianalyysin laajennukset

Kaksisuuntainen varianssianalyysi

Yksisuuntainen varianssianalyysi sisältää vain yhden selittävän muuttujan. Menetelmää voidaan kuitenkin laajentaa kattamaan myös useampia luokittelu- tai järjestysasteikon selittäviä

muuttujia. Kaksisuuntaisessa varianssianalyysissa (*two-way analysis of variance*) selittäviä muuttujia on kaksi. Tällöin voidaan tutkia sitä, vaikuttavatko molemmat selittävät muuttujat selitettävän muuttujan arvoihin yksittäin sekä onko niillä yhteisvaikutusta (eli interaktiovaikutusta).

Kaksisuuntaisessa varianssianalyysissa voisi esimerkkitutkimusongelmana olla, vaikuttaako sukupuoli ja koulutus keskimääräiseen palkkatasoon tutkimuksen kohdeyrityksessä. Tulokset kertovat, onko näillä kahdella selittävällä muuttujalla tilastollisesti merkitsevää vaikutusta palkkatasoon sekä sen, onko sukupuolella ja koulutuksella yhteisvaikutusta. Tässä esimerkissä yhteisvaikutus voi tarkoittaa esimerkiksi sitä, että yliopistotutkinnon suorittaneiden naisten keskimääräinen palkkataso on selvästi huonompi kuin saman koulutustason miesten keskipalkka, mutta muissa koulutusluokissa tällaista sukupuolten välistä eroa ei ole.

Periaatteessa varianssianalyysissa voidaan käyttää useampaakin kuin kahta selittävää muuttujaa. Silloin mahdollisten yhteisvaikutusten määrä kuitenkin kasvaa suureksi, mikä tekee tulkinnan monimutkaisemmaksi.

Kovarianssianalyysi

Samoin kuin kaksiulotteisessa varianssianalyysissa myös kovarianssianalyysissa (*covariance analysis*) lisätään varianssianalyysiin yksi tai useampia selittäviä muuttujia. Erona on kuitenkin se, että kovarianssianalyysissa lisättävä muuttuja on mittaustasoltaan välimatka- tai suhdeasteikollinen. Varianssianalyysin yhteydessä tällaista muuttujaa kutsutaan kovariaatiksi.

Oletetaan edellisen esimerkin tapaan, että tutkija on kiinnostunut sukupuolten välisistä palkkaeroista tutkimuksen kohteena olevassa yrityksessä. Hän kuitenkin epäilee, että sukupuolen lisäksi työntekijöiden ikäerot voivat vaikuttaa keskimääräiseen palkkatasoon. Ikämuuttujan vaikutus voidaan ottaa varianssianalyysissa huomioon lisäämällä se kovariaattina analyysiin. Saadut tulokset osoittavat, vaikuttaako sukupuoli tilastollisesti merkitsevästi keskimääräiseen palkkatasoon silloin, kun miesten ja naisten keski-iän erot on otettu huomioon.

Kovarianssianalyysi lähenee menetelmänä regressioanalyysia, jossa luokittelumuuttujat voidaan sisällyttää analyysiin ns. dummy-muuttujien avulla. Erona on, että kovarianssianalyysissa (ja varianssianalyysissa yleensäkin) otetaan automaattisesti huomioon selittävien muuttujien interaktiovaikutukset, kun taas regressioanalyysissa tutkija voi erikseen lisätä analyysiin ns. interaktiomuuttujat, jotta muuttujien mahdollinen yhteisvaikutus tulisi esille.

Monen muuttujan varianssianalyysi

Monen muuttujan varianssianalyysi eli MANOVA (*multivariate analysis of variance*) eroaa edellisistä varianssianalyysin laajennuksista siinä, että MANOVAssa on useita selitettäviä muuttujia. MANOVAA voidaan käyttää tilanteissa, joissa selitettävät muuttujat ovat teoreettisesti ja empiirisesti toisiinsa sidoksissa. Esimerkiksi työilmapiiritutkimuksissa voitaisiin kyselyn avulla muodostaa useita toisiinsa liittyviä työpaikan ilmapiiriä kuvaavia summamuuttujia ja tutkia eroja näissä muuttujissa yhtäaikaisesti.

Yleisesti voidaan todeta, että MANOVA on melko monimutkainen menetelmä ja ehkä siksi sen sovellukset yhteiskuntatieteissä ovat jääneet verraten harvinaisiksi.

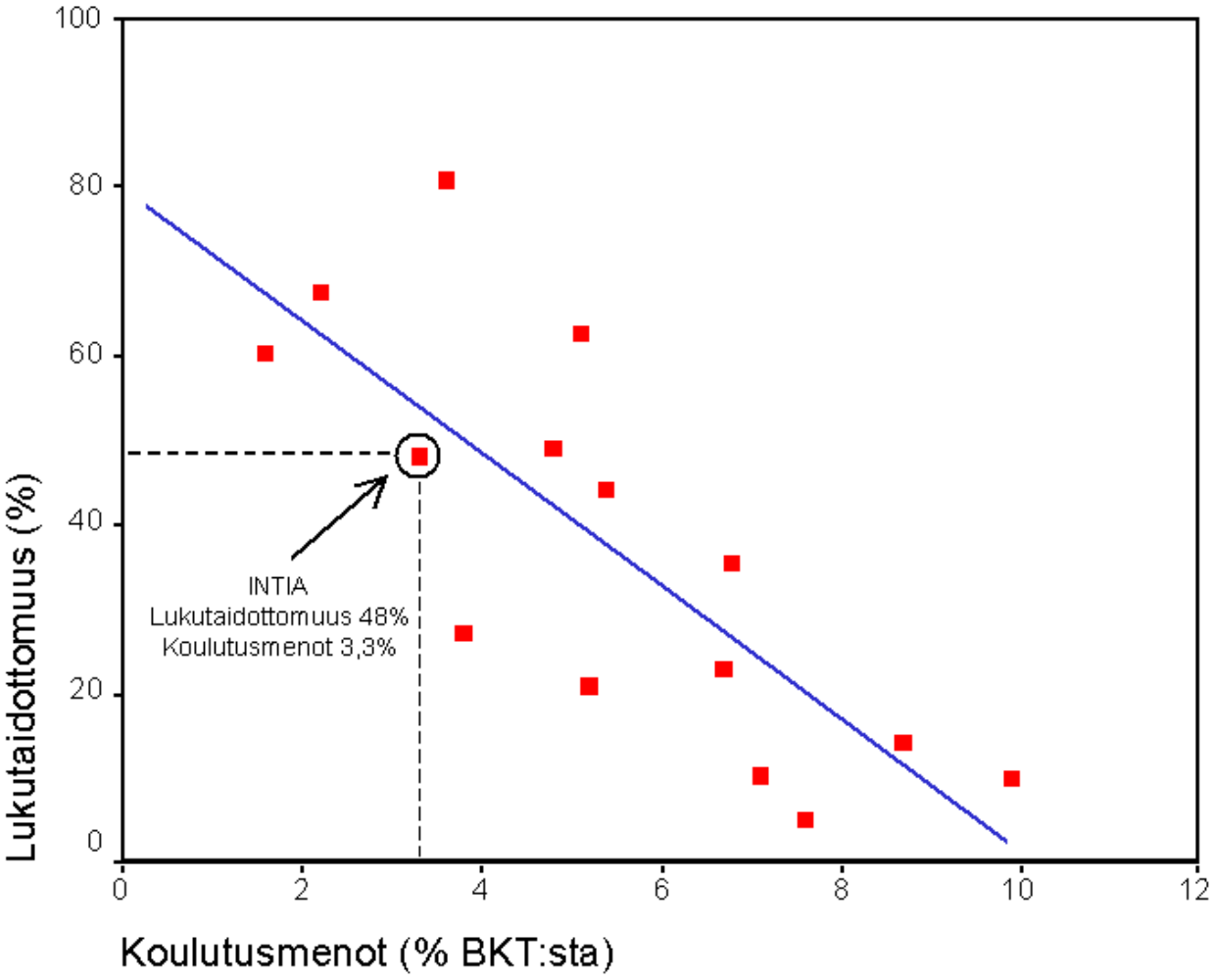
Regressioanalyysi

Regressioanalyysin (*regression analysis*) avulla tutkitaan yhden tai useamman selittävän muuttujan vaikutusta selitettävään muuttujaan. Sen avulla voidaan pyrkiä vastaamaan esimerkiksi siihen vaikuttaako koulutuksen pituus saadun palkan suuruuteen ja jos vaikuttaa, niin kuinka voimakas tämä vaikutus on. Regressioanalyysin erityinen etu on, että siinä voidaan tutkia yhtä aikaa monen selittävän muuttujan vaikutusta selitettävään muuttujaan. Tällöin tuloksen kertovat, mikä on yksittäisen selittävän muuttujan osuus silloin kuin muiden vaikuttavien tekijöiden vaikutus selitettävään muuttujaan on otettu huomioon.

Regressioanalyysi on monipuolinen ja joustava menetelmä muuttujien välisten kausaalisuhteiden tutkimukseen. Sen edellytyksenä on, että selitettävä muuttuja on vähintään välimatka-asteikollinen (katso muuttujien mittaustaso). Selittävät muuttujat ovat yleensä myös vähintään välimatka-asteikollisia, mutta myös luokittelu- ja järjestysasteikollisia muuttujia voidaan sisällyttää analyysiin. Tällöin niistä täytyy tehdä ns. dummy-muuttujia.

Regressiosuora ja -kerroin

Regressioanalyysin perusperiaatteet voidaan esittää havainnollisesti kuvion 1 avulla. Hajontakuviossa on esitetty 15 valtion lukutaidottomuusprosentti ja valtion panostus koulutukseen prosenttiosuutena bruttokansantuotteesta. Jokainen kuvion piste viittaa yhteen maahan. Esimerkiksi Intiassa oli vuonna 1999 lukutaidottomia noin 48% väestöstä ja maan bruttokansantuotteesta käytettiin 3,3% koulutusmenoihin. Kannattaa huomata, että kuviossa esitetyt maat ja luvut ovat oikeita, mutta niiden valinta perustui tarkoituksenmukaisuusharkintaan. Näin esitetyt empiiriset tulokset ovat yleistettävyyden kannalta parhaimmassakin tapauksessa vain suuntaa-antavia.



Kuvio 1. Lukutaidottomuusprosentti (1991) ja koulutusmenot (% BKT:sta, 1995). Lähde: Tilastokeskus, Maailma numeroina.

Kuviosta näkee selvästi, miten lukutaidottomuus ja panostus koulutukseen ovat yhteydessä toisiinsa. Mitä suurempi osuus maan bruttokansatuotteesta sijoitetaan koulutukseen, sitä vähemmän maassa on lukutaidottomia. Regressioanalyysin avulla voidaan tutkia, onko näiden kahden muuttujan välinen yhteys tilastollisesti merkitsevä. Lisäksi regressioanalyysi kertoo, kuinka vahva yhteys on, eli kuinka paljon lukutaidottomuus vähenee, kun koulutusmenojen osuus kasvaa.

Kuvioon piirretty viiva on ns. **regressiosuora** (*regression line*). Se osoittaa muuttujien välisen yhteyden voimakkuuden. Jos regressiosuora laskee alaspäin, on muuttujilla negatiivinen yhteys ja jos se nousee ylöspäin, on niillä positiivinen yhteys. Mitä lähempänä vaakatasoa suora on, sitä vähemmän muuttujilla on yhteyttä toisiinsa.

Regressiosuora voidaan merkitä kaavan avulla seuraavasti: $Y = a + bX$

Kaavassa Y tarkoittaa selitettävän muuttujan arvoa, a on ns. vakiotekijä, X on selittävän muuttujan arvo ja b on **regressiokerroin** (*regression coefficient*). Regressiokerroin on regressiosuoran kulmakerroin. Jos se saa negatiivisen arvon, on suora laskeva ja jos regressiokerroin on positiivinen, on suora nouseva. Jos regressiokerroin on nolla, ei muuttujien välillä ole lineaarista eli suoraviivaista yhteyttä. Vakiotekijä kertoo, minkä arvon selitettävä muuttuja saa silloin, kun selitettävän muuttujan X arvo on nolla. Se siis kertoo, missä kohtaa regressiosuora leikkaa kuvion y-akselin.

Regressioanalyysin avulla voidaan selvittää kaavan vakiotekijän ja regressiokertoimen arvot. Esimerkiksi kuvion 1 aineiston perusteella saadaan regressioyhtälö: $Y = 80 - 7{,}9X$

Yhtälön regressiokerroin (eli b:n arvo) on -7,9. Regressiokerroin kertoo, kuinka paljon selitettävä muuttuja muuttuu, kun selittävä muuttuja kasvaa yhden yksikön. Esitetty yhtälö voidaan tulkita seuraavasti. Kun koulutusmenoja lisätään yhdellä prosenttiyksiköllä bruttokansantuotteesta, vähenee lukutaidottomien määrä 7,9 prosenttiyksikköä. Vakiotekijä kertoo, kuinka paljon maassa olisi lukutaidottomia, jos koulutusmenot olisivat nolla eli maassa ei panostettaisi laisinkaan rahaa koulutukseen. Tällöin lukutaidottomia olisi maassa 80%. Tämä on tietenkin vain hypoteettinen arvio, koska maailmasta tuskin löytyy sellaista maata, missä koulutukseen ei panostettaisi ollenkaan.

Regressiomallin eli -yhtälön pätevyyttä voidaan arvioida sen mukaan, kuinka lähelle kuvion pisteet sijoittuvat regressiosuoraa. Mitä lähempänä suoraa ne sijaitsevat, sitä parempi on regressiomallin selitysvoima ja päinvastoin. Jos kuvion pisteen sijoittuvat hyvin lähelle suoraa, on mallilla hyvä ennustevoima, koska sen avulla voidaan hyvin tarkasti arvioida, mikä on jonkin yksittäisen maan lukutaidottomuusprosentti silloin, kun tiedetään kuinka paljon maassa sijoitetaan koulutukseen. Mitä kauempana pisteet suorasta sijaitsevat, sitä epävarmempia ovat ennusteet.

Yksittäisen havainnon arvon etäisyyttä regressiosuorasta kutsutaan havainnon **virhetermiksi** tai **residuaaliksi** (*residual*). Esimerkiksi kuviosta 1 tiedämme, että Intiassa lukutaidottomuuden taso on 48%. Regressioyhtälön avulla voidaan myös laskea regressiomallin ennusteen Intian lukutaidottomuudelle. Se saadaan sijoittamalla regressiokaavaan selitettävän muuttujan eli koulutukseen menevien varojen bruttokansantuoteosuus, joka on Intian kohdalla 3,3. Näin saadaan regressiomallin ennusteeksi Intian osalta 53,9 (=80-7,9*3,3). Tämä osoittaa, että regressiomalli ei ole aivan tarkka yksittäisten havaintojen kohdalla. Intian virhetermi mallissa on 48-53,9=-5,9. Mitä suuremmat mallin virhetermit itseisarvoltaan ovat, sitä huonompi ennustearvo regressiomallilla on ja päinvastoin.

Regressioanalyysin tulosten tulkinta

Seuraavaksi käytetään Tilastokeskuksen keräämää Maailma numeroina -aineistoa regressioanalyysin tulosten esittelemiseksi (katso aineiston kuvaus verkkoversiosta). Selitettävänä muuttujana on maakohtainen elinajan odote eli väestön keskimääräisen odotettavissa olevan eliniän pituus. Elinajan odotteeseen vaikuttaa tietenkin useat eri tekijät, mutta esimerkkiregressioanalyysissa käytetään keskeisenä selittävänä tekijänä HIV-taudin levinneisyyttä. HI-virus ja siitä seuraava AIDS-tauti on 1990-luvulla kääntänyt monessa maassa aikaisemmin kasvussa olleet elinajan odotteet laskuun. Suurimmillaan tämä vaikutus näkyy Afrikassa. Arvioiden mukaan esimerkiksi Zimbabwessa odotettavissa oleva elinikä on laskenut AIDSin vaikutuksesta jopa 26 vuotta (U.S. Bureau of Census 1998). AIDS vaikuttaa elinajan odotteeseen kahdella eri tavalla. Ilman kallista lääkitystä sairaus tappaa aikuiset potilaat nopeasti. Lisäksi sairaus kasvattaa lapsikuolleisuutta, koska taudin voi saada myös HI-virusta kantavalta äidiltä. Näiden kahden tekijän kautta AIDSilla on suuri vaikutus odotettavissa olevaan elinikään.

Aineistossa on 165 maata, joista on saatavilla tiedot sekä elinajan odotteesta että HIV-potilaiden määrästä. Vuonna 1999 eliniän odote vaihteli 36.3 (Malawi) ja 83.5 (Andorra) vuoden välillä. HIV-tapausten yleisyyttä mitataan suhteuttamalla ne väestön kokoon niin, että muuttuja mittaa HIV-tapausten yleisyyttä suhteessa 1000 henkilöön. Tämä muuttuja vaihtelee lähes nollan (esimerkiksi Suomessa 0,21) ja 182 (Botswana) välillä.

Taulukossa 1 on esitetty regressioanalyysin tulokset. Taulukon yläosassa ovat analyysin selittävät muuttujat, niiden regressiokertoimet, t-arvot ja merkitsevyystiedot. Taulukon alaosa sisältää regressiomallin pätevyyden arviointiin sopivia tunnuslukuja.

	Regressiokerroin	t-arvo	Merkitsevyys
Vakio	68,4**	91,5	p<0,001
HIV tapaukset (/1000 henkilöä)	-0,27**	-11,3	p<0,001
R^2	0,44		
Korjattu R^2	0,44		
F-testi	128,0**		p<0,001
Estimaatin keskivirhe	8,7		

Taulukko 1. Regressioanalyysi HIV:n yleisyyden vaikutuksesta elinajan odotteeseen (**p<0,01, n=165).

Ennen regressiokertoimien varsinaista tulkintaa kannattaa kiinnittää huomiota niiden tilastolliseen merkitsevyyteen. Regressioanalyysin yhteydessä testataan jokaisen selittävän muuttujan osalta onko niillä vaikutusta selitettävään muuttujaan eli eroavatko ne tilastollisesti merkitsevästi nollasta (katso tilastollinen päättely ja hypoteesien testaus). Tällaiseen tarkoitukseen sopiva testimenetelmä on ns. t-testi. Testin tuloksena jokaiselle selittävälle muuttujalle saadaan t-arvo, jonka suuruus ratkaisee sen, voidaanko muuttujan kerrointa pitää nollaa suurempana tilastollisten kriteerien mukaan. Taulukon viimeisessä sarakkeessa on esitetty t-testien merkitsevyystasot. Ne osoittavat, että sekä vakiotermi että HIV-tapausten laajuuden regressiokerroin eroavat tilastollisesti selvästi nollasta. Kaikki regressioanalyysiin sopivat ohjelmat tuottavat nämä tunnusluvut automaattisesti.

Taulukon 1 tulokset siis osoittavat, että HIV-tapausten levinneisyys laskee odotettavissa olevaa elinikää (regressiokertoimen etumerkki on negatiivinen). Kerroin on arvoltaan -0,27, mikä tarkoittaa sitä, että HIV-tapausten suhteellisen osuuden kasvu yhdestä hengestä kahteen henkeen tuhannesta laskee elinajan odotetta 0,27 vuotta. Tämä on suuri muutos. Jos Suomessa (0,21 tapausta / 1000 henkilöä) HIV olisi yhtä yleinen kuin Ranskassa (2,21/1000 henkilöä), suomalaisten keskimääräinen elinajan odote olisi noin puoli vuotta matalampi ((2,21-0,21)*0,27=0,54). Jos HIV-tapauksia olisi suhteellisesti yhtä paljon kuin Tansaniassa (39,6/1000 henkilöä), suomalaisten elinajan odote olisi peräti 11 vuotta lyhyempi ((39,6-0,21)*0,27=10,6).

Taulukon 1 alalaidassa on esitetty tärkeimmät regressioanalyysin selitysvoimaa kuvaavat testit. Tällaisia testejä on useita, mutta R^2-luku ja F-testi ovat yleisemmin käytetyt.

R^2-luku on regressiomallin selitysosuus. Se kertoo kuinka suuren osuuden selitettävän muuttujan vaihtelusta regressionanalyysin selittävät muuttujat pystyvät selittämään. R^2-luku vaihtelee nollan ja yhden välillä. Se saadaan laskemalla selitettävän muuttujan arvojen ja mallin tuottamien ennustearvojen korrelaation neliö. Jos R^2-luku on pieni regression selittävät muuttujan pystyvät selittämään vain vähän selitettävän muuttujan vaihtelusta ja päinvastoin. Taulukossa 1 R^2-luku on 0.44. Tämä tarkoittaa, että HIV-tapausten levinneisyydellä pystytään siis kohtuullisen hyvin selittämään elinajan odotteen vaihtelua. Regressiomallin avulla 44% elinajan odotteen vaihtelusta voidaan selittää pelkästään HIV-tapausten suhteellisella määrällä. On kuitenkin huomattava, että selitysosuutta kuvaavat luvut ovat merkityksellisiä nimenomaan regressiomallin asettamassa kontekstissa. Jos elinajan odotetta selitettäisiin lisäksi muilla siihen vaikuttavilla tekijöillä, HIVin levinneisyyden selitysosuus olisi luultavasti pienempi.

Korjattua R^2-lukua (*adjusted R^2*) käytetään silloin, kun halutaan verrata kahden regressioanalyysin tuloksia keskenään. Korjattu R^2-luku ottaa huomioon mallin sisältämien selittävien muuttujien lukumäärän. Se on arvoltaan aina pienempi tai yhtä suuri kuin varsinainen R^2-luku. Korjaus R^2-lukuun tarvitaan sen vuoksi, että uusien selittävien muuttujien lisääminen regressioanalyysiin nostaa aina R^2-lukua, vaikka nämä lisätyt muuttujat eivät todellisuudessa pystyisikään lisäämään selityskykyä. Silloin kun tarkasteltavana on vain yksi regressiomalli, ei

korjatun R^2-luvun käyttäminen ole tarpeellista, mutta regressiomalleja verratessa siitä on hyötyä. Jatkossa taulukon 1 regressioanalyysia laajennetaan uusilla muuttujilla. Siksi korjattu R^2-luku on raportoitu myös tässä yhteydessä, jotta vertaileminen myöhemmin esitettyihin laajennettuihin regressiomalleihin on mahdollista.

F-testi on tilastollinen testi, joka kertoo pystytäänkö regressioanalyysissa olevilla muuttujilla ylipäänsä selittämään selitettävän muuttujan vaihtelua. Koska se on tilastollinen testi, saadaan sille myös merkitsevyystaso. Taulukossa 1 F-testin tulos on erittäin merkitsevä. Tämä ei sinänsä ole yllätys, koska myös selittävän muuttujan regressiokerroin on tilastollisesti merkitsevä. On kuitenkin mahdollista, että yhdenkään selittävän muuttujan regressiokerroin ei ole tilastollisesti merkitsevä, mutta F-testin tulos on. Tämä tarkoittaa sitä, että regressioanalyysin muuttuja pystyvät yhdessä selittämään selitettävän muuttujan vaihtelua, vaikka yksittäin katsoen ne eivät ole tilastollisesti merkitseviä. Tällaiset tapaukset ovat kuitenkin harvinaisia.

Viimeinen regressiomallin onnistuneisuutta kuvaava tunnusluku on **estimaatin keskivirhe** (*standard error of estimate*). Tämä luku ilmoittaa regressiomallin virhetermien keskihajonnan (katso hajontaluvut). Mitä suurempi se on, sitä suurempi on virhetermien hajonta ja samalla sitä pienempi mallin selitysvoima. Estimaatin keskivirheen suuruus riippuu aina regressiomallin hyvyyden lisäksi selitettävän muuttujan mittaluokasta. Taulukossa 1 se on 8,7, mikä on kohtalaisen suuri luku, kun se suhteutetaan elinajan odotteen vaihteluväliin (36-84 vuotta). Tämä osoittaa, että HIV-tapausten yleisyydestä tietyssä maassa ei pystytä kovinkaan tarkasti ennustamaan maan väestön odotettavissa olevaa keskimääräistä elinikää.

Usean muuttujan regressioanalyysi

Edellisissä regressioanalyysin esimerkeissä oli vain yksi selittävä muuttuja. Regressioanalyysin etu on kuitenkin se, että siihen voi sisällyttää useita selittäviä muuttujia yhtäaikaisesti. Tällöin muuttujien regressiokertoimet kertovat, kuinka paljon selitettävän muuttujan arvo muuttuu, kun selittävän muuttujan arvo muuttuu yhdellä yksiköllä ja kaikkien muiden muuttujien arvo pysyy samana. Toisin sanoen usean muuttujan regressioanalyysissa regressiokertoimet ilmoittavat selittävän muuttujan vaikutuksen selitettävään muuttujaan niin, että muiden mallin muuttujien vaikutus on vakioitu.

Kahden selittävän muuttujan regressioanalyysin kaava voidaan esittää seuraavasti:
$Y = a + b_1X_1 + b_2X_2$

Kaavassa Y on selitettävän muuttujan arvo, a vakiotekijä, X_1 ja X_2 selittävät muuttujat sekä b_1 ja b_2 niiden regressiokertoimet.

Usean muuttujan regressioanalyysin kuvaamiseen voidaan käyttää edellistä esimerkkiä HIV-taudin yleisyyden ja elinajan odotteen yhteydestä. HIV ei ole ainoa tekijä, joka vaikuttaa keskimääräiseen odotettavissa olevaan elinikään. Yksi tällainen tekijä on maan yleinen taloudellinen kehitystaso, joka vaikuttaa muun muassa siihen, kuinka paljon lääkäreitä ja sairaaloita maassa on, kuinka paljon on mahdollista käyttää kalliita lääkkeitä jne. Usein taloudellista kehitystasoa mitataan suhteuttamalla maan bruttokansantuote väkilukuun. Seuraavaksi tämä muuttuja lisätään HIV-taudin yleisyyden lisäksi regressioanalyysiin. BKT-muuttuja mittaa henkeä kohden laskettua bruttokansatuotetta vuonna 1997 tuhansina dollareina (eli 1000 US$/henkilöä). Muuttuja vaihtelee välillä 0,09 (Kongon demokraattinen tasavalta) ja 40,6 (Brunei).

Taulukossa 2 on esitetty tämän regressioanalyysin tulokset. Uuden muuttujan lisääminen analyysiin ei muuttanut paljoakaan HIV-muuttujan kerrointa. Tämä tarkoittaa sitä, että HIV-taudin yleisyydellä on selvä vaikutus elinajan odotteeseen, vaikka maan taloudellinen kehitystaso otetaankin analyysissa huomioon. BKT-muuttujan regressiokerroin on myös tilastollisesti merkitsevä ja sen arvo on 0,57. Kertoimen tulkinta kertoo, että maan henkeä kohden lasketun bruttokansantuotteen kasvaessa 1000 yhdysvaltain dollarilla elinajan odote kasvaa noin puolella vuodella, jos maan HIV-tilanne pysyy samana.

	Regressiokerroin	t-arvo	Merkitsevyys
Vakio	64,4**	87,0	p<0,001
HIV tapaukset (/1000 henkilöä)	-0,23**	-11,6	p<0,001
BKT /henkilö	0,57**	9,44	p<0,001
R^2	0,64		
Korjattu R^2	0,63		
F-testi	143,2**		p<0,001
Estimaatin keskivirhe	7,04		

Taulukko 2. Regressioanalyysi HIV:n yleisyyden vaikutuksesta elinajan odotteeseen (**p<0,01, n=165).

Taulukon 2 korjattu R^2-luku luku osoittaa, että BKT-muuttujan lisääminen regressiomalliin paransi mallin selityskykyä huomattavasti verrattuna taulukon 1 tuloksiin. Taulukossa 1 korjattu R^2-luku on 0,44 ja taulukossa 2 vastaava tunnusluku on 0,63. Lisäksi estimaatin keskivirhe pieneni 8,7:stä 7,0:an. Nämä molemmat tunnusluvut kertovat, että käyttämällä BKT-muuttujaan HIV-muuttujan ohella analyysissa, pystytään eri maiden odotettavissa olevaa elinikää ennustamaan paremmin kuin tyytymällä ainoastaan HIV-muuttujan käyttöön.

Dummy-muuttujat

Dummy-muuttujaksi kutsutaan sellaista muuttujaa, joka voi saada vain kaksi eri arvoa, jotka on koodattu nollaksi ja yhdeksi. Tyyppiesimerkki tällaisesta muuttujasta on vastaajan sukupuoli, mutta vaihtoehtoja on helppo keksiä lisää (onko vastaaja opiskelija vai ei, onko maa liittovaltio vai ei jne.). Dummy-muuttujien avulla regressioanalyysiin voidaan helposti sisällyttää luokittelu- tai järjestysasteikollisia muuttujia.

Oletetaan, että afrikkalaisissa maissa elinajan odote on jostakin syystä alhaisempi kuin muissa maissa. Tätä hypoteesia voi tutkia lisäämällä regressioanalyysiin dummy-muuttujan, joka saa arvon yksi silloin kun maa sijaitsee Afrikassa ja muutoin arvoksi tulee nolla. Kaavan avulla tämä voidaan esittää seuraavasti: $Y = a + b_1X_2 + b_2X_2 + b_3X_3$

Kaavassa X_3 on uusi dummy-muuttuja, joka saa arvon yksi silloin kun kyseessä on afrikkalainen maa. Muut muuttujat ovat samat kuin edellisessä esimerkissä.

Dummy-muuttujien regressiokertoimien tulkinta on erittäin yksinkertaista. Kerroin ilmoittaa, kuinka muuttujalla arvon yksi saava havaintoryhmä eroaa niistä havainnoista, jotka saavan arvon nolla. Jos kerroin on positiivinen, se ilmaisee kuinka paljon suurempi elinajan odote on Afrikassa kuin Afrikan ulkopuolisissa maissa. Jos se on negatiivinen, kertoo se kuinka paljon lyhyempi elinikä Afrikassa on.

Taulukko 3 sisältää tulokset regressioanalyysista, jossa Afrikkaa koskeva dummy-muuttuja on mukana. Se saa arvon -11, mikä tarkoittaa sitä, että Afrikan maissa elinajan odote on noin 11 vuotta lyhyempi kuin muissa maissa, vaikka HIVin levinneisyys ja maan taloudellisen kehityksen tila on otettu huomioon. Lisäksi kannattaa huomioida, että HIV-muuttujan kerroin pieneni huomattavasti dummy-muuttujan lisäyksen jälkeen. Tässä tapauksessa dummy-muuttujan käyttö ei itse asiassa selitä miksi elinikä on Afrikassa lyhyempi kuin muualla, vaan se ainoastaan tuo esillä tämän empiirisen yhdenmukaisuuden. Analyysin seuraavana askeleena tulisikin pohtia, mitkä mahdolliset elinikään vaikuttavat tekijät ovat yleisempiä Afrikassa kuin muualla maailmassa. Tämän teoreettistakin pohdintaa vaativan arvioinnin jälkeen analyysiin voitaisiin ehkä lisätä uusia muuttujia tulosten parantamiseksi.

	Regressiokerroin	t-arvo	Merkitsevyys
Vakio	67,3**	98,8	p<0,001
HIV tapaukset (/1000 henkilöä)	-0,14**	-7,1	p<0,001
BKT /henkilö	0,44**	8,4	p<0,001
Afrikkaa kuvaava dummy-muuttuja	-11,02**	-8,76	p<0,001
R^2	0,76		
Korjattu R^2	0,75		
F-testi	165,7**		p<0,001
Estimaatin keskivirhe	5,81		

Taulukko 3. Regressioanalyysi HIV:n yleisyyden vaikutuksesta elinajan odotteeseen.

Dummy-muuttujia voidaan käyttää myös tilanteessa, jossa laatu- tai järjestysasteikon muuttuja saa useampia kuin kaksi vaihtoehtoa. Tällaisessa tilanteessa yleinen periaate on, että uusia dummy-muuttujia täytyy luoda yksi vähemmän kuin laatu- tai järjestysasteikon muuttujassa on vastausvaihtoehtoja. Jos esimerkiksi laatueroasteikon muuttuja voi saada neljä eri arvoa, täytyy regressioanalyysia varten luoda kolme uutta dummy-muuttujaa.

Oletetaan, että tutkija haluaa regressioanalyysin avulla selvittää henkilöiden iän ja koulutuksen vaikutusta heidän palkkatasoonsa. Koulutus on mitattu kolmiasteisella mittarilla, jonka vaihtoehdot ovat peruskoulu, keskiasteen tutkinto ja korkeakoulututkinto. Regressioanalyysin tarpeisiin tästä muuttujasta täytyy luoda kaksi uutta dummy-muuttujaa. Ensimmäinen muuttuja voisi olla peruskoulu-dummy, joka saa arvon yksi jos vastaaja on suorittanut vain peruskoulun. Muutoin muuttuja saa arvon nolla. Toinen muuttuja olisi keskiaste-dummy, joka saa arvon yksi silloin kun vastaajalla on keskiasteen tutkinto ja arvon nolla muutoin. Tutkija laskee regressioanalyysin, jossa selitettävänä muuttujana on vastaajan palkan suuruus markkoina ja selittävinä muuttujina vastaajan ikä sekä kaksi edellä mainittua dummy-muuttujaa.

Useamman dummy-muuttujan tapauksessa niiden regressiokertoimien tulkinta tulee hiukan hankalammaksi, koska ne täytyy tulkita toisiinsa suhteuttaen. Oletetaan, että regressioanalyysin tuloksissa peruskoulu-dummyn regressiokerroin on -5000 ja keskiaste-dummyn -2000. Nämä kertoimet tulee tulkita suhteessa korkeakoulututkinnon suorittaneiden palkkaan. Ne kertovat, että ainoastaan peruskoulun suorittaneiden palkka on keskimäärin 5000 mk pienempi kuin korkeakoulun suorittaneiden palkat. Keskiasteen tutkinnon suorittaneiden keskimääräinen palkka on 2000 mk pienempi kuin korkeakoulututkinnon suorittaneiden. Dummy-muuttujien regressiokertoimet ilmoittavat siis ryhmän keskimääräisen poikkeaman siitä ryhmästä, jolle ei tehty omaa dummy-muuttujaa.

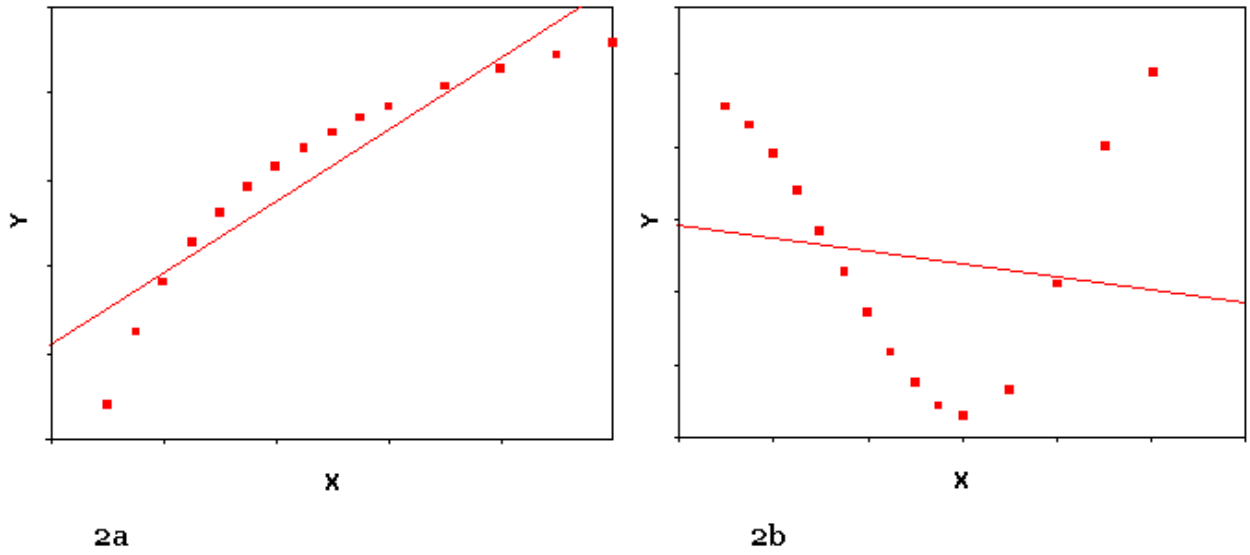
Päätökset siitä, mille vastausvaihtoehdoille omat dummy-muuttujat luodaan ja mikä vaihtoehto jätetään analyysista pois eivät ole kovin ratkaisevia. Ne toki vaikuttavat dummy-muuttujien regressiokertoimien arvoihin, mutta niistä tehtävät tulkinnat ovat kuitenkin samoja. Jos edelliseen regressiomalliin olisikin lisätty keskiaste- ja korkeakoulu-dummyt, olisivat niiden regressiokertoimet olleet +3000 ja +5000 mk. Ne siis kertovat, että korkeakoulun käyneiden ja peruskoulun käyneiden keskimääräinen ero palkoissa on 5000 mk sekä korkeakoulun käyneiden ja keskiasteenkoulutuksen saaneiden 2000 mk.

Regressioanalyysin rajoitteet

Regressioanalyysi on joustavuudessaan erinomainen menetelmä muuttujien riippuvuussuhteiden tarkasteluun. Siihen liittyy kuitenkin rajoitteita, joista menetelmän käyttäjän

on hyvä olla tietoinen. Tässä yhteydessä rajoitteet esitellään vain lyhyesti. Regressioanalyysi tarjoaa myös monia mahdollisia tapoja ottaa rajoitteet huomioon ja "korjata" niiden vaikutukset regressioanalyysissa. Lisätiedot osuudessa listataan useita kirjoja, joista saa tarkempia tietoja näistä mahdollisuuksista.

a) Lineaarisuusoletus. Regressioanalyysin avulla voidaan tutkia muuttujien välisiä lineaarisia eli suoraviivaisia kausaalisuhteita. Jos regressioanalyysin tulokset osoittavat, että selittävällä muuttujalla ei ole tilastollisesti merkitsevää yhteyttä selitettävään muuttujaan, tarkoittaa tämä tarkasti ottaen ainoastaan sitä, ettei lineaarista yhteyttä esiinny. Muuttujilla voi kuitenkin olla epälineaarinen yhteys. Kuviossa 2 on esitetty kaksi tilannetta, joissa x- ja y-muuttujien välillä on epälineaarinen yhteys.



Kuvio 2. Esimerkkejä muuttujien epälineaarista yhteyksistä.

Kuvion 2 kummassakin esimerkissä pisteet tarkoittavat muuttujien havaittuja arvoja ja suora on niiden pohjalta piirretty regressiosuora. Kuvion 2a tilanteessa x- ja y-muuttujien yhteys on epälineaarinen, mutta poikkeama lineaarisuudesta ei ole suuri. Tässä tilanteessa muuttujan x regressiokerroin olisi positiivinen ja se antaisi kohtuullisen hyvän likiarvon muuttujien välisestä suhteesta.

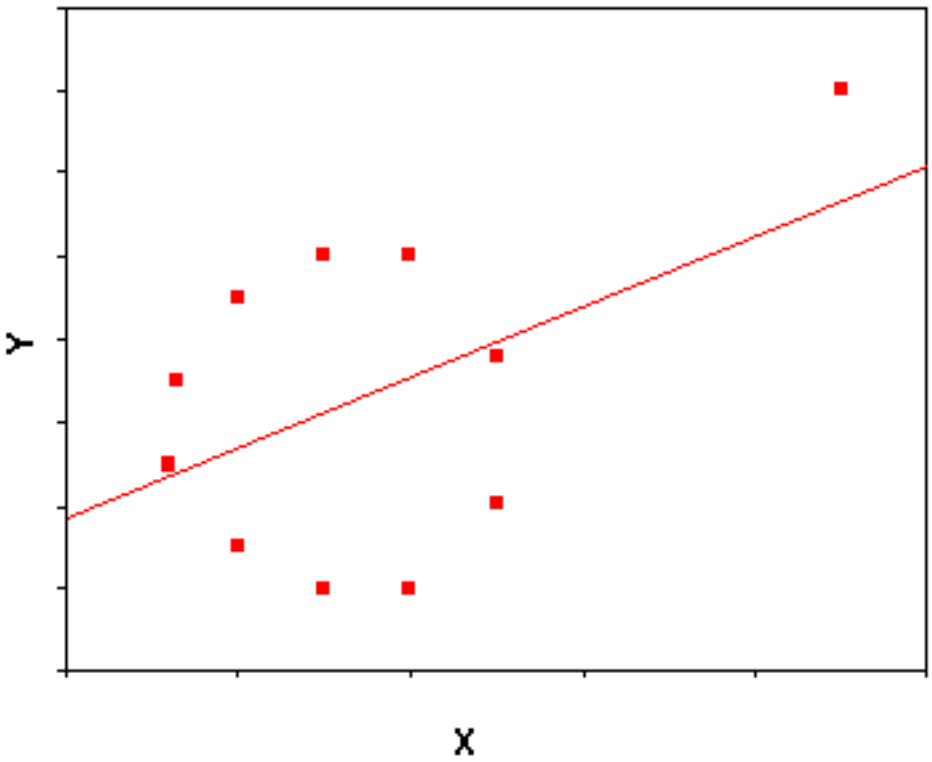
Esimerkki kuviossa 2b kuvaa tilannetta, jossa x- ja y-muuttujan suhde on erittäin epälineaarinen. Regressiosuora on lähes vaakasuora (eli regressiokerroin on lähellä nollaa), mikä ilmaisee sen, että muuttujilla ei ole lineaarista yhteyttä toisiinsa. Jos tutkija tällaisen analyysin pohjalta toteaa, että x-muuttujan avulla ei voida selittää y-muuttujan arvoja, tekee hän kuitenkin virheen, koska muuttujilla on selkeä epälineaarinen yhteys toisiinsa.

Regressioanalyysin avulla voi kuitenkin tarkastella myös muuttujien epälineaarisia suhteita. Tämä tapahtuu muuttujien muunnosten avulla. Muunnoksen kohteena voi olla sekä selitettävä tai selittävät muuttujat tilanteen mukaan. Lievien epälineaarisuuksien korjaamiseen käytetään logaritmi- tai neliöjuurimuunnosta. Jos kuvion esimerkissä 2a x-muuttujasta otetaan luonnollinen logaritmi ja tämä uusi muuttuja sisällytetään regressioanalyysiin alkuperäisen x-muuttujan sijasta, paranee mallin selitysosuus huomattavasti. Tämä johtuu siitä, että y-muuttujalla ja uudella selittävällä muuttujalla (x:n logaritmi) on lähes täydellinen lineaarinen riippuvuus toisistaan.

Esimerkissä 2b epälineaarisuus on niin vahva, että yksinkertaisilla muuttujamuunnoksilla siitä ei selvitä. Muuttujien välinen yhteys on kuitenkin sellainen, että se voidaan kuvata toisen asteen yhtälöllä. Käytännössä tämä tarkoittaa sitä, että regressioanalyysia varten luodaan uusi muuttuja, joka saa arvoksi X-muuttujan arvon neliön (eli X^2). Kun nämä molemmat muuttujat

lisätään regressioanalyysiin selittävinä muuttujina, voidaan esimerkin mukainen epälineaarinen yhteys analysoida regressioanalyysin avulla.

b) Poikkeavat havainnot eli outlier-tapaukset (*outliers*). Joskus yksittäisillä poikkeavilla havainnoilla voi olla suuri vaikutus regressioanalyysiin tuloksiin. Tällaisia havaintoja kutsutaan niiden englanninkielisen nimen mukaan outlier-tapauksiksi. Asia on havainnollistettu kuviossa 3. Kuvion oikeassa ylälaidassa oleva havainto on outlier-tapaus. Jos se poistetaan kuviosta, x- ja y-muuttujilla ei ole laisinkaan lineaarista riippuvuutta toisistaan.

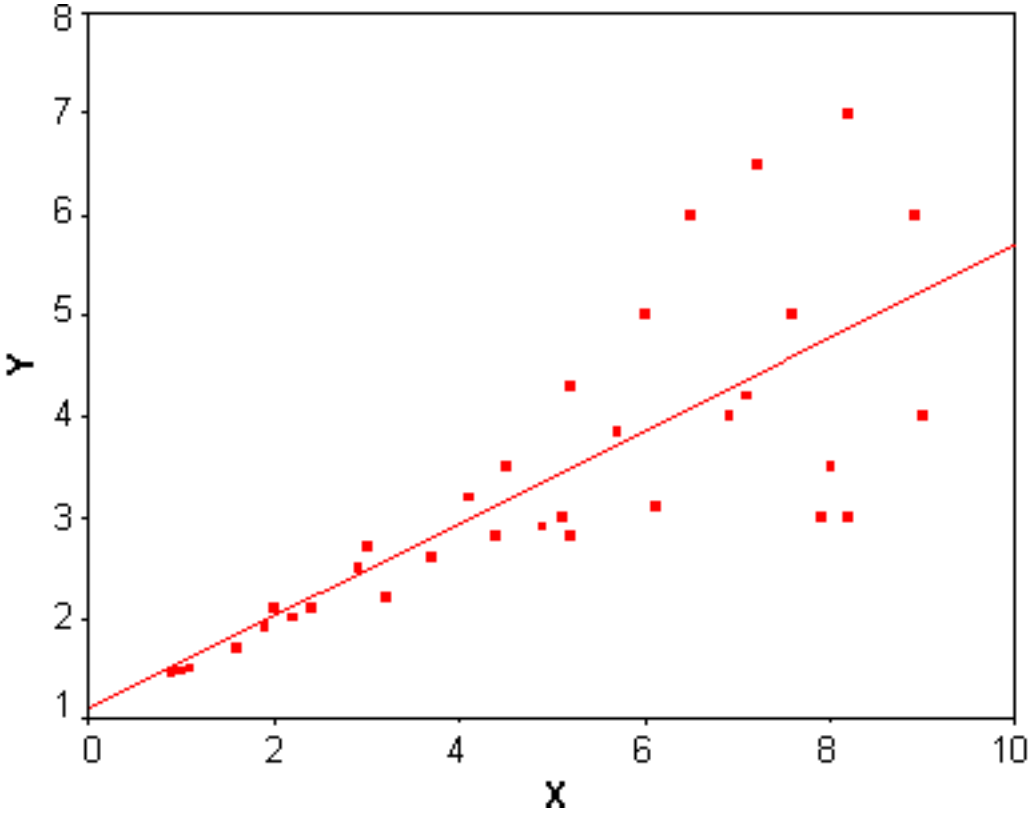


Kuvio 3. Esimerkki tilanteesta, jossa yksittäinen poikkeava havainto vääristää regressioanalyysin tuloksia.

Joskus poikkeavien havaintojen taustalla voi olla yksinkertaisesti koodausvirhe, joka voidaan helposti korjata. Useimmiten kyse on kuitenkin siitä, että jokin tai jotkut havainnot saavat todellisuudessa muista huomattavasti poikkeavia arvoja. Tällaisessa tilanteessa kannattaa pohtia, mikä tekijä aiheuttaa havainnon poikkeavuuden. Jos sille löytyy hyvä selitys joka voidaan mitata, voidaan tämä tekijä sisällyttää analyysiin uutena muuttujana, jolloin se ei enää vääristä analyysin tuloksia. Poikkeavien havaintojen löytämiseksi on kehitetty erilaisia tunnuslukuja (esimerkiksi Mahalanobisin ja Cookin etäisyysmittarit). Näistä luvuista ja niiden tulkinnasta löytyy tietoa lisätietoja-kohdassa suositelluista kirjoista (katso esimerkiksi Tabachnickin ja Fidellin kirja).

c) Multikollineaarisuus ja heteroskedastisuus. Regressioanalyysissa on aivan luonnollista, että selittävät muuttujat korreloivat keskenään. Joskus niiden keskinäinen korrelaatio voi kuitenkin olla niin suuri, että se aiheuttaa ongelmia regressioanalyysin tulosten tarkkuuden kannalta. Tällaista tilannetta kutsutaan multikollineaarisuudeksi. Yleensä multikollineaarisuusongelmia ei synny, jollei selittävien muuttujien välillä ole todella suuria riippuvuuksia (esimerkiksi korrelaatiokerroin yli 0,9). Ongelmana on, että kaikkia multikollineaarisuusongelmia ei voi havaita tarkastelemalla pelkästään selittävien muuttujien välisiä korrelaatiokertoimia. Tämän vuoksi on kehitetty erilaisia multikollineaarisuusmittareita, jotka ilmaisevat ongelman mahdollisen vakavuuden (esimerkiksi VIF-mittari).

Heteroskedastisuus viittaa tilanteeseen, jossa regressiomallin virhetermien hajonta vaihtelee suuresti ja systemaattisesti x-muuttujien arvojen muuttuessa. Kuviossa 4 havainnollistetaan heteroskedastisuutta. Kuvion y-akseli kuvaa selitettävän muuttujan arvoja ja x-akseli selittävän muuttujan arvoja. Kuvion esittämässä tilanteessa on kyse heteroskedastisuudesta siksi, että virhetermit vaihtelevat regressiosuoran ympärillä huomattavasti enemmän silloin kun x-muuttuja saa suuria arvoja.



Kuvio 4. Esimerkki heteroskedastisuudesta.

Heteroskedastisuudella ei oikeastaan ole haitallisesta vaikutusta regressiokertoimien arvoon. Sen sijaan sillä voi olla vaikutusta niiden tilastolliseen merkitsevyyteen. Tämä voi johtaa esimerkiksi tilanteeseen, jossa tietty muuttuja ei näytä olevan tilastollisesti merkitsevä Y:n selittäjä vaikka se todellisuudessa sellainen onkin. Heteroskedastisuusongelmien havainnoimiseksi on kehitetty erilaisia testejä, joita ei kuitenkaan esitellä tässä yhteydessä. Yksinkertaisin tapa havainnoida mahdollisia heteroskedastisuusongelmia on tehdä aineistosta alustavan regressioanalyysin jälkeen kuvion 4 kaltaisia hajontakuvioita jokaisen selittävän muuttujan osalta. Jos hajontakuviot tai testit osoittavat, että aineistossa on heteroskedastisuutta, voidaan regressioanalyysin tulosten estimointiin käyttää sellaista menetelmää, joka pystyy ottamaan huomioon nämä ongelmat.

d) Havaintojen aikariippuvuus. Yksi regressioanalyysin perusolettamuksista on, että havaintojen virhetermit ovat toisistaan riippumattomia. Jos analysoitavana on aikasarja-aineisto (katso tutkimusasetelmat), tämä oletus ei useinkaan ole pätevä. Tämä johtuu siitä, että eri ajankohtina kerättyjen havaintojen virhetermit korreloivat keskenään. Jos analysoitavana on esimerkiksi työttömyyden taso jossain maassa eri vuosina, on tietyn vuoden työttömyystaso osittain riippuvainen edellisen vuoden tasosta. Jos tätä riippuvuutta ei oteta huomioon, regressioanalyysin tulokset vääristyvät. Havaintojen aikariippuvuuden korjaamiseksi on useita eri tapoja. Näistä kerrotaan esimerkiksi Ostromin kirjassa sekä ekonometrian oppikirjoissa (ks. lisätietoja-linkki).

Lähteet

- U.S Bureau of Census (1998): *World Population Profile: 1998*. Washington.

Faktorianalyysi

Tutkiessaan potilasta lääkäri kysyy häneltä erilaisista taudin oireista: särkeekö päätä tai vatsaa, onko kuumetta tai väsymystä jne. Nämä oirekuvaukset johtavat diagnoosiin eli päätelmään siitä, mikä sairaus potilasta vaivaa. Lääkäri ei siis suoraan pysty tunnistamaan tautia, vaan hän tekee sen epäsuorasti taudin aiheuttamien oireiden perusteella. Tällainen päättelytapa kuvaa hyvin faktorianalyysia. Faktorianalyysissa pyritään löytämään havaintoyksikön ominaisuuksia kuvaavasta muuttujajoukosta piileviä yhdenmukaisuuksia eli faktoreita. Ajatuksena on, että tiettyjä havaintoyksikköjen ominaisuuksia ei pystytä havainnoimaan suoraan, vaan niistä saadaan ainoastaan epäsuoraa tietoa. Faktori voidaan käsittää eräänlaisena hypoteettisena konstruktiona tai teoreettisena käsitteenä, jonka olemassaolo päätellään konkreettisista havainnoista. Yhteiskuntatieteiden alalla tällaisia käsitteitä ovat esimerkiksi 'konservatiivisuus', 'sosiaalisuus', 'älykkyys' tai 'työpaikan ilmapiiri'.

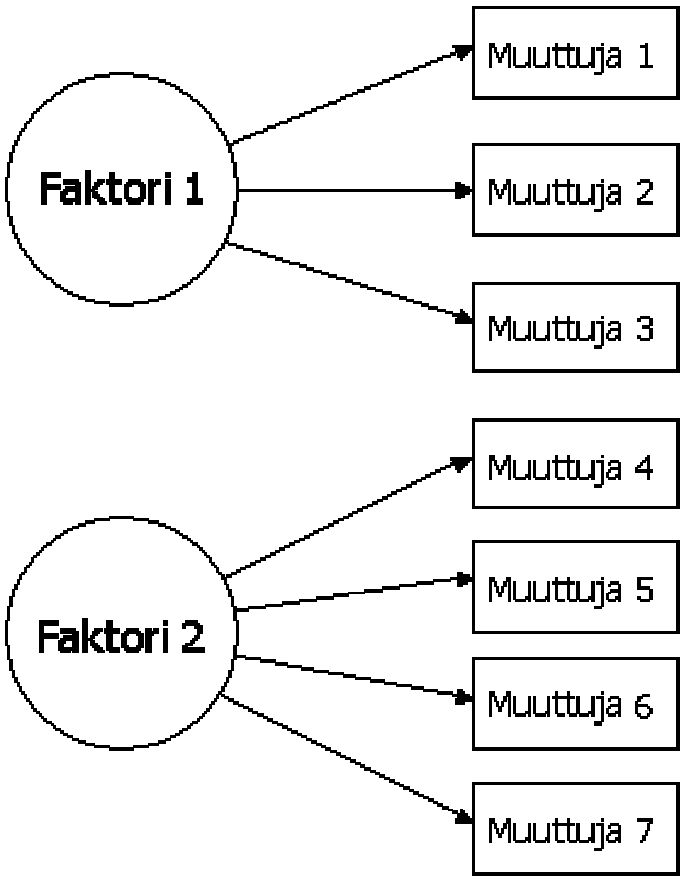
Tyyppiesimerkki faktorianalyysista on älykkyyden mittaaminen. Älykkyyttä ei voida suoraan havainnoida, mutta erilaisten älykkyystestien tuloksena ihmisen älykkyydestä on tehty päätelmiä. Lähtökohtana on ollut oletus siitä, että ihmisillä on tietty latentti eli piilevä ominaisuus (älykkyys), josta seuraa havaittavia ilmiöitä (oikeat vastaukset älykkyystestin eri osioihin). Yhteiskuntatieteissä faktorianalyysia käytetään usein mielipidekysymysten analysoinnissa. Jos tutkija haluaa esimerkiksi tutkia ihmisten konservatiivisuutta, ei tätä ominaisuutta voida tarkasti mitata ainoastaan yhden kysymyksen perusteella, vaan ilmiötä monipuolisesti mittaavia kysymyksiä tarvitaan useita. Jotkin näistä kysymyksistä liittyvät suoremmin konservatiivisuuden käsitteeseen, jotkin vähemmän suorasti. Faktorianalyysin avulla voidaan tutkia, muodostavatko annetut vastaukset yhteisen faktorin, joka sitten voidaan tulkita konservatiivisuusfaktoriksi. Tuloksena voi myös olla useita faktoreita, jotka kuvastavat konservatiivisuuden eri ulottuvuuksia.

Faktorianalyysin perusidea

Faktorianalyysissa voidaan erottaa kaksi toisistaan poikkeavaa lähestymistapaa. **Eksploratiivinen faktorianalyysi** pyrkii etsimään muuttujajoukosta faktoreita, jotka pystyvät selittämään havaittujen muuttujien vaihtelua ilman, että tutkijalla on etukäteen vahvoja odotuksia löydettävien faktoreiden määrästä tai niiden tulkinnasta. Eksploratiivinen faktorianalyysi on siis aineistolähtöinen tutkimusmenetelmä. Analyysin tuloksena voidaan löytää yksi tai useampia faktoreita, joita käytetään hyväksi tulosten tulkinnassa. **Konfirmatorisessa faktorianalyysissa** tutkijalla on jo etukäteen teorian pohjalta muodostettu käsitys aineiston faktorirakenteesta ja analyysin tehtävänä on joko varmistaa tai kumota tämä käsitys empiirisen aineiston pohjalta. Eksploratiivinen faktorianalyysi on näistä kahdesta faktorianalyysin muodosta yleisempi, joten tässä yhteydessä esittely keskittyy lähinnä siihen. Lopussa käsitellään lyhyesti myös konfirmatorista faktorianalyysia.

Kuviossa 1 on esitetty faktorianalyysin perusidea yksinkertaistetun kaavion muodossa. Kuviossa on kaksi faktoria ja seitsemän havaittua muuttujaa. Muuttujat voivat olla esimerkiksi kyselylomakkeen seitsemän väittämää, joilla on pyritty mittaamaan vastaajien konservatiivisuutta. Faktorit ovat tavallaan 'piilomuuttujia', koska niitä ei voida suoraan havainnoida, vaan niiden olemassaolo päätellään ainoastaan havaittujen muuttujien avulla. Käytännössä faktorin muodostaa joukko muuttujia, jotka korreloivat vahvasti keskenään, mutta vähän muiden muuttujien kanssa. Kuviossa faktoreista lähtee nuolia havaittuihin muuttujiin. Ne kuvaavat faktorianalyysin pohjana olevaa oletusta, jonka mukaan piilevät faktorit aiheuttavat havaitut ilmiöt, eikä päinvastoin.

Faktorianalyysi tuottaa jokaista kuvion nuolen 'vahvuutta' kuvaavan arvon eli **faktorilatauksen** (*factor loading*). Latauksen suuruus kertoo kuinka paljon faktorin avulla pystytään selittämään havaitun muuttujan vaihtelusta. Lataukset saavat arvoja -1 ja 1 välillä. Mitä lähempänä latauksen itseisarvo on yhtä (1) sitä vahvemmin muuttuja latautuu faktorilla (eli sitä paremmin faktori selittää muuttujan vaihtelua). Jos muuttujan lataus on arvoltaan negatiivinen, kertoo se ainoastaan sen, että muuttujan arvot korreloivat negatiivisesti faktorin arvojen kanssa. Jos faktori kuvaa esimerkiksi konservatiivisuutta ja yksi muuttuja saa vahvan, mutta negatiivisen latauksen, tarkoittaa tämä sitä, että konservatiivisia piirteitä omaavat vastaajat ovat vastanneet kysymykseen pienillä arvoilla, kun taas muihin kysymyksiin he ovat vastanneet suurilla arvoilla.



Kuvio 1. Faktorianalyysin idea yksinkertaistettuna.

Faktorimallin toimivuutta voidaan arvioida faktoreiden ominaisarvojen ja havaittujen muuttujien kommunaliteettien avulla. **Ominaisarvot** (*eigenvalue*) ilmoittavat, kuinka hyvin faktorit pystyvät selittämään havaittujen muuttujien hajontaa. Mitä suurempi faktorin ominaisarvo on, sitä paremmin se selittää muuttujien hajontaa ja päinvastoin. Kun faktorin ominaisarvo jaetaan havaittujen muuttujien määrällä, saadaan faktorin suhteellinen selitysosuus, joka saa arvoja nollan ja yhden välillä. Selitysosuus kertoo, kuinka suuri osuus kaikkien mallissa mukana olevien havaittujen muuttujien hajonnasta voidaan faktorin avulla selittää. Mitä suurempi osuus on, sitä parempi on faktorin selitysvoima. Kun kaikkien faktoreiden selitysosuudet lasketaan yhteen, saadaan koko analyysin selitysosuus. Se kertoo siis, kuinka suuri osuus kaikkien havaittujen muuttujien hajonnasta voidaan selittää kaikilla löydetyillä faktoreilla.

Kommunaliteetti (*communality*) puolestaan kertoo, kuinka suuri osuus yksittäisen havaitun muuttujan vaihtelusta selittyy löydettyjen faktorien avulla. Jos muuttujan kommunaliteetti on lähellä yhtä, pystyvät faktorit selittämään sen vaihtelun lähes kokonaan.

Toisaalta mitä pienempiä arvo kommunaliteetti saa, sitä huonommin faktorit muuttujaa selittävät. Jos yksittäisen muuttujan kommunaliteetti on pieni, kannattaa harkita, onko muuttujaa ylipäänsä syytä sisällyttää analyysiin.

Esimerkki faktorianalyysista

Faktorianalyysin periaatteet ovat helpoimmin ymmärrettävissä esimerkin avulla. Seuraavassa tehdään faktorianalyysi vuoden 1996 World Values -kyselyn Suomen osa-aineistosta (katso aineistonkuvaus verkkoversiosta). Aineiston yksi osuus koostuu 11 asiasta, joiden hyväksyttävyyttä vastaajat arvioivat yksittäin. Kysymys kuului: "Voisitteko sanoa jokaisesta seuraavaksi luettelemastani asiasta, ovatko ne aina hyväksyttäviä, ei koskaan hyväksyttäviä vai jotain siltä väliltä?". Vastaajat ilmaisivat hyväksymisensä asteen valitsemalla vastauksensa väliltä 1-10 niin, että arvo 1 tarkoitti "ei koskaan hyväksyttävä" ja arvo 10 "aina hyväksyttävä". Taulukossa 1 on esitetty kaikki 11 arvioitavaa asiaa, niitä vastaavan muuttujan nimi sekä lyhenne, jota käytetään jatkossa muuttujien ilmaisemiseksi.

Muut-tuja	**Lyhenne**	**Koko kysymys**
v192	sosiaaliturva	Vaatia sairauskorvaus tai sosiaaliturvaetu, johon ei ole oikeutta
v193	kulkuneuvo	Jättää maksamatta julkisessa kulkuneuvossa
v194	verovilppi	Tehdä verovilppiä, jos tilaisuus sallii
v195	varastettu tavara	Ostaa tavaraa, jonka tietää olevan varastettua
v196	lahjukset	Lahjusten ottaminen virkatehtävien hoidossa
v197	homoseksuaalisuus	Homoseksuaalisuus
v198	prostituutio	Prostituutio
v199	abortti	Abortti
v200	avioero	Avioero
v201	eutanasia	Eutanasia, parantumattomasti sairaiden elämän lopettaminen
v202	itsemurha	Itsemurha

Taulukko 1. Faktorianalyysin muuttujat.

Taulukon 1 kaikkiin muuttujiin liittyy moraalisia näkökohtia. Kysymyksen sanamuodossa ei puhuta sanatarkasti siitä, hyväksyvätkö vastaajat asiat omalla kohdallaan. Yleinen sanamuoto on silti kategorinen, koska vastausvaihtoehtojen ääripäät kuvaavat täydellistä hyväksymistä tai kieltämistä. Faktorianalyysi paljastaa muodostavatko kaikki 11 muuttujaa yhteisen faktorin, vai onko tuloksena useita faktoreita. Jos tuloksena on useita faktoreita, on mielenkiintoista nähdä, voidaanko näille faktoreille antaa mielekäs tulkinta. Kuvaavatko ne jäsentyneesti suhtautumisen eri ulottuvuuksia? Aina faktorianalyysin tuloksena ei löydy järkevästi tulkittavaa faktorirakennetta.

Eksploratiivisessa faktorianalyysissa tehtäessä tutkijan tulee aluksi päättää analysoitavien faktorien määrä. Tällaisessa induktiivisesti etenevässä analyysissa faktorien määrää ei ole etukäteen rajattu, joten valinta täytyy tehdä jonkin muun kriteerin perusteella. Yleensä tilasto-ohjelmistot laskevat faktorianalyysin aluksi aineistosta yhtä monta faktoria kuin analyysiin on valittu muuttujia. Kaikkien näin syntyvien ulottuvuuksien käyttö jatkoanalyysissa ei ole kuitenkaan järkevää, koska faktorianalyysin perusideana on tiivistää muuttujien sisältämä informaatio suppeahkoon määrään faktoreita. Edellä mainittiin, että faktorin ominaisarvo kuvaa sitä, kuinka hyvin se pystyy selittämään havaittujen muuttujien hajontaa. Yleisesti käytetty nyrkkisääntö faktorien määrän valinnalle on, että jatkoanalyysiin otetaan vain sellaiset faktorit, joiden ominaisarvo on suurempi kuin yksi. On kuitenkin hyvä muistaa, että tämä sääntö perustuu

vain käytäntöön, eikä sille ole mitään vahvaa tilastotieteellistä perustetta. Tilasto-ohjelmistot sisältävät yleensä useita vaihtoehtoisia tapoja faktoreiden lukumäärän määrittelemiseksi.

Seuraava vaihe faktorianalyysissa on niin sanottu faktoreiden **rotaatio** (*rotation*). Rotaatiolla (eli faktoriakselien kiertämisellä) viitataan prosessiin, jonka tarkoituksena on tehdä faktorianalyysin tulosten tulkinta helpommaksi. Rotaatio ei juurikaan muuta tuloksia sisällöllisesti, se tekee niistä vain helpommin tulkittavia. Rotaatiomenetelmät voidaan jakaa kahteen pääluokkaan. **Suorakulmarotaatiomenetelmät** (*orthogonal rotation*) tuottavat sellaisia faktoreita, jotka eivät korreloi keskenään ja **vinokulmarotaatiomenetelmät** (*oblique rotation*) puolestaan faktoreita, jotka voivat korreloida keskenään. Tässä yhteydessä ei käsitellä näiden menetelmien yksityiskohtia. "Lisätietoja" -osuudessa on esitelty kirjoja, joissa kerrotaan tarkemmin rotaation yksityiskohdista (esimerkiksi Nummenmaa ym. 1996). Yleisesti rotaation käyttämistä faktorianalyysin yhteydessä voidaan suositella, koska se lähes poikkeuksetta tekee faktorilatausten teoreettisen tulkinnan helpommaksi.

Palataan taulukon 1 esimerkkiin. Aineisto tuotti kaksi faktoria, joiden ominaisarvo oli suurempi kuin yksi. Taulukkoon 2 on kirjattu esimerkkifaktorianalyysin lopulliset tulokset suorakulmaisen rotaation jälkeen. Rotaatiomenetelmänä käytettiin ns. varimax-rotaatiota.

Analyysi tuotti kahden faktorin mallin, jossa viisi ensimmäistä muuttujaa latautuu vahvimmin toisella faktorilla ja loput muuttujista ensimmäisellä faktorilla. Taulukossa on tummennettu kaikki faktorilataukset, joiden arvo on vähintään 0,5. Entä kuinka suuri latauksen täytyy olla ollakseen merkittävä? Tähän ei ole yksikäsitteistä vastausta. Jotkut oppikirjat suosittelevat alarajaksi arvoa 0,3, toiset taas esimerkiksi 0,5.

Muuttuja	**Faktori 1 'Vapaamielisyys'**	**Faktori 2 'Lainkuuliaisuus'**	**Kommunaliteetti**
sosiaaliturva	0,02	**0,52**	0,27
kulkuneuvo	0,24	**0,69**	0,53
verovilppi	0,18	**0,68**	0,50
varastettu tavara	0,14	**0,75**	0,58
lahjukset	0,03	**0,55**	0,30
homoseksuaalisuus	**0,64**	0,03	0,41
prostituutio	**0,63**	0,23	0,45
abortti	**0,80**	0,09	0,65
avioero	**0,75**	0,02	0,56
eutanasia	**0,61**	0,14	0,39
itsemurha	**0,50**	0,13	0,27
Ominaisarvo	2,75	2,17	
Selitysosuus	25,0%	19,8%	

Taulukko 2. Faktorianalyysin tulokset (rotaation jälkeen).

Faktorianalyysista ei ole paljoakaan hyötyä, jos sen tuottamille faktoreille ei pystytä antamaan mielekästä sisällöllistä tulkintaa. Tämän vuoksi tulkinta onkin keskeinen osa faktorianalyysia. Taulukon 2 tulokset viittaavat siihen, että eri asioiden hyväksynnässä voidaan erottaa kaksi ulottuvuutta. Ensimmäisellä faktorilla latautuvat vahvasti sellaiset muuttujat, jotka kuvaavat vastaajien suhtautumista yhteiskunnallisiin ja uskonnollisväritteisiin arvokysymyksiin. Abortti, itsemurha, homoseksuaalisuus jne. ovat perinteisiä ja vuosien saatossa paljon keskustelua herättäneitä yhteiskunnallisia ilmiöitä. Nämä kaikki asiat (eutanasiaa lukuun ottamatta) ovat nyky-Suomessa laillisia, mutta silti jotkut niistä ovat yhä kiistanalaisia. Ensimmäistä faktoria kutsutaan 'vapaamielisyydeksi'.

Toisella faktorilla latautuvat asiat ovat kaikki laittomia tekoja, vaikka useimmat niistä ovat ainakin rikoslain mukaan vain suhteellisen pieniä rikkomuksia (ehkä lahjusten vastaanottoa lukuun ottamatta). Arkielämässä monet ihmiset eivät pidä tällaisia pikkuvilppejä kovinkaan vakavina rikkomuksina. Toinen faktori siis kuvastaa suhtautumista pikkulaittomuuksia kohtaan. Yksinkertaisinta lienee kutsua sitä 'lainkuuliaisuus'-faktoriksi.

Faktorianalyysin tulosten tulkinnassa tulee muistaa, että faktorilatauksista ei voida tulkita mitään siitä, kuinka suuri osa vastaajista hyväksyy tai paheksuu kysymyksissä esitettyjä asioita. Jos tämä on mielenkiinnon kohteena, kannattaa analysoida esimerkiksi muuttujien keskiarvoja ja hajontoja. Faktorianalyysin tulokset kertovat ainoastaan sen, että vastaajat omassa mielessään jaottelevat kysymyksissä luetellut asiat kahden löydetyn ulottuvuuden (faktorin) mukaan.

Tuloksia voidaan tarkastella tarkemmin myös tekniseltä kannalta. Ensimmäisen faktorin ominaisarvo on 2,75 ja sen selitysosuus 25%. Tämä tarkoittaa, että faktori pystyy selittämään neljänneksen kaikkien havaittujen muuttujien hajonnasta, mitä voidaan pitää kohtuullisen hyvänä tuloksena. Faktorin ominaisarvo saadaan ottamalla kaikista faktorilatauksista neliö ja laskemalla saadut arvot yhteen ($0,02^2+0,24^2+...+0,61^2+0,50^2=2,75$). Selitysosuus puolestaan saadaan jakamalla ominaisarvo muuttujien määrällä (2,75/11=0,25). Toisen faktorin selitysosuus on noin 20%, joten kaiken kaikkiaan molemmat faktorit selittävä yhteensä 45% havaittujen muuttujien hajonnasta. Tätä voidaan pitää suhteellisen tyydyttävänä tuloksena.

Muuttujien kommunaliteetit kertovat sen, kuinka hyvin faktorit selittävät yksittäisen muuttujan hajontaa. Kommunaliteetti lasketaan korottamalla muuttujan faktorilataukset neliöön ja laskemalla ne yhteen. Esimerkiksi sosiaaliturvamuuttujan kommunaliteetti saadaan laskemalla $0,02^2+0,52^2=0,27$. Taulukon 2 kaikki kommunaliteetit ovat arvoltaan 0,27 tai suurempia. Tämä tarkoittaa sitä, että analyysista ei tarvinne pudottaa mitään muuttujaa pois. Mitään täsmällistä tilastollista kriteeriä kommunaliteetin arvon riittävälle tasolle ei ole. Muuttujien poistamisessa analyysista täytyy aina käyttää tapauskohtaista harkintaa.

Faktoripisteet

Edellä esitellyt faktorianalyysin tulosten tulkinnat voivat olla sinänsä riittäviä yksittäisen tutkimuksen tarpeisiin. Joskus on kuitenkin mielenkiintoista tietää, miten eri vastaajaryhmät sijoittuvat faktoreiden suhteen. Tällaisessa analyysissa voidaan käyttää hyväksi **faktoripisteitä** (*factor scores*), jotka kuvaavat jokaisen aineiston havainnon sijoittumista eri faktoreilla. Faktoripisteet saadaan laskemalla painotettu keskiarvo alkuperäisten muuttujien standardoiduista arvoista. Painoina käytetään faktorilatauksia. Tällä menetelmällä saatujen uusien faktoripistemuuttujien keskiarvo on aina nolla. Toinen ja yksinkertaisempi vaihtoehto käyttää faktorianalyysin tuloksia hyväksi jatkoanalyysissa on muodostaa summamuuttujat niistä muuttujista, jotka latautuvat vahvasti kullakin faktorilla (eli ns. kärkimuuttujista).

Tilasto-ohjelmistot laskevat tarvittaessa faktoripisteet automaattisesti. Faktoripisteitä voidaan käyttää jatkoanalyysissa joko selittävinä tai selitettävinä muuttujina. Faktoripisteiden käytön havainnollistamiseksi taulukossa 3 on esitetty muutaman vastaajaryhmän sijoittuminen edellisen faktorianalyysin kahdella faktorilla. Tuloksilla on tarkoitus vain konkretisoida esimerkinomaisesti faktoripisteiden käyttömahdollisuuksia. Tarkempi analyysi edellyttäisi selittävien muuttujien huolellista seulontaa sekä havaittujen erojen tilastollisen merkitsevyyden analyysia esimerkiksi varianssianalyysin tai regressioanalyysin keinoin. Kannattaa myös huomata, että 'lainkuuliaisuus' -faktorilla pienet arvot kuvaavat sellaista vastaajaa, joka ei hyväksy laittomia toimia ja suuret arvot vastaajaa, joka on valmiimpi hyväksymään faktorilla latautuneita asioita.

Ryhmä	Keskiarvo ensimmäisellä faktorilla ('Vapaamielisyys')	Keskiarvo toisella faktorilla ('Lainkuuliaisuus')
Alle 35-vuotiaat	0,17	0,17
60-vuotta tai yli	-0,43	-0,34
Nainen	-0,01	-0,14
Mies	0,02	0,13
Ei ammatillista koulutusta	-0,15	0,20
Korkeakoulututkinto	0,20	-0,14

Taulukko 3. Valittujen ryhmien faktoripisteiden keskiarvot.

Taulukosta 3 nähdään, että vastaajan iällä vaikuttaisi olevan suuri vaikutus siihen, kuinka he ajattelevat vapaamielisyys-faktorilla latautuvista kysymyksistä. Alle 35-vuotiaat saavat molemmilla faktoreilla suuremmat keskiarvot kuin 60-vuotiaat tai vanhemmat, mikä osoittaa, että he ovat suhteellisesti valmiimpia hyväksymään sekä arvofaktorilla latautuneet asiat että toisella faktorilla latautuneet laittomuudet. Vastaajan sukupuolella ei ole suurtakaan vaikutusta vapaamielisen suhtautumisen yleisyyteen. Sen sijaan toisella faktorilla miesten faktoripisteiden keskiarvo on hiukan suurempi kuin naisten. Miehet ovat siis valmiimpia hyväksymään pikkuvilpin kuin naiset. Myös koulutuksella näyttäisi olevat vaikutusta. Hyvin koulutetut vastaajat ovat vapaamielisempiä kuin vähän koulutusta saaneet. Lainkuuliaisuusfaktorilla keskiarvot ovat päinvastaisia, eli vähän koulutetut ovat vähemmän lainkuuliaisia pikkuvilppiasioissa.

Konfirmatorinen faktorianalyysi

Konfirmatorisen faktorianalyysin lähtökohtana on, että tutkijalla on jo ennen analyysin suorittamista teoriaan perustuva oletus aineiston faktorirakenteesta. Näin konfirmatorinen faktorianalyysi on teorialähtöinen analyysimenetelmä. Sen käyttö edellyttää, että tutkijalla on käytettävissään hyvin muotoiltuja hypoteeseja muuttujien välisistä suhteista ja siitä, kuinka monta faktoria havaitut muuttujat muodostavat. Tämä on lähes päinvastainen lähestymistapa eksploratiiviseen faktorianalyysin verrattuna. Eksploratiivinen faktorianalyysi on luonteeltaan aineistolähtöinen eikä sen käyttö edellytä vahvoja ennakko-oletuksia aineiston faktorirakenteesta.

Käytännössä konfirmatorisen faktorianalyysi edellyttää sitä, että tutkijan on tiedettävä etukäteen, mitkä muuttujat latautuvat milläkin faktorilla ja korreloivatko faktorit keskenään vai ei. Kun faktorit ja niihin kuuluvat muuttujat on valittu, voidaan konfirmatorinen faktorianalyysi toteuttaa. Analyysin tulokset kertovat, miten hyvin alkuperäiset odotukset aineiston faktorirakenteesta pitävät paikkansa. Arviointi perustuu useisiin tilastollisiin tunnuslukuihin, jotka kuvaavat faktorimallin soveltuvuutta aineistoon.

Rakenneyhtälömallit (structural equation models) ovat eräänlainen konfirmatorisen faktorianalyysin laajennus. Niissä yhdistyvät sekä faktori- että regressioanalyysi. Hieman yksinkertaistettuna rakenneyhtälömallien ideana on tarkastella regressioanalyysin avulla faktorien välisiä kausaalisuhteita. Tutkimuksessa on esimerkiksi voitu muodostaa faktorianalyysin avulla vastaajien masentuneisuutta ja itsetuntoa koskevat faktorit. Rakenneyhtälömallien avulla voidaan tutkia, minkälainen kausaalinen vaikutus itsetunnolla on masentuneisuuteen.

Useimpien yleiskäyttöisten tilastoanalyysiohjelmien perusmodulit eivät sisällä konfirmatoriseen faktorianalyysiin ja rakenneyhtälömallien analysointiin tarkoitettuja työkaluja. Esimerkiksi SPSS-tilasto-ohjelman yhteydessä käytetään AMOS-nimistä konfirmatoriseen faktorianalyysiin erikoistunutta ohjelmaa, joka täytyy hankkia erikseen. 'Lisätietoja'-osuudessa on linkkejä eri ohjelmiin, joilla konfirmatorisia faktorimalleja ja rakenneyhtälömalleja voidaan toteuttaa.

Logistinen regressio

Logistinen regressioanalyysi on tavanomaisen regressioanalyysin erityistyyppi. Sitä käytetään silloin, kun selitettävä muuttuja voi saada vain kaksi arvoa. Voidaan esimerkiksi pyrkiä selittämään sitä, miten eri tekijät vaikuttavat siihen, onko vastaaja naimisissa vai ei.

Tavallisessa regressioanalyysissa selitettävän muuttujan arvot voivat vaihdella paljonkin. Regressioanalyysi ei kuitenkaan ole käyttökelpoinen silloin, kun selitettävän muuttujan arvot rajoittuvat vain kahteen vaihtoehtoon. Logistinen regressioanalyysi ei pyri ennustamaan määriä, vaan todennäköisyyksiä. Kyse on siis siitä, millä todennäköisyydellä tarkasteltavana oleva asia tapahtuu tai pätee. Tulokset kertovat, vaikuttavatko selittävät muuttujat tapahtuman todennäköisyyteen ja kuinka suuri vaikutus on. Esimerkiksi äänestämistutkimuksen tulokset voivat kertoa, että naisilla on suurempi todennäköisyys äänestää kuin miehillä tai että iän kasvaessa osallistumistodennäköisyys kasvaa.

Logistisen regressiomallin idea

Logistisessa regressioanalyysissa selitettävä muuttuja täytyy koodata niin, että se voi saada ainoastaan arvon yksi tai nolla. Oletetaan, että tutkimuksessa on tarkoitus selvittää, mitkä tekijät vaikuttavat ihmisten äänestysaktiivisuuteen. Selitettävä muuttuja mittaa sitä, äänestikö vastaaja viime vaaleissa. Se saa arvon nolla, jos vastaaja ei äänestänyt (eli Y=0) ja arvon yksi jos hän äänesti (Y=1).

Logistisen regressioanalyysin ymmärtämiseksi täytyy tietää, mitä riskisuhteella tarkoitetaan. Oletetaan, että äänestystutkimuksen otoksessa naisista 70 % ja miehistä 60 % ilmoitti äänestäneensä viime vaaleissa. Näiden lukujen avulla voidaan naisille ja miehille laskea ns. **riskisuhde** (*odds ratio*). Riskisuhdetta käytetään yleisesti esimerkiksi kuvattaessa vedonlyönnin voittosuhteita. Riskisuhde saadaan suhteuttamalla naisten äänestämistodennäköisyys miesten vastaavaa lukuun. Näin saadaan tulokseksi 1,17 (=0,7/0,6), mikä tarkoittaa sitä, että naisilla on 1,17 kertaa suurempi todennäköisyys äänestää kuin miehillä. Riskisuhde voidaan laskea myös toisinpäin. Miesten todennäköisyys äänestää on 0,86-kertainen (=0,6/0,7) naisten vastaavaan todennäköisyyteen verrattuna.

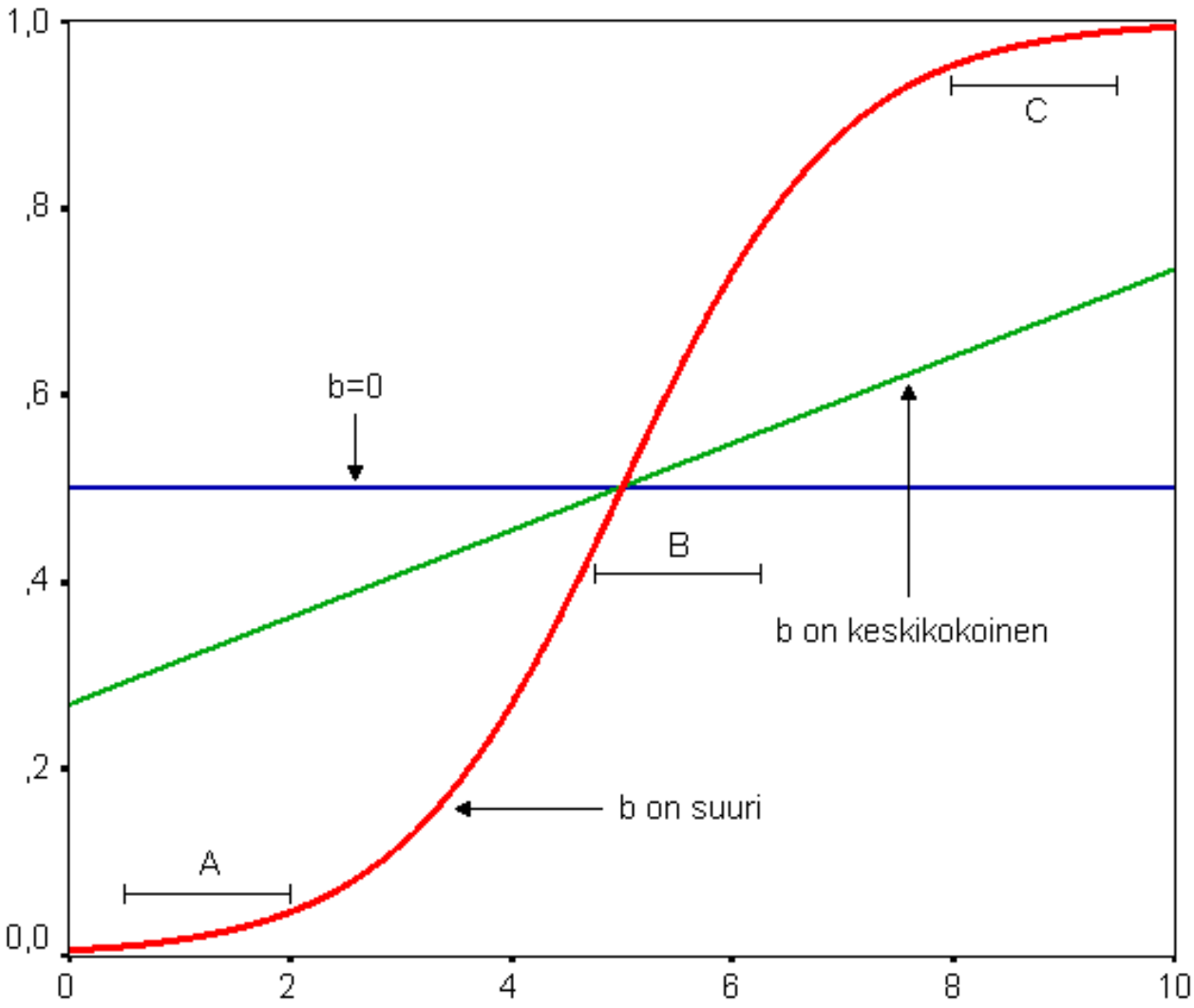
Riskisuhde voi saada arvoja nollan ja äärettömän välillä. Tavanomainen regressioanalyysi soveltuu kuitenkin parhaiten tilanteeseen, missä selitettävän muuttujan arvoja ei ole rajattu millekään ennalta määrätylle välille. Siksi logistista regressioanalyysia varten riskisuhteesta otetaan vielä logaritmi. Tämä varmistaa sen, että saatu luku vaihtelee äärettömän pienien ja äärettömän suurien lukujen välillä.

Yksinkertaistettuna logistinen regressiomalli on tavallinen regressiomalli, jossa selitettävänä muuttujana on riskisuhteen logaritmi. Tämä voidaan ilmaista kaavalla seuraavasti:

$$\log\left[\frac{P(Y=1)}{1-P(Y=1)}\right] = a + bx$$

Kaavassa P(Y=1) on todennäköisyys sille, että selitettävä muuttuja saa arvon yksi, a on vakiotekijä, b regressiokerroin ja x selittävän muuttujan arvo. Logistisen regressiomallin kaavan lauseke a+bx on täsmälleen sama kuin normaalissa regressioanalyysissä. Siksi logistisen regressiomallin tulkinta ja siihen liittyvät ongelmat ovat lähes samat kuin regressioanalyysissa.

Tulkinnassa täytyy kuitenkin ottaa huomioon se, että logistisessa regressiomallissa selittävien ja selitettävän muuttujan suhde ei ole lineaarinen, vaan siinä oletetaan suhteen seuraavan niin sanotun s-käyrän (eli logistisen käyrän) muotoa. Kuviossa 1 on esitetty kuvitteellinen esimerkki logistisista käyristä. Esimerkissä selittävä muuttuja x-akselilla saa arvoja nollasta kymmeneen. Logistisen regressioanalyysin tulos on y-akselilla. Logistisessa regressioanalyysissa selitettävän tapahtuman todennäköisyys saa arvoja nollan ja yhden välillä.



Kuvio 1. Esimerkkejä logistisesta s-käyrästä.

Jos selittävällä ja selitettävällä muuttujalla ei ole lainkaan yhteyttä toisiinsa logistisessa regressiomallissa, saa regressiokerroin b itseisarvoltaan hyvin pienen arvon. Kuten kuviosta 1 nähdään, on muuttujien yhteyttä kuvaava käyrä täysin vaakasuora silloin, kun b saa arvon nolla. Tämä osoittaa sen, että selitettävän muuttujan mittaaman tapahtuman todennäköisyys ei muutu ollenkaan selittävän muuttujan arvojen vaihdellessa. Silloin kun kerroin b saa suuren arvon, on selittävän muuttujan arvojen ja tapahtuman todennäköisyyden yhteyttä kuvaava käyrä s-kirjaimen muotoinen. Tämä tarkoittaa sitä, että jos selittävän muuttujan pieni arvo kasvaa hiukan, ei tämä muuta paljoakaan selitettävän muuttujan mittaaman tapahtuman todennäköisyyttä (väli A). Sen sijaan selittävän muuttujan saadessa arvoja vaihteluvälin keskivaiheilta pienikin muutos aiheuttaa suuren muutoksen selitettävän ilmiön tapahtumistodennäköisyydessä (väli B). Selittävän muuttujan ollessa lähellä ylärajaa muutoksilla on jälleen pienempi vaikutus (väli C).

Kun kertoimen b arvo on keskikokoinen, on sen muoto vaakasuoran ja s-käyrän välimailla. Jos kertoimen arvo on negatiivinen, laskee selitettävän muuttujan mittaaman tapahtuman todennäköisyys selittävän muuttujan arvon kasvaessa. Tällöin logistiset käyrät ovat samanmuotoisia kuin kuviossa 1, mutta ne laskevat vasemmalta oikealle.

Logistisen regressiomallin kertoimien tulkinta eroaa tavallisen regressiomallin kertoimien tulkinnasta siinä, että tavallisessa regressiomallissa yhden yksikön muutos selittävässä muuttujassa aiheuttaa aina samansuuruisen muutoksien selitettävässä muuttujassa. Sen sijaan

logistisessa regressioanalyysissa selitettävän todennäköisyyden muutos riippuu b-kertoimen lisäksi selittävän muuttujan arvosta. Tämän takia logistisen regressiomallin tulosten tulkinta on aina hankalampaa kuin tavallisessa regressiomallissa.

Esimerkki logistisesta regressioanalyysista

Logistisen regressioanalyysin esimerkissä tutkitaan, mitkä tekijät vaikuttavat suomalaisten protektionismin kannatukseen. Vuoden 1996 World Values Surveyn Suomen osa-aineistossa (katso aineistokuvaus verkkoversiosta) on kysymys, jossa vastaajien piti valita kahdesta vaihtoehdosta, kumpi on heidän mielestään parempi (v133). Nämä vaihtoehdot olivat 1) "Muissa maissa valmistettuja tuotteita voidaan tuoda tänne ja myydä täällä, jos ihmiset haluavat ostaa niitä" ja 2) "Ulkomaisten tuotteiden myynnille Suomessa pitäisi olla enemmän esteitä, jotta voitaisiin suojella tämän maan ihmisten työpaikkoja". Näistä jälkimmäinen edustaa protektionistista ajattelutapaa.

Vastaajista noin 40 prosenttia valitsi ensimmäisen ja noin 60 prosenttia jälkimmäisen vaihtoehdon. Analyysia varten muuttuja on koodattu niin, että ensimmäinen vaihtoehto saa arvon nolla ja jälkimmäinen arvon yksi. Näin logistisen regressioanalyysin avulla voidaan tutkia siis, mitkä tekijät vaikuttavat vastaajien todennäköisyyteen valita protektionistinen vaihtoehto.

Analyysin selittäjinä käytetään viittä eri muuttujaa. Demografisista muuttujista mukana ovat vastaajan ikä (v216) ja sukupuoli (v214, koodattuna dummy-muuttujaksi seuraavasti: mies=0, nainen=1). Vastaajan tulotasoa mitataan 10-luokkaisella muuttujalla (v227), jossa suuret arvot tarkoittavat korkeampia tuloja. Asennemuuttujista mukana on vastaajien ylpeys suomalaisuudestaan (v205). Se on mitattu neliportaisella asteikolla, jossa pienet arvot kuvaavat suurempaa ylpeyttä. Hypoteesina on, että ne vastaajat, jotka ovat ylpeitä suomalaisuudestaan ovat valmiimpia kannattamaan protektionismia. Lisäksi analyysissa on mukana muuttuja, joka kuvaa vastaajan sijoittumista politiikan vasemmisto-oikeisto -ulottuvuudella (v123). Se saa arvoja yhdestä kymmeneen pienten arvojen kuvastaessa sijoittumista vasemmalle. Oletuksena on, että vasemmalle identifioituvat vastaajat todennäköisemmin hyväksyvät protektionistiset ajatukset ulottuvuuden oikeaan laitaan sijoittuvat vastaajat.

Muuttuja	**Regressiokerroin**	**Merkitsevyys**
Vakio	-0.00	p=0,99
Sukupuoli (nainen=1, mies=0)	0,48**	p=0,001
Ikä	0,02**	p<0,001
Ylpeys suomalaisuudesta (1=suuri...4=heikko)	-0,10	p=0,33
Sijoittuminen vasemmisto-oikeisto – ulottuvuudella (1-10)	-0,07	p=0,11
Tuloluokka (1-10)	-0,08*	p=0,01

Taulukko 1. Logistinen regressioanalyysi protektionismin kannatukseen vaikuttavista tekijöistä.

Logistisen regressioanalyysin tulokset ovat taulukossa 1. Mallin toimivuuden tarkastelu kannattaa aloittaa muuttujien merkitsevyystasojen analyysilla. Vastaajien poliittista sijoittumista ja heidän ylpeyttään suomalaisuudesta kuvaavat muuttujat eivät ole tilastollisesti merkitseviä tekijöitä protektionismin selittäjinä. Sen sijaan muut muuttujat ovat tilastollisesti merkitseviä. Ikä-muuttujan regressiokerroin on positiivinen, mikä kertoo sen, että vanhemmat ihmiset ovat valinneet protektionistisen vaihtoehdon nuorempia todennäköisemmin. Myös sukupuolimuuttuja on positiivinen eli naiset valitsevat miehiä todennäköisemmin protektionistisen vaihtoehdon. Tuloluokkamuuttuja saa negatiivisen kertoimen. Se kertoo, että suurituloisilla on pienituloisempia vähäisempi todennäköisyys kannattaa protektionistista vaihtoehtoa.

Logistisen regressiomallin ennustearvoa voidaan tarkastella katsomalla, kuinka hyvin sen avulla pystytään luokittelemaan vastaajat oikeisiin luokkiin heidän vastaustensa mukaan. Taulukon 1 regressiomalli ennustaa oikein 80 prosenttia niistä vastaajista, jotka valitsivat protektionistisen vaihtoehdon. Toisaalta malli ennustaa oikein vain 37 prosenttia niistä, jotka valitsivat vapaata kauppaa arvostavan vaihtoehdon. Näin mallin ennustekyky on parhaimmillaankin vain kohtalainen. Toisin sanoen taulukon 1 sisältämien muuttujien avulla ei pystytä ennustamaan kovinkaan tarkasti vastaajien kantaa protektionismiin. Samalla on huomattava, että selitettävänä muuttujana ollut protektionismimittari on hyvin karkea, ja suhtautumista olisikin kannattanut mitata laajemmalla skaalalla. Logistista regressioanalyysia onkin tarkoituksenmukaisinta käyttää silloin, kun selitettävää ilmiötä ei ole mitattu tai ei voida mitata tarkemmin kuin kaksijakoisesti.

Multinomiaalinen logistinen regressio

Multinomiaalinen logistinen regressio (*multinomial logistic regression*) on tavallisen logistisen regressioanalyysin laajennus, jossa selitettävä muuttuja voi saada useampia kuin pelkästään kaksi vaihtoehtoa. Kuvitellaan esimerkiksi tilanne, jossa luokitteluasteikolla mitattu selitettävä muuttuja voi saada kolme eri vaihtoehtoa: A, B ja C. Multinomiaalisessa logistisessa regressioanalyysissa tutkitaan, mitkä tekijät vaikuttavat siihen, että vastaaja on valinnut tietyn vaihtoehdon suhteessa muihin vaihtoehtoihin. Käytännössä tämä tarkoittaa sitä, että tässä esimerkkitapauksessa tuloksena saadaan kolme erilaista mallia. Yhdessä verrataan vaihtoehdon A valintaa suhteessa vaihtoehtoon B, toisessa A:n valintaa suhteessa C:hen ja kolmannessa B:n valintaa suhteessa C:hen.

Tässä yhteydessä ei käsitellä multinomiaalista logistista regressioanalyysia tarkemmin. Menetelmästä kiinnostuneen kannattaa katsoa 'Lisätietoja' -osuudesta kirjallisuusvinkkejä.

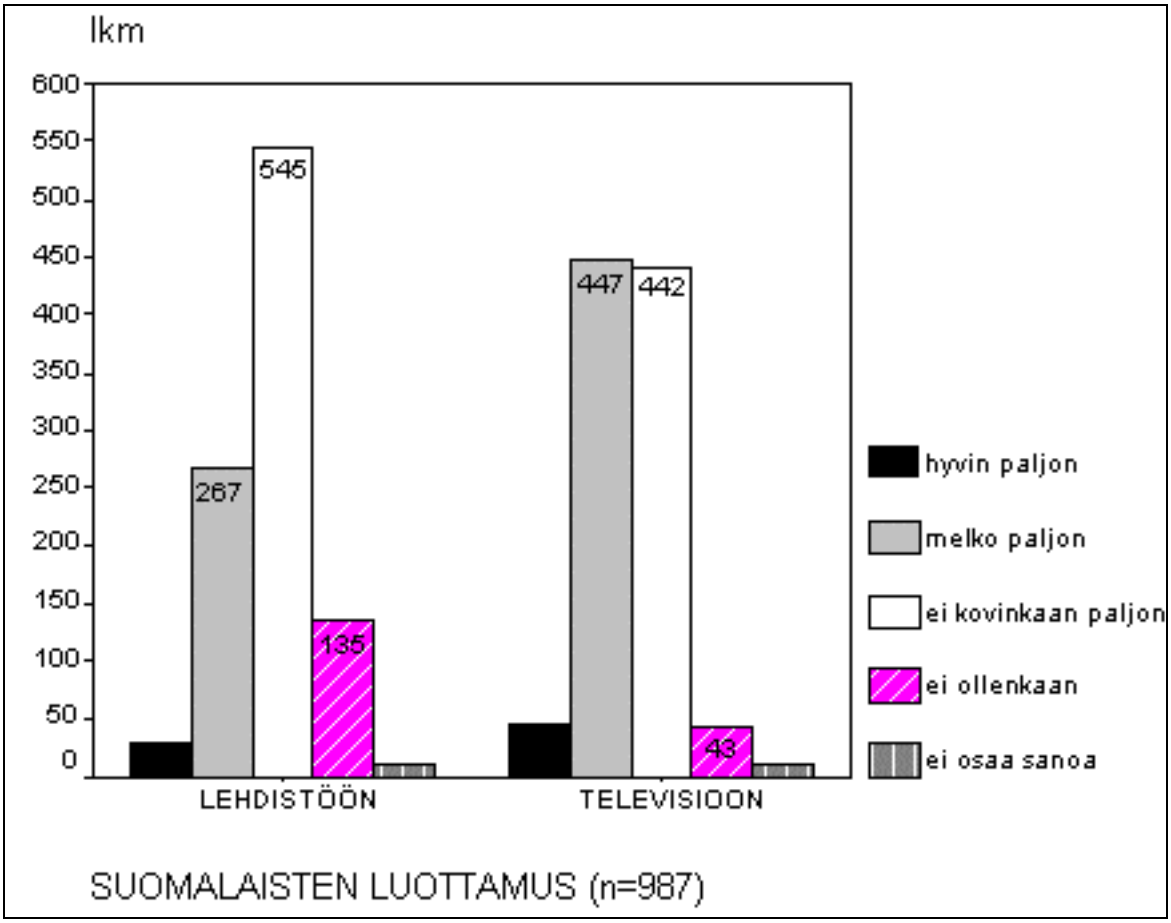
Graafinen esitys (kuviot)

Kuviot ovat visuaalinen tapa havainnollistaa ilmiöitä. Niiden tarkoituksena on helpottaa oleellisen informaation hahmottamista. Tässä tekstissä kuviolla tarkoitetaan nimenomaan tilasto-ohjelmistolla aikaansaatavaa kuviota, ei esim. rakennekaaviota. Kuvioiden tekemisen taustalla on tietty näkökulma ja tarkasteltavat muuttujat, joiden jakaumia kuvataan, ovat luonteeltaan erilaisia. Ne asettavat kuviolle tiettyjä vaatimuksia, joiden huomioiminen lisää kuvioiden pätevyyttä. On olemassa vaara, että rutinoidutaan käyttämään yhtä kuviotyyppiä kaikissa tilanteissa. Tyypillisimpiä graafisia esitystapoja ovat pylväsdiagrammi ja viivadiagrammi. Joskus pylväsdiagrammi voitaisiin korvata sektoridiagrammilla ja joskus voisi olla informatiivisempaa käyttää laatikko-jana -esitystä tai korrelaatiodiagrammia. Analysointivaiheessa luonnollisesti tutkitaan muuttujien välisiä yhteyksiä useilla eri tavoilla, mutta julkaistavaksi valitaan kuvio, joka on luonteenomaisin ja selkein kussakin tilanteessa.

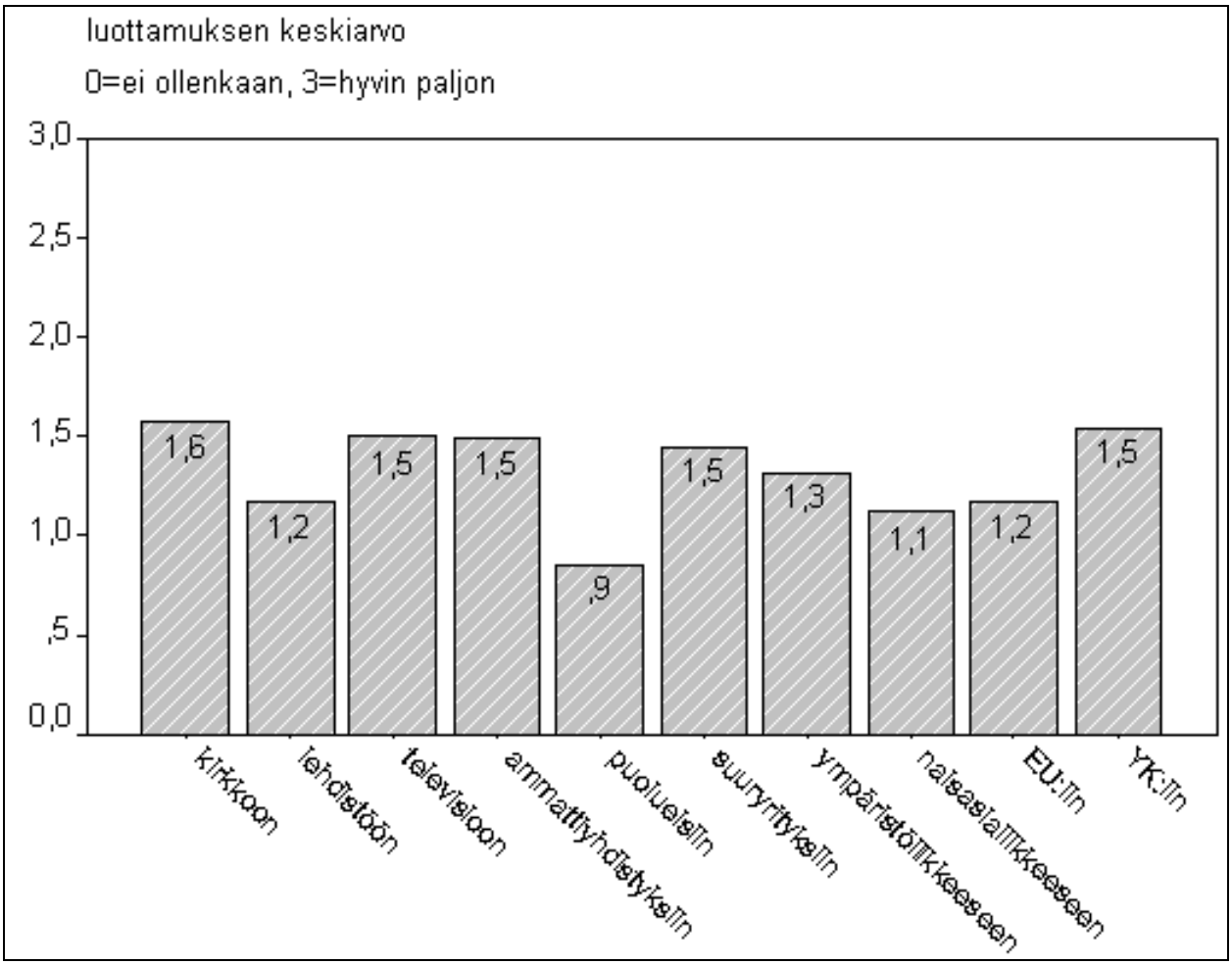
Yksiulotteisen jakauman eli yhden muuttujan arvojen jakautumisen tarkasteluun liittyy oleellisesti jakauman sijainnin ja hajonnan kuvaaminen. Kahta muuttujaa tarkasteltaessa ollaan yleensä kiinnostuneita niiden yhteisjakaumasta ts. halutaan tietää, onko muuttujien välillä keskinäistä riippuvuutta. Tutkittavana voi olla esimerkiksi, millainen yhteys näyttäisi tuloilla olevan mielipiteeseen siitä, kuinka paljon henkilö on valmis käyttämään kunnan rahoja omaishoidon tukeen. Riippuvuutta voidaan tarkastella pylväs-, viiva- ja laatikko-jana -kuviolla sekä korrelaatiodiagrammilla.

Kahden muuttujan tilanteessa näkökulma voi olla myös se, että toinen muuttuja jakaa aineiston osa-aineistoiksi ja toisen muuttujan jakaumaa tarkastellaan näissä osa-aineistoissa. Verrataan esimerkiksi eri kuntien asukkaiden mielipiteitä: eroavatko tamperelaisten ja oululaisten asenteet suhteessa omaishoidon tukeen. Tällöin verrataan asennemuuttujien absoluuttisia- tai prosenttijakaumia tai vertailua tehdään tunnusluvuilla. Mikäli käytetään pylväskuvioita, muuttujien jakauma- tai tunnuslukupylväät (esim. keskiarvopylväät) voidaan tehdä vierekkäin samaan kuvioon tai ne voivat olla erillisinä kuvioina.

Näiden kahden em. näkökulman raja ei selviä aina suoraa muuttujien perusteella, mutta tutkija voi itse valita näkökulman. On tärkeää tiedostaa tietyn näkökulman esille nostaminen myös kuviossa. Näkökulmapohdintoja voi harjoittaa esimerkkikuvioiden tilanteissa.



Kuvio 1. Kahden muuttujan absoluuttisten jakaumien vertailun mahdollistava pylväskuvio. (Suomalaisten luottamus eri instituutioihin.)



Kuvio 2. Jopa kymmenen eri muuttujan sijainnin vertailu onnistuu keskiarvokuviolla.

Kolmea muuttujaa tarkasteltaessa voidaan myös valita kahden eri näkökulman välillä. Yhtäältä muuttujien väliset suhteet voidaan nähdä siten, että halutaan tarkastella kahden, yleensä taustamuuttujien, yhdysvaikutusta kolmanteen muuttujaan. Tässä voidaan hyödyntää ns. typologioiden muodostamista. Esimerkiksi iän ja sukupuolen ollessa taustamuuttujia ja asennemuuttujan riippuva muuttuja, saadaan selvitettyä nuorten naisten, vanhojen naisten, nuorten miesten ja vanhojen miesten välisten asenteiden eroja ja samanlaisuuksia, vaikkapa laatikko-jana -kuvioilla. Konkreettisesti typologiat saadaan muodostamalla tilasto-ohjelmassa uusi muuttuja, joka on yhdistelmä kahdesta muuttujasta: esimerkin tapauksessa tämän uuden muuttujan arvoja ovat 'nuoret naiset', 'vanhat naiset', 'nuoret miehet' ja 'vanhat miehet'. Tällöin palataan kuvioiden tekemisessä kahden muuttujan tilanteeseen.

Toisaalta voi olla tilanne, jossa yksi muuttuja jakaa tarkasteltavan aineiston ryhmiin, osa-aineistoihin, ja näissä ryhmissä vertaillaan kahden muuttujan riippuvuuksia. Verrataan esimerkiksi Tampereella ja Oulussa iän ja asenteiden välisiä riippuvuuksia. Jos riippuvuuksia halutaan verrata kahdessa tai useammassa osajoukossa, vertailuja voidaan tehdä suoraa jakaumakuvioilla tai tunnuslukukuvioilla (esim. keskiarvo tai mediaani, kvartiilit). Tällöin voidaan esimerkiksi verrata eri ryhmille (tamperelaisille ja oululaisille) tehtyjä 100 %:n pylväskuvioita toisiinsa.

Kuten kahta muuttujaa yhtäaikaa tarkasteltaessa, myös kolmen muuttujan tilanteessa, sekä osa-aineisto- että yhdysvaikutusnäkökulmasta katsottaessa, voidaan käyttää samojakin kuvioita. Myös osa-aineistotarkasteluissa voidaan käyttää mm. typologioita, esimerkiksi nuoret tamperelaiset, vanhat tamperelaiset ja nuoret oululaiset ja vanhat oululaiset. Jos halutaan korostaa osa-aineistoja, niille tehdään erilliset kuviot.

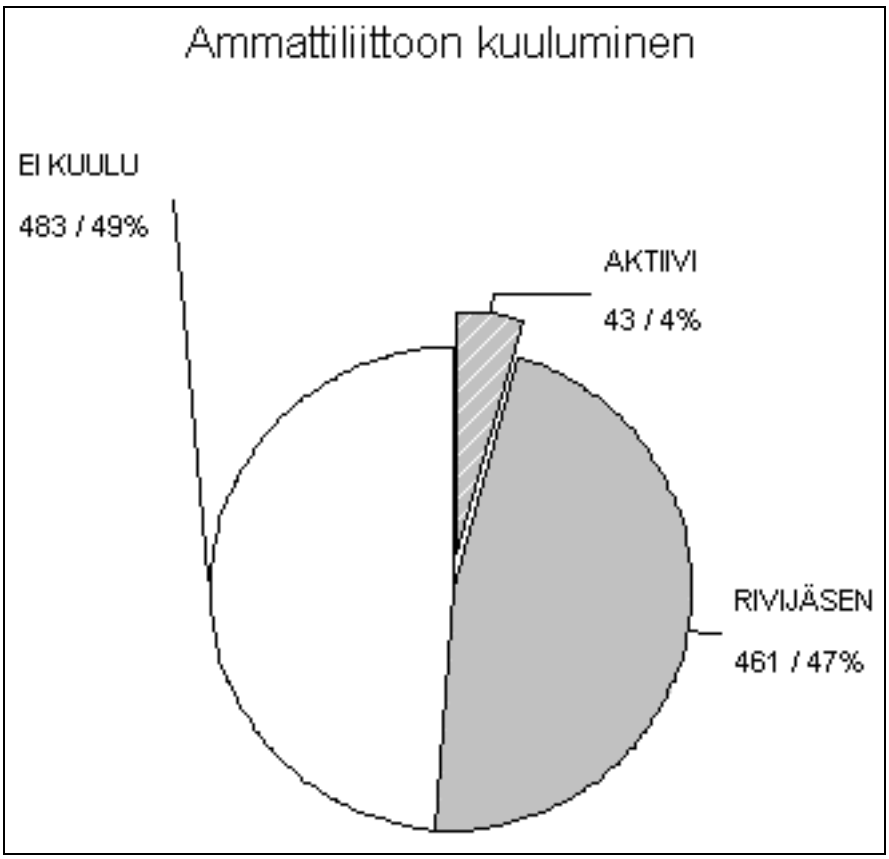
Kahden tai useamman muuttujan kuvioiden tulkintaa helpottaa, jos vaaka-akselilla on taustamuuttuja tai riippumaton muuttuja, mikäli tällainen asetelma muuttujien välillä on mahdollinen. Samoin prosenttiosuudet on syytä määritellä taustamuuttujan tai riippumattoman muuttujan ryhmissä. Näissä ryhmissä kussakin prosenttien summa on 100. Vaikka muuttujien välillä ei varsinaisesti voi määritellä, kumpi on taustamuuttuja, kysymyksen asettelu määrää, miten prosenttiosuudet lasketaan: Ollaanko kiinnostuneita ikäjakaumista sukupuolittain vai sukupuolista ikäluokittain. On myös tilanteita, joissa prosenttiosuus kokonaismäärästä on sisällöllisesti paras vaihtoehto.

Kuvion informatiivisuutta ajatellen siihen ei ole syytä laittaa liikaa tietoa - ei siis liian monta muuttujaa eikä liian useita luokkia. Julkaisuun valinnassa kannattaa erityisesti pohtia, onko kuviolla todella sille kuuluva erityismerkitys, jolla se palvelee lukijaa. Myös kuvioihin liittyvillä muotoseikoilla voidaan parantaa luettavuutta. Esimerkiksi keskenään vertailtaviksi tarkoitettujen kuvioiden asteikkojen on oltava samoja tai mahdollisimman vertailukelpoisia. On myös olemassa joitakin vakiintuneita ja hyväksi havaittuja käytäntöjä, kuten se, että kuvioiden otsikot kirjoitetaan julkaisuissa kuvioiden alapuolelle. (Ks. aiheesta luettavaa lisätiedoista.)

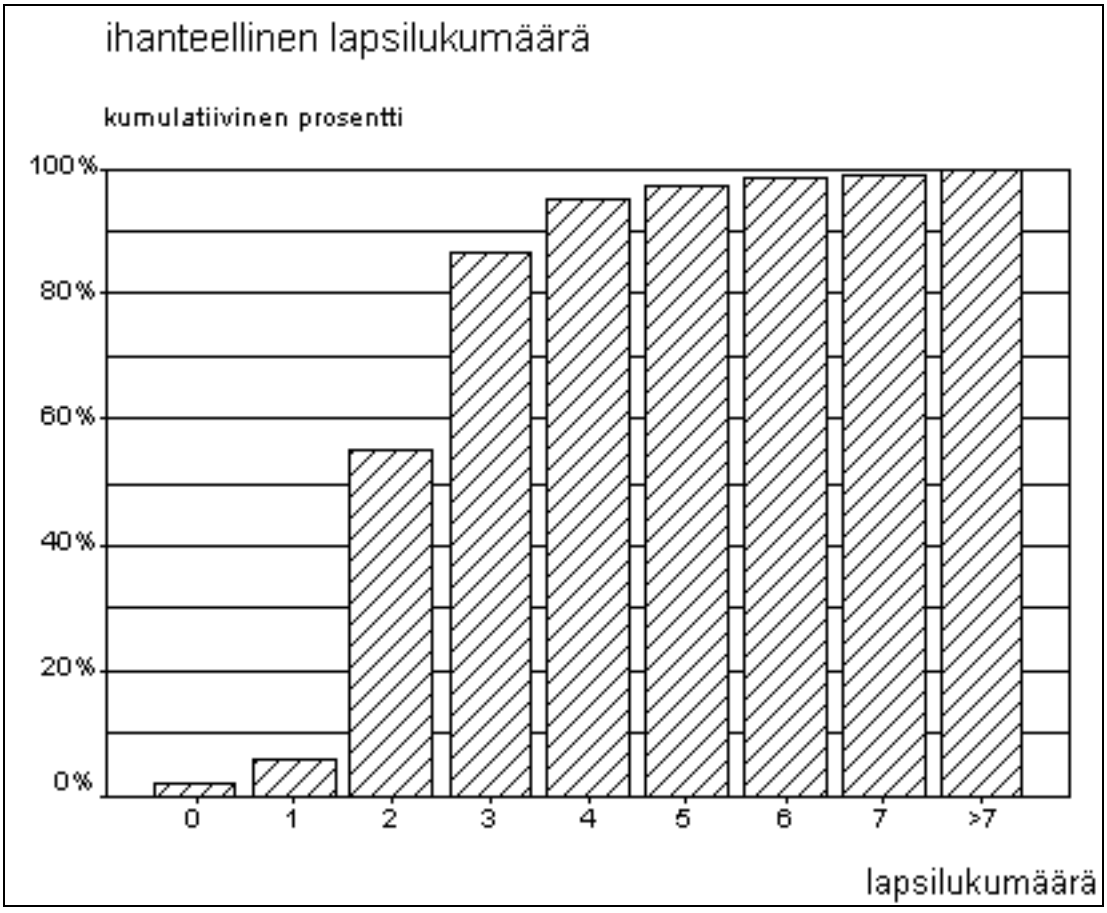
Seuraavassa muutamia pohdintoja erilaisiin kuvioihin liittyen sekä harvinaisemman laatikko-jana -kuvion tulkinnasta.

Sektoridiagrammi vai pylväskuvio?

Kun mietitään sektoridiagrammia ja pylväskuviota vaihtoehtoisina yksiulotteisen jakauman kuvaajina, voidaan huomioida seuraavia seikkoja. Pylväsdiagrammissa korostuu muuttujien arvojen järjestys enemmän kuin sektoridiagrammissa. Siinä on selvästi ensimmäinen ja viimeinen pylväs - olemmehan tottuneet lukemaan vasemmalta oikealle. Sektoridiagrammissa ei sen sijaan ole selvää alku- ja loppukohtaa. Pylväsdiagrammiin voidaan valita joko lukumäärät tai prosentit, mutta sektoridiagrammissa korostuvat prosenttiosuudet. Luokittelutasoiselle muuttujalle käytetään mielellään sektoridiagrammia, erityisesti silloin, kun halutaan korostaa prosenttiosuuksia: ympyrän koko ala on koko aineisto, 100 %, ja sen sektorien pinta-alat kuvaavat tarkasteltavan muuttujan arvojen jakautumista. Mikäli luokkia on kovin paljon, pylväskuvio on selkeämpi kuin sektoridiagrammi.



Kuvio 3. Sektoridiagrammi soveltuu hyvin kuvaamaan sellaista muuttujaa, joka ei saa kovin paljon eri arvoja, ja arvojen järjestystä ei haluta erityisesti korostaa.



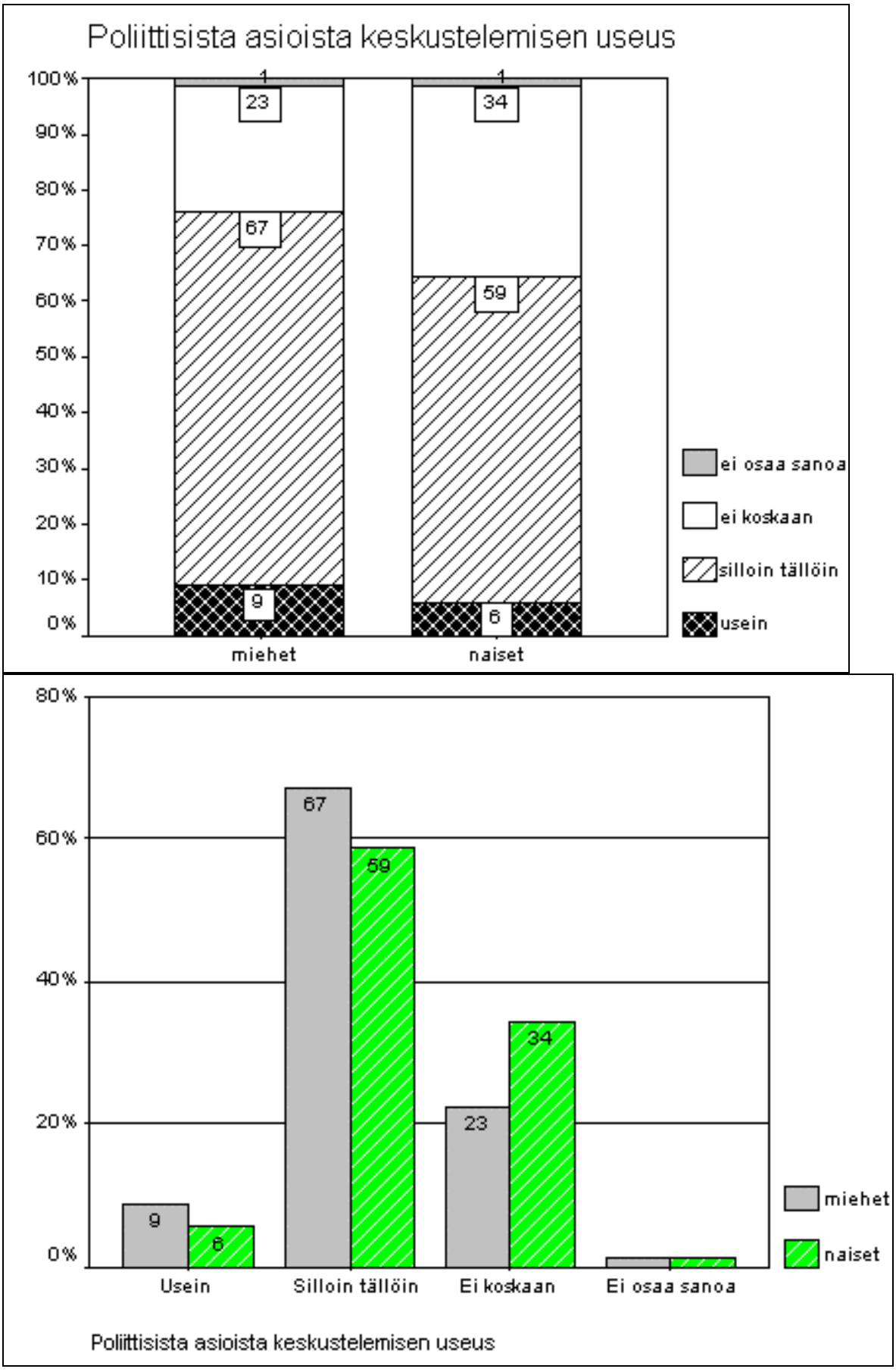
Kuvio 4. Pylväsdiagrammi soveltuu muuttujalle, joka saa useita arvoja. Se korostaa muuttujan diskreettisyyttä ja muuttujan arvojen järjestystä.

Pylväskuvio vai viivakuvio?

Mikäli halutaan kuvata kumulatiivisia eli summautuvia lukumääriä tai prosentteja, voidaan käyttää joko pylväskuviota tai viivakuviota. Pylväskuviota voidaan pitää näyttävämpänä, mutta viivoja paksuntamalla myös viivakuvioon saadaan voimaa. Muuttujan muutosta ajassa luonnehtii paremmin viivakuvio kuin pylväskuvio, sillä aika on ilmiönä jatkuva. Mittaukset, joihin kuvio perustuu, on luonnollisesti tehty tiettyinä ajanhetkinä. Jatkuvaa muuttujaa voi jatkuvuuden korostamiseksi myös kuvata yhteen liitetyillä pylväillä, joista käytetään nimitystä histogrammi. Erillisiä pylväitä käytettäessä aika ikään kuin pysähtyy tiettyinä ajanhetkinä. Kuvien tekemiseen käytettävä ohjelmisto voi kuitenkin asettaa rajoituksia esim. luokitusten tekemisessä histogrammiin.

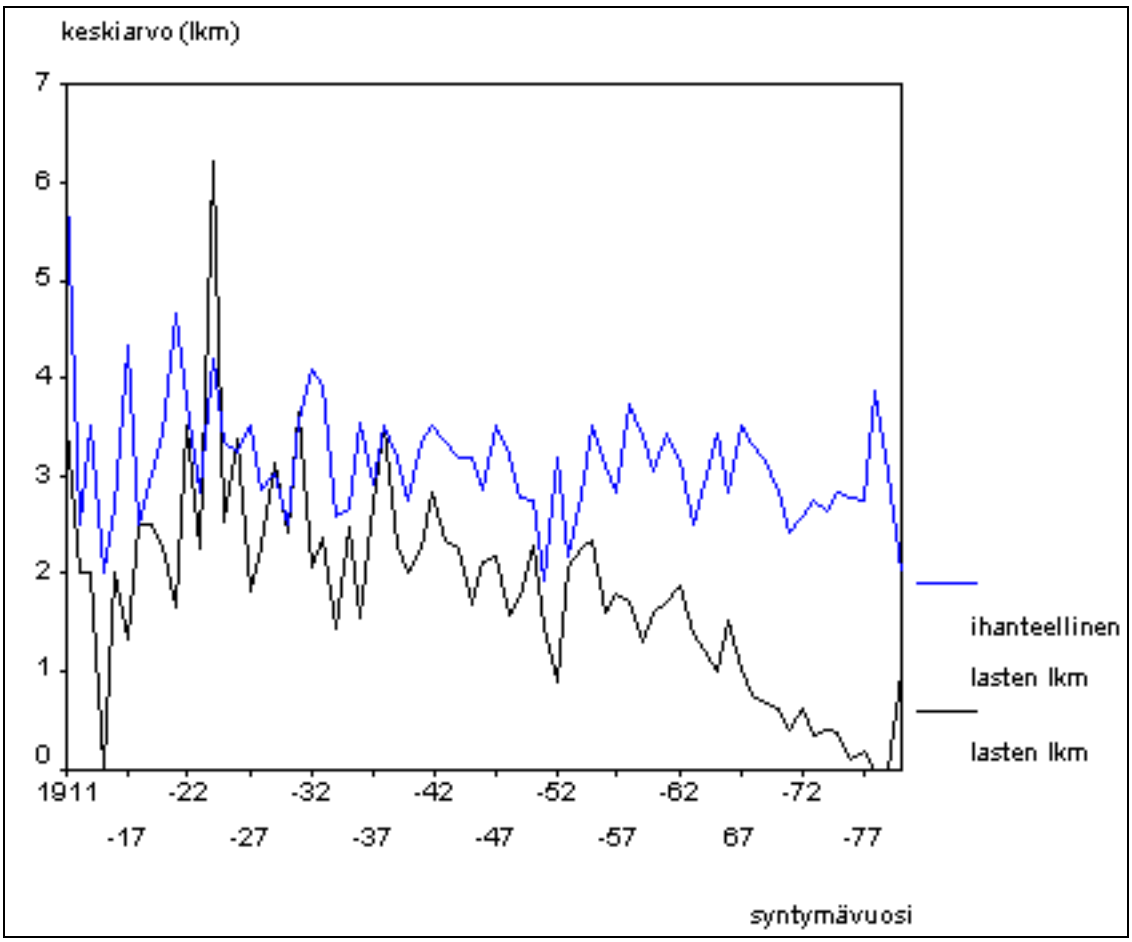
Kumulatiivisten kuvioiden ideana on se, että vasemmalta oikealle siirryttäessä lukumäärä tai prosentti sisältää myös vasemmalla puolella olevat määrät. Voidaan esimerkiksi ilmoittaa, että enintään kaksilapsista perhettä pitää ihanteena hiukan yli puolet suomalaisista. Tällaisesta kuviosta ei ole päätarkoitus nähdä, kuinka moni ihannoi kahden lapsen perhettä, vaan nimenomaan lapsettoman, yhden lapsen ja kahden lapsen perhettä ihannoivien "kasautunut" eli yhteismäärä.

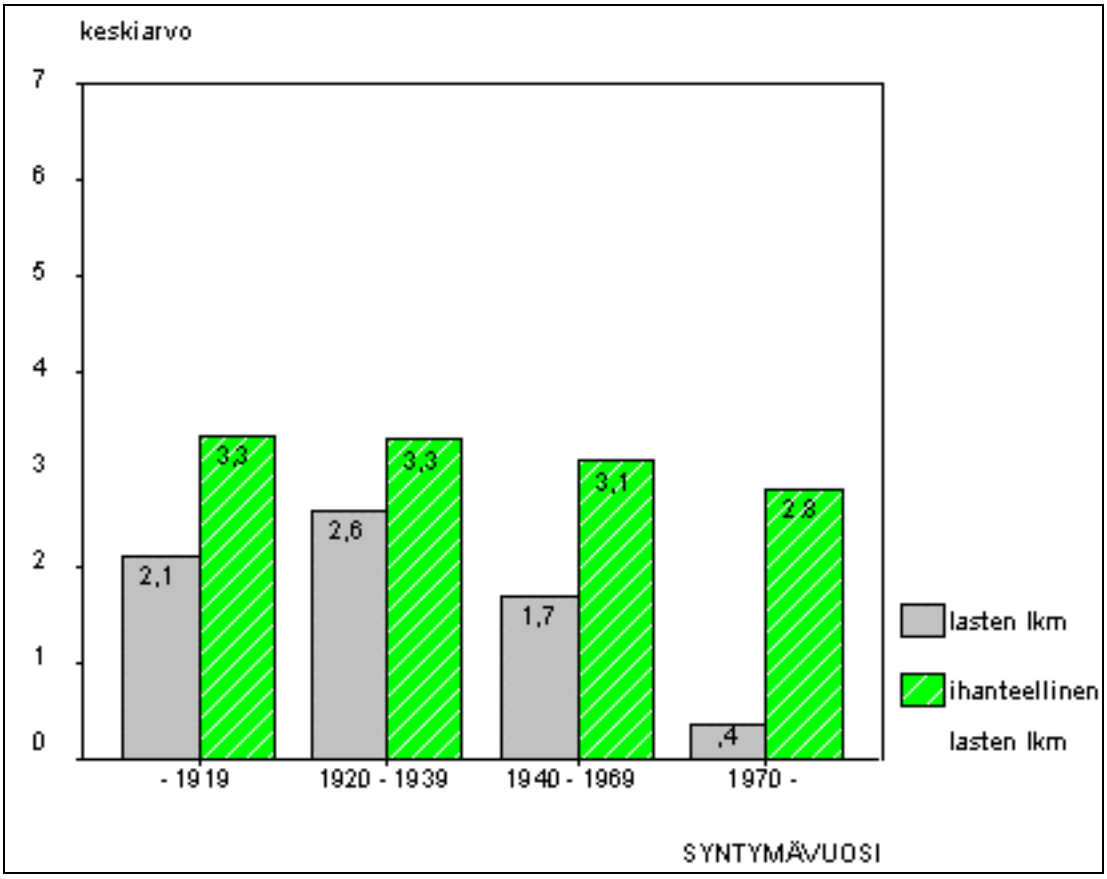
Kahden muuttujan välistä riippuvuuden tarkastelua voidaan havainnollistaa prosenttipylväillä, joko 100 %:n pylväskuvioina tai erillisistä prosenttipylväistä koostuvilla pylväiköillä. Tällöin vertaillaan toisen muuttujan luokissa toisen muuttujan prosenttijakaumia, esimerkiksi ikäluokittaisia asennejakaumia. Jakaumien vertaaminen lukumäärien avulla on hankalaa erityisesti silloin, kun ryhmittelevän muuttujan luokissa, esimerkiksi ikäluokissa on hyvin eri määrät tapauksia. Koska kahden muuttujan pylväsdiagrammissa on luettavuuden säilyttämiseksi oltava kohtuullinen määrä eri luokkia, paljon eri arvoja saavat muuttujat, esim. ikä, luokitellaan pylväsdiagrammin tekemistä varten.



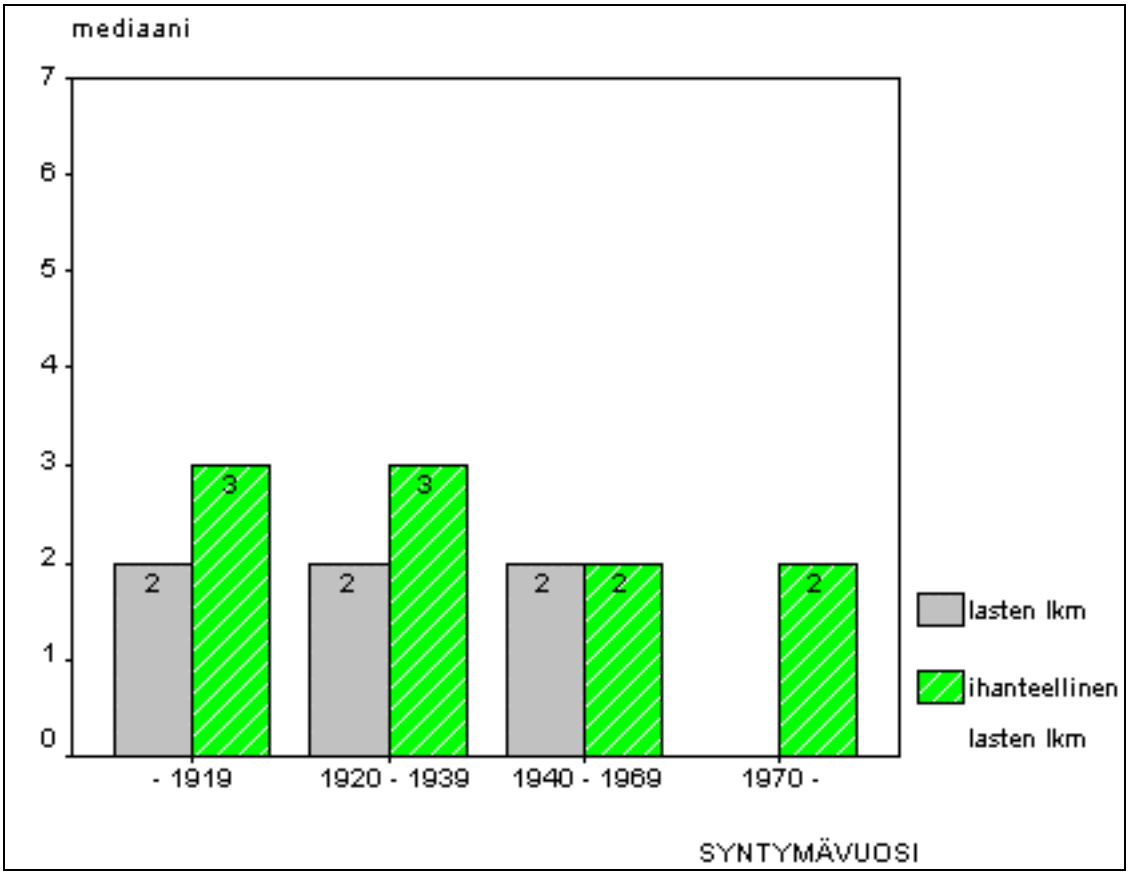
Kuvio 5. Kaksi vaihtoehtoista pylväskuviota prosenttijakaumien vertailuun.

Keskiarvo-, mediaani- ja moodipylväillä nähdään helposti jakaumien keskisijainti. Käytetyin ja kuvaavin on keskiarvokuvio. Vaikka tilastollisessa mielessä sitä ei voitaisi hyväksyä järjestystasoisille muuttujille, on kuitenkin todettava, että kuvattava ilmiö tulee yleensä paremmin esille keskiarvokuviossa kuin mediaani- tai moodikuviossa. Tästä syystä keskiarvo on yleisesti hyväksytty yhteiskuntatieteellisissä tutkimuksissa kuvaamaan järjestystasoisten muuttujien jakaumien sijaintia.





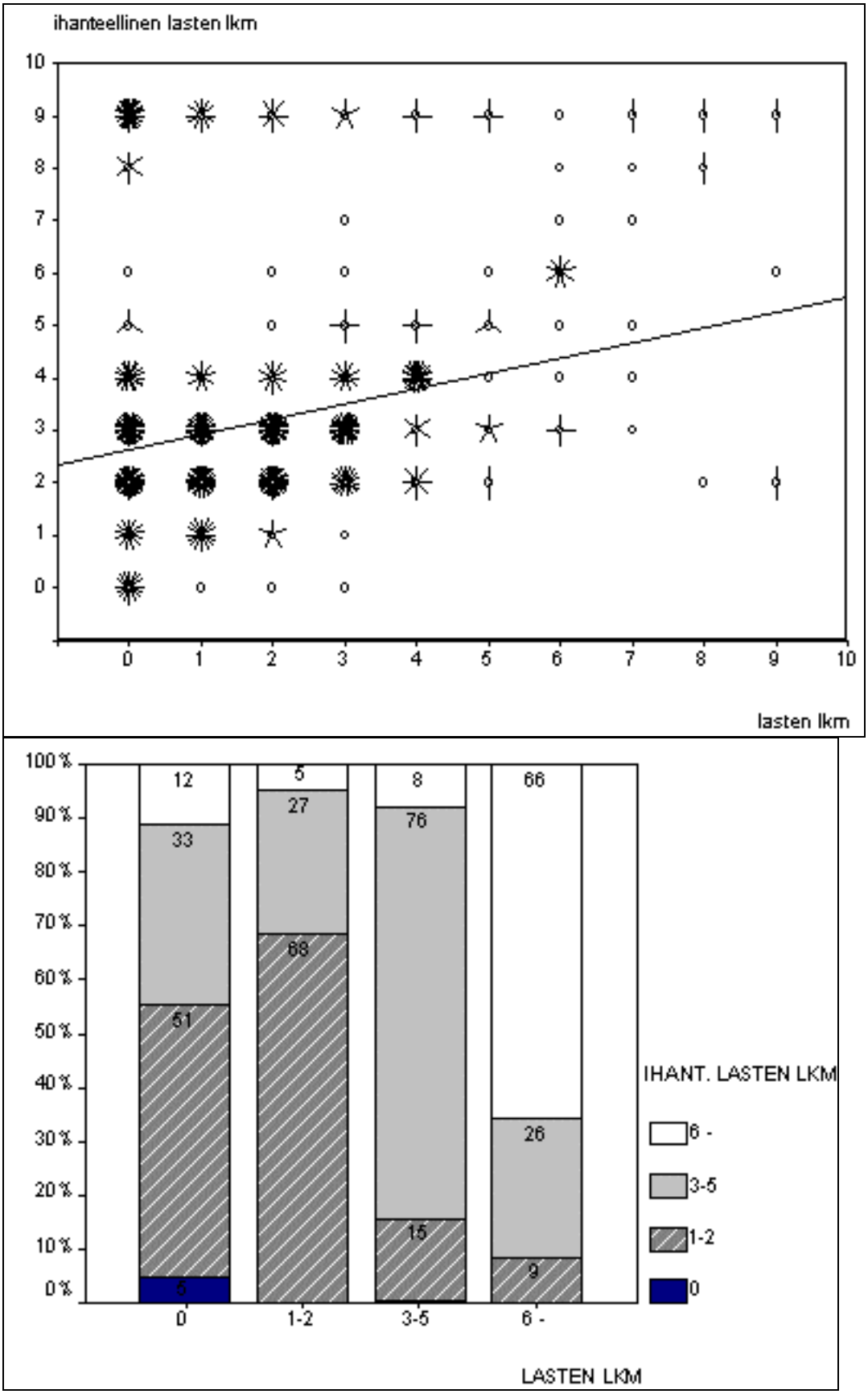
Kuvio 6. Samojen muuttujien keskiarvot kuvattuna viivakuviona ja pylväskuviona. Vaaka-akselilla on vastaajan syntymäaika. Lasten keskimääräistä lukumäärää on kuvattu keskiarvolla, vaikka ylimmät arvot on yhdistetty luokaksi, jota edustaa lukumäärä 8.



Kuvio 7. Lasten keskimääräinen lukumäärä on kuvattu mediaanilla. Tätä kuviota verrattaessa edelliseen, voidaan pohtia keskiarvon ja mediaanin antaman informaation eroja.

Korrelaatiodiagrammi vai pylväskuvio?

Korrelaatiodiagrammissa näkyy kahden muuttujan arvojen yhteisjakauma. Kutakin tilastoyksikköä vastaa yksi piste. Isossa aineistossa useat pisteet menevät päällekkäin. Korrelaatiodiagrammissa tarkastellaan nimenomaan muuttujien alkuperäisiä jakaumia, jolloin esim. ikää ei luokitella. Muuttujien on oltava vähintään järjestystasoisia. Järjestystason muuttujien yhteydessä on hyvä muistaa, että mittayksikköä ei todellisuudessa ole olemassa. Näin ollen korrelaatiodiagrammissa suoraviivaiselta, lineaariselta näyttävä järjestystasoisten muuttujien välinen riippuvuus voidaan yhtä asteikkoväliä pidentämällä muuttaa käyräviivaiseksi, joka saattaakin paremmin vastata todellisuutta. Tilasto-ohjelmalla piirretyssä korrelaatiodiagrammissa kaikki asteikkovälit ovat samanpituisia, ja jos niitä muutettaisiin erimittaisiksi, muutosten täytyisi perustua muuttujan arvoihin. Joka tapauksessa korrelaatiodiagrammi antaa suuntaa muuttujien välisestä riippuvuudesta. Aina muuttujien välinen yhteys ei tule selkeästi esille, mikä saattaa johtua useista päällekkäisistä pisteistä tai riippuvuuden luonteesta. Tällöin kannattaa harkita jotakin muuta tapaa kuvata muuttujien välistä riippuvuutta. Joskus korrelaatiodiagrammi paljastaa mielenkiintoisesti muuttujien välisen riippuvuuden. Vaikka korrelaatiokertoimen arvo on likipitäen nolla, saattaa korrelaatiodiagrammista paljastua selkeä riippuvuus, joka on esimerkiksi alas- tai ylöspäin aukeavan paraabelin muotoista.



Kuvio 8. Todellisen ja ihanteellisen lasten lukumäärän riippuvuutta on kuvattu sekä korrelaatiodiagrammilla että 100 %:n pylväskuviolla. Lukijan tehtäväksi jää arvioida kuvioiden sopivuutta ja informatiivisuutta.

Laatikko-jana -kuvio

Laatikko-jana -kuvio on hyvin havainnollinen esitystapa tarkasteltaessa muuttujan jakauman sijaintia ja hajontaa. Se perustuu järjestysasteikon tasoisiin tunnuslukuihin ja sopii erityisesti silloin, kun muuttuja saa paljon eri arvoja. Esimerkiksi asenneväittämistä muodostettu summamuuttuja voi olla tällainen.

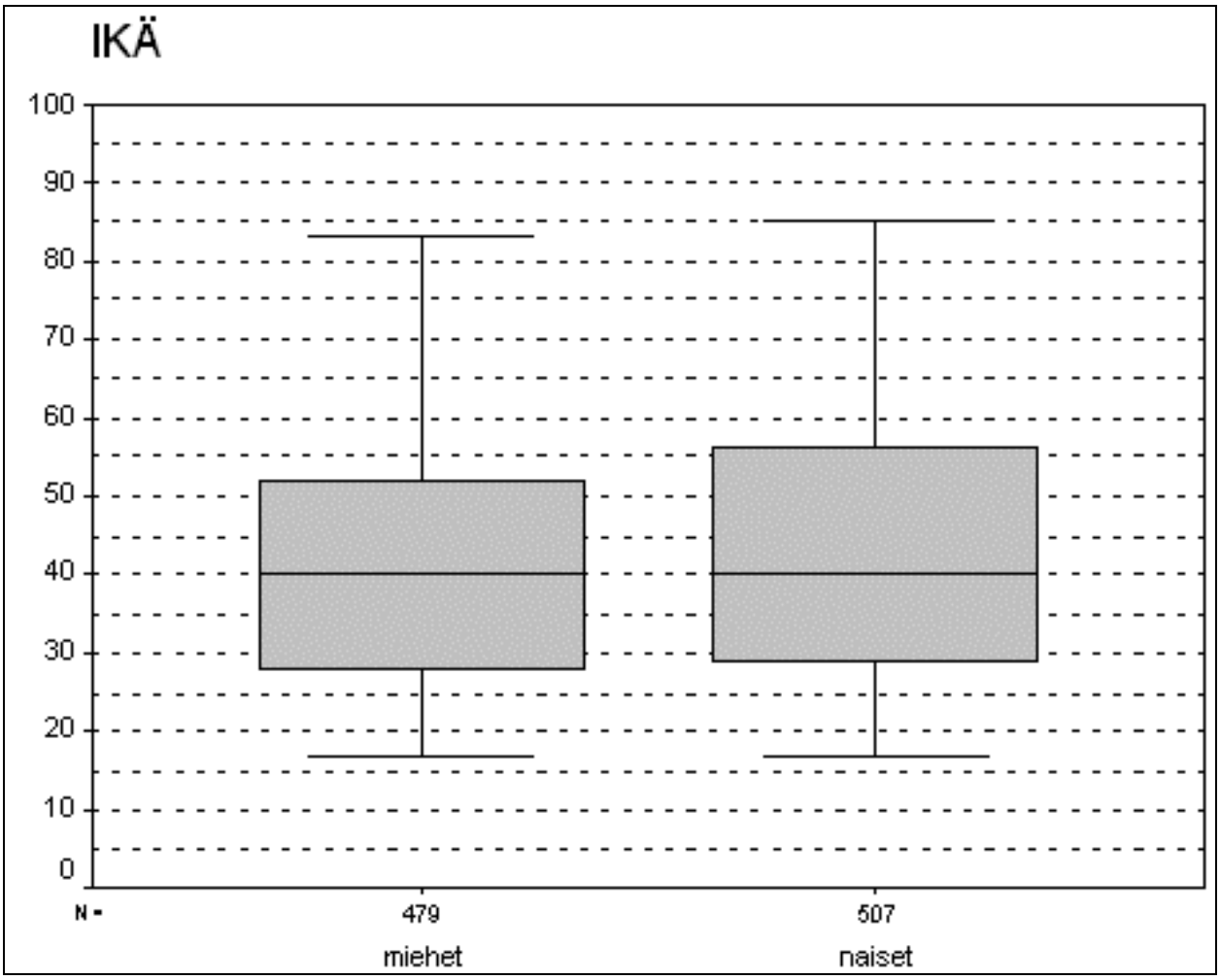
Laatikko-jana -kuvio sopii erityisesti jakaumien vertailuun. Vertailu tapahtuu toisen muuttujan ryhmissä, esim. asenteita tarkastellaan sukupuolittain. Ryhmitteleviä muuttujia voi olla kaksikin, jolloin voidaan tarkastella yhdysvaikutusta. Myös erillisten muuttujien kuvaaminen vierekkäisillä laatikko-janoilla on mahdollista. Tällöin helpottuu samaan ilmiöön liittyvien muuttujien jakaumien keskinäinen vertaaminen. Muutoksen tarkastelu esim. paneelitutkimuksissa on laatikko-jana -kuvioilla helppoa: samaa asiaa eri ajankohtina mittaavista muuttujista tehdään vierekkäiset laatikko-janat.

Kuviossa 9 tarkastellaan laatikko-jana -kuviolla naisten ja miesten ikäjakaumia. Laatikko-janat ovat lähes identtiset, mikä kertoo, että naisten ja miesten ikäjakaumissa ei ole suurta eroa. Tämän tuttuihin muuttujiin liittyvän esimerkkikuvion avulla perehdytään laatikko-jana -kuvioon.

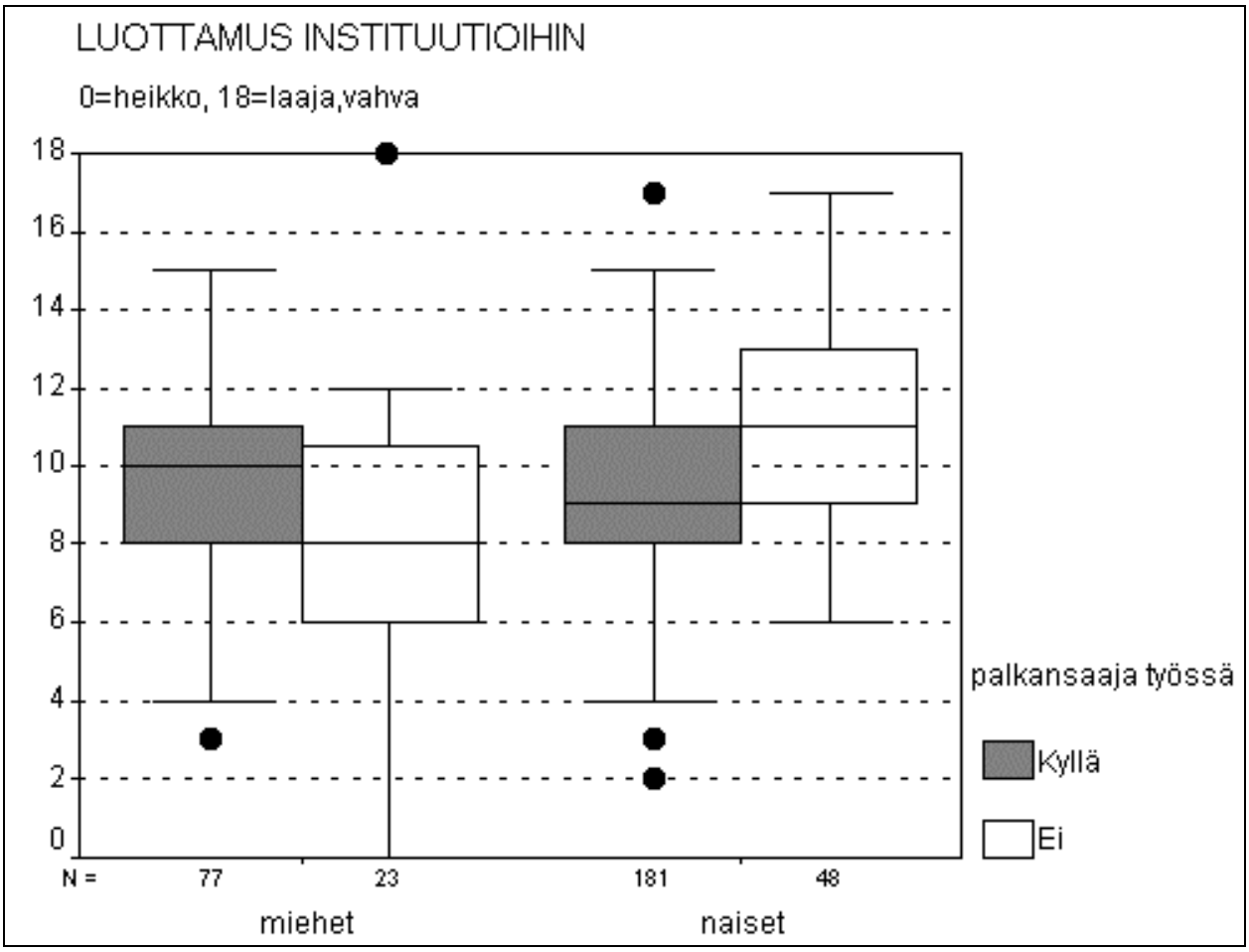
Laatikko-jana -kuvio perustuu tunnuslukuihin, jotka jakavat tarkasteltavan ryhmän neljään yhtä suureen joukkoon. Tunnusluvut on siten minimi, alakvartiili, mediaani, yläkvartiili ja maksimi. Kuvion avulla voidaan ensin hahmottaa hajontaa minimien ja maksimien vertailulla. Yksinkertaisimmillaan laatikko-jana -kuviossa janojen päät kertovat minimin ja maksimin. Tilasto-ohjelma voi merkitä todelliset minimit ja maksimit erityismerkillä, kun arvo poikkeaa muista oleellisesti. Kuviosta näkee, että sekä miehistä että naisista nuorin on 17-vuotias, vanhin mies on 83 ja nainen 85 vuotta.

Ikäjakauman sijainnista kertoo tiivistetysti mediaani, joka on merkitty viivalla ja sijaitsee yleensä laatikon sisällä. Joskus se on sama kuin ala- tai yläkvartiili. Sekä miesten että naisten keski-ikä mediaanilla ilmoitettuna on 40 vuotta, ts. vähintään puolet miehistä ja naisista on alle 41-vuotiaita.

Alakvartiili-ikä on 28 vuotta eli vähintään 25 % miehistä on 28-vuotiaita tai nuorempia. Vastaavasti naisista on vähintään neljännes 29-vuotiaita tai nuorempia. Yläkvartiili miehillä on 52 vuotta ja naisilla 56 vuotta. Yleistäen naiset ovat siis hiukan vanhempia kuin miehet. Se näkyy myös siinä, että "keskimmäiset" 50 % eli kuvion laatikko-osuus on naisilla hiukan ylempänä ja hiukan korkeampi. Ala- ja yläkvartiilin rajoittama laatikko kertoo sekä jakauman sijainnista että hajonnasta.



Kuvio 9. Naisten ja miesten ikäjakaumat laatikko-jana -kuviolla esitettynä.



Kuvio 10. Luottamus instituutioihin.

Laatikko-jana -kuvioon voidaan ottaa kaksi taustamuuttujaa, joiden muodostamissa typologioissa tarkastellaan kolmannen muuttujan jakaumaa. Tässä esimerkissä taustamuuttujiksi on valittu sukupuoli ja se, onko perheen pääasiallinen palkansaaja työssä vai työtön. Näiden muuttujien muodostamissa ryhmissä tarkastellaan luottamusta julkisen vallan instituutioihin. Kuvion perusteella näyttää siltä, että naiset, joiden perheessä pääasiallinen palkansaaja on työttömänä, eivät ole menettäneet luottamustaan, mutta miehillä tilanne on toinen.

Jos julkaistavan laatikko-jana -kuvion oletetaan olevan lukijakunnalle outo, on ensimmäisen kuvion yhteydessä syytä kirjoittaa alaviite, jossa kerrotaan kuvion tulkinnasta yleisesti.

Lähteet:

- Kuviot on tehty SPSS-ohjelmalla käyttäen Suomen Gallupin kokoamaa World Value Survey 1996 -aineistoa.

Lisätiedot (linkit, kirjallisuusviitteet)

Tutkimusprosessi

Hyvä suomenkielinen lähde tieteenfilosofian laajaan ongelmakenttään on Ilkka Niiniluodon kaksiosainen perusteos, joka käsittelee mm. tieteellisten teorioiden luonnetta, tieteellisen päättelyn eri tapoja sekä tieteellisen selittämisen luonnetta.

- Niiniluoto, Ilkka (1999, alkup. 1980): Johdatus tieteenfilosofiaan. Käsitteen- ja teorianmuodostus. Otava, Keuruu.
- Niiniluoto, Ilkka (1983): Tieteellinen päättely ja selittäminen. Otava, Keuruu.

Edellä mainitut Niiniluodon teokset käsittelevät tieteenfilosofiaa yleisesti. Yhteiskuntatieteelliseen selittämiseen liittyvää keskustelua on viime aikoina Suomessa käyty erityisesti sosiologien piirissä. Ainakin seuraavat kirjat käsittelevät yhteiskuntatieteelliseen selittämiseen liittyviä asioita:

- Alasuutari, Pertti (1999): Laadullinen tutkimus. Kolmas uudistettu painos. Vastapaino, Tampere.
- Alkula, Tapani & Pöntinen, Seppo & Ylöstalo, Pekka (1994): Sosiaalitutkimuksen kvantitatiiviset menetelmät. WSOY, Juva.
- Raunio, Kyösti (1999): Positivismi ja ihmistiede. Sosiaalitutkimuksen perustat ja käytännöt. Gaudeamus, Tampere.
- Toivonen, Timo (1999): Empiirinen sosiaalitutkimus: filosofia ja metodologia. WSOY, Porvoo.
- Töttö, Pertti (2000): Pirullisen positivismin paluu. Laadullisen ja määrällisen tarkastelua. Vastapaino, Tampere.

Englannin kielellä hyvä käytännöllinen esittely aloittelijalle yhteiskuntatieteellisen (erityisesti määrällisen) tutkimusprosessin vaiheista löytyy kirjasta:

- De Vaus, D.A. (1994): Surveys in Social Research. Third edition. UCL Press, Guildford.

Hieman yleisemmällä tasolla keskeisistä yhteiskuntatieteellisen tutkimuksen periaatteista sekä laadullisen ja määrällisen tutkimuksen erovaisuuksista ja samankaltaisuuksista kannattaa lukea kirjasta:

- King, Gary & Keohane, Robert O. & Verba, Sidney (1994): Designing Social Inquiry. Scientific Inference in Qualitative Research. Princeton University Press, Princeton.

Verkosta löytyy myös hyvä tieteenfilosofian bibliografia osoitteessa:

- http://www.herts.ac.uk/humanities/philosophy/scibib.html

Tutkimusasetelma

Hyvä aloittelijalle sopiva yleiskatsaus erilaisiin yhteiskuntatieteiden käyttämiin tutkimusasetelmiin löytyy kirjasta:

- De Vaus, D.A. (1994): Surveys in Social Research. Third edition. UCL Press, Guildford.

Yleisesti erilaisiin tutkimusasetelmiin liittyviä teoksia ovat esimerkiksi:

- Cook, Thomas D. & Cambell, Donald T. (1979): Quasi-Experimentation. Design & Analysis Issues for Field Settings. Houghton Mifflin Company, Boston.
- Spector, Paul E. (1981): Research Designs. Sage, Beverly Hills.
- Brown, Steven R. & Melamed, Lawrence E. (1990): Experimental Design and Analysis. Sage, Beverly Hills.

Hyvä johdatus yhteiskuntatieteellisiin muutoksen analyysiin soveltuviin asetelmiin ja menetelmiin on

- Dale, Angela & Davies, Richard B. (1994): Analyzing Social & Political Change. A Casebook of Methods. Sage, Lontoo.

Tapaustutkimusasetelmasta löytyy lisätietoja Yinin kirjasta:

- Yin, Robert K. (1990): Case Study Research. Design and Methods. Revised Edition. Sage, Newbury Park.

Verkossa lisätietoa tutkimusasetelmista löytyy mm. G. David Garsonin "Statnotes: an online textbook" –sivustoilta (valitse kohta "Research Designs") osoitteesta:

- http://www2.chass.ncsu.edu/garson/pa765/statnote.htm
- sivuston tutkimusasetelmia koskeva aineisto on sivulla: http://www2.chass.ncsu.edu/garson/pa765/design.htm

Mittaminen

- Esimerkkitapauksen lähde: Aikio Marjut, Saamelaiset kielenvaihtdon kierteessä. Kielisosiologinen tutkimus viiden saamelaiskylän kielenvaihdosta 1910-1980. Suomalaisen kirjallisuuden seuran toimituksia 479. Helsinki. Mäntän Kirjapaino Oy - Mänttä 1988. ISBN 951-717-476-4 ISSN 0355-1768

Kappaleen "Mittareiden luotettavuus" lähteet ja lisäluettevaa:

- Alkula Tapani, Pöntinen Seppo, Ylöstalo Pekka: Sosiaalitutkimuksen kvantitatiiviset menetelmät. WSOY:n graafiset laitokset, Juva 1995. ISBN 951-0-19286-4

Operationalisointi

Lähteet:

- Vilkuna Ilpo: Sosiaalisen taidon metsästys - Lähtökohtia sosiaalisen kyvykkyyden ymmärtämiseen. Artikkeli kirjassa Sosiaalinen vuorovaikutus. Toimittaneet Anja Riitta Lahikainen ja Anna-Maija Pirttilä-Vackman. Otavan kirjapaino, Keuruu 1998. ISBN 951-1-15790-6
- Liiman Raigo: Virolaisten ja venäläisten uskonnollisuus ja etnisyys. Empiirinen tutkimus Viron uskonnollisuudesta ja Virossa asuvien venäläisten etnisestä identiteetistä 1990-luvulla. Yleisen käytännöllisen teologian pro gradu -tutkielma, Teologinen tiedekunta, huhtikuu 2000.

Luettavaa:

- Kajamies Anu: Mitä on älykkyys? Artikkeli Psykologia-lehdessä 04/2000. Suomen psykologisen seuran julkaisu. 35. vuosikerta. ISSN 0355-1067
- Rose David, Sullivan Oriel: Introducing Data analysis for Social Scientists. Second Edition. Printed in Great Britain by Redwood Books, Trowbridge. 1996. ISBN 0-335-19617-9

Validiteetti

Lähteet:

- Procter, Michael, Measuring attitudes, luku kirjassa 'Researching social Life'. Edited by Gilbert, Nigel. Printed in Great Britain by The Cromwell Press, Trowbridge, Wiltshire 1998. First published 1993. ISBN 0-8039-8682-3
- Nummenmaa Tapio, Konttinen Raimo, Kuusinen Jorma, Leskinen Esko: Tutkimusaineiston analyysi. WSOY Kirjapainoyksikkö, Porvoo 1997. ISBN 951-0-21369-1

Luettavaa:

- Ronkainen Suvi, Äärimmäisen ihanaa vai suhteellisen mukavaa: ääri- ja keskirekisterin käyttö. Luku s.157 kirjassa Ajan ja paikan merkitsemät; subjektiviteetti, tieto ja toimijuus. Gaudeamus. Oy yliopistokustannus University Press Finland Ltd / Gaudeamus. Tammer-Paino Oy, Tampere 1999. ISBN 951-662-761-7

Reliabiliteetti

Lähteet:

- Procter, Michael, Measuring attitudes, luku kirjassa 'Researching social Life'. Edited by Gilbert, Nigel. Printed in Great Britain by The Cromwell Press, Trowbridge, Wiltshire 1998. First published 1993. ISBN 0-8039-8682-3
- SPSS Base 9.0 Application Guide 1999. Printed in the United States of America. ISBN 0-13-020401-3.
- Wright, Sonia R.: Quantitative Methods and Statistics - A Guide to Social Research. Sage Publications Beverly Hills London 1979. ISBN 0-8039-1294-3.

Otantamenetelmät

Yhteiskuntatieteelliseltä kannalta otantamenetelmiä sekä niihin liittyviä mahdollisuuksia ja ongelmia käsitellään muun muassa seuraavissa kirjoissa. Otantamenetelmien lisäksi De Vausin kirjassa käsitellään myös sopivan otoskoon valintaan vaikuttavia tekijöitä.

- Alkula, Tapani & Pöntinen, Seppo & Ylöstalo, Pekka (1994): Sosiaalitutkimuksen kvantitatiiviset menetelmät. WSOY, Juva.
- De Vaus, D.A. (1994): Surveys in Social Research. Third edition. UCL Press, Guildford.

Tilastotieteelliseltä kannalta otantamenetelmiä ja –teoriaa käsitellään Pahkisen ja Lehtosen kirjassa:

- Pahkinen, Erkki & Lehtonen, Risto (1989): Otanta-asetelmat ja tilastollinen analyysi. Gaudeamus, Helsinki.

Soveltuvan kokoisen otoskoon määrääminen riippuu monista asioista. Yksi otoskoon valintaan vaikuttava tekijä on se, millä tarkkuudella saadut tuloksen halutaan yleistää koko perusjoukkoa koskeviksi (katso tilastollinen päättely). Verkosta löytyy useita laskureita, jotka voivat auttaa otoskoon määräämisessä. Laskureita löytyy muun muassa seuraavista osoitteista:

- http://ebook.stat.ucla.edu/calculators/sampsize.phtml
- http://www.researchinfo.com/docs/calculators/samplesize.cfm

Postikyselyaineiston kokoaminen

Kirjallisuus:

- Dillman, Don A.: Mail and Internet Surveys. The Tailored Design Method. (2. painos). Wiley & Sons, 2000.
- Lotti, Leila: Markkinointitutkimuksen käsikirja: Helsinki, WSOY, 1998.

Lomakkeen laatiminen

Kirjallisuutta

- Alkula, Tapani - Pöntinen, Seppo - Ylöstalo, Pekka: Sosiaalitutkimuksen kvantitatiiviset menetelmät. Helsinki: WSOY 1994.
- Jyrinki, Erkki: Kysely ja haastattelu tutkimuksessa (2. painos). Helsinki: Gaudeamus 1976.
- Uusitalo, Hannu: Tiede, tutkimus, tutkielma. Johdatus tutkielman maailmaan. WSOY, Juva 1991.
- Valkonen, Tapani: Haastattelu- ja kyselyaineiston analyysi sosiaalitutkimuksessa. Helsinki: Ylioppilastuki ry. 1971.

Linkkejä

American Association for Public Opinion Research tarjoaa sivustoillaan laaja-alaisesti linkkejä kyselytutkimuksen ja lomakesuunnittelun metodologiaan ja tutkimuseettisiin kysymyksiin:

- http://www.aapor.org/main.html

Yksittäisiä verkosta löytyviä linkkejä kyselylomakkeen suunnittelun tueksi:

- Research Methods Knowledge Base (William M. Trochim): http://trochim.human.cornell.edu/kb/contents.htm
- American Statistical Association/Survey Research: http://www.amstat.org/sections/srms/whatsurvey.html

Muuttujien muunnokset

Suomeksi lisätietoja muuttujien koodauksesta ja muunnoksista voi katsoa Alkulan ym. teoksesta:

- Alkula, Tapani & Pöntinen, Seppo & Ylöstalo, Pekka (1994): Sosiaalitutkimuksen kvantitatiiviset menetelmät. WSOY, Juva.

Englanniksi kannattaa katsoa:

- De Vaus, D.A. (1994): Surveys in Social Research. Third edition. UCL Press, Guildford.

Joskus muuttujien jakauma voi olla sellainen, että analyysin parantamiseksi sen muunnos jollain tavalla on tarpeellinen. Muunnostapoja on erilaisia riippuen käytetyn menetelmän vaatimuksista ja teoreettisista olettamuksista. Tällaisista muunnoksista voi lukea lisää seuraavasta kirjasta:

- Tabachnick, Barbara G. & Fidell, Linda S. (1996): Using Multivariate Statistics. HarperCollins, New York.

Summamuuttujat

Kirjallisuutta:

- Alkula Tapani; Pöntinen Seppo, Ylöstalo Pekka: Sosiaalitutkimuksen kvantitatiiviset menetelmät. WSOY:n graafiset laitokset, Juva 1995. ISBN 951-0-19286-4

Puuttuvat havainnot

Puuttuvien havaintojen aiheuttamia ongelmia ja näiden ongelmien ratkaisuyrityksiä käsitellään määrällisten menetelmien perusoppikirjoissa yllättävän vähän. De Vausin kirjassa aihetta käsitellään jonkin verran, perustuen kuitenkin lähinnä Hertelin artikkeliin.

- De Vaus, D.A. (1994): Surveys in Social Research. Third edition. UCL Press, Guildford.
- Hertel, Bradley R. (1976): Minimizing Error Variance Introduced by Missing Data Routines in Survey Analysis. Sociological Methods & Research 4: 459-474.

Havaintojen painottaminen

Lisäinformaatiota kyselytutkimuksen painotusmenetelmistä:

- Holt, D. and Smith, T. M. F. (1979): Post Stratification. Journal of the Royal Statistical Society.
- Journal of the American Statistical Association (verkkoversion lisätiedot-osiossa on suorat linkit seuraaviin PDF-muotoisiin dokumentteihin):
    - Deville & Särndal (1992): Calibration Estimators in Survey Sampling.
    - Little, R.J.A. (1993): Post-stratification: A modeler's perspective.
    - Zieschang, Kimberly D. (1990): Sample Weighting Methods and Estimation of Totals in the Consumer Expenditure Survey

Tilastollinen päättely

Suomeksi tilastollista päättelyä on käsitelty mm. Nummenmaan ym. kirjassa:

- Nummenmaa, Tapio & Konttinen, Raimo & Kuusinen, Jorma & Leskinen, Esko (1996): Tutkimusaineiston analyysi. WSOY, Porvoo.

Tilastollisen päättelyn periaatteet löytyvät useimmista tilastotieteen perusoppikirjoista. Suomenkielellä katso esimerkiksi:

- Vasama, Pyry-Matti & Vartia, Yrjö (1980): Johdatus tilastotieteeseen I. Neljäs korjattu painos. Gaudeamus, Pori.

Englanninkielellä tilastollisen päättelyn perusteita voi opiskella esimerkiksi seuraavista kirjoista:

- Bohrnstedt, George W. & Knoke, David (1988): Statistics for Social Data Analysis. Toinen pianos. F.E. Peacock Publishers, Itasca.
- Cohen, Louis & Holliday, Michael (1996): Practical Statistics for Students. Paul Chapman Publishing, Lontoo.
- Moore, David S. (1995): The Basic Practice of Statistics. W.H. Freeman and Company, New York.

Suomenkielellä verkosta löytyy ns. "Internetix-oppimisympäristöstä" kaksi tilastotieteen peruskurssia. Molemmat perustuvat Simo Kivelän materiaaliin. "Tilastot ja todennäköisyys" –kurssi löytyy osoitteesta:

- http://www.internetix.ofw.fi/opinnot/opintojaksot/5luonnontieteet/matematiikkal/mb3/

ja "Tilastotiedettä ja todennäköisyyslaskentaa" –kurssi osoitteesta:

- http://www.internetix.ofw.fi/opinnot/opintojaksot/5luonnontieteet/matematiikka/tilastot/index.htm

Englanninkielistä lisätietoa tilastollisesta päättelystä löytyy mm. Hyperstat Online -palvelusta, jonka osoite on:

- http://davidmlane.com/hyperstat/index.html

Toinen hyvä verkkoresurssi on Gene V. Glassiin pitämän "Intro to Quant Methods" –kurssin sivut osoitteessa (valitse kohta "Lesson six: Sampling and Statistical Inference"):

- http://glass.ed.asu.edu/stats/

Keskiluvut

Keskiluvut on esitelty kaikissa tilastotieteiden ja kvantitatiivisten menetelmien perusoppaissa. Hyvä suomenkielinen opastus keskilukuihin on esimerkiksi:

- Heikkilä, Juha (1993): Tilastotieteen ABC-kirja 1. Yliopistopaino, Jyväskylä.

Englannin kielellä keskiluvuista ja niiden sovelluksista yhteiskuntatieteellisessä tutkimuksessa voi lukea esimerkiksi seuraavista teoksista:

- Jones, Laurence F. & Olson, Edward C. (1996): Political Science Research. A Handbook of Scope and Method. Addison Wesley Longman, New York.
- De Vaus, D.A. (1994): Surveys in Social Research. Third edition. UCL Press, Guildford.

Verkosta lisätietoa keskiluvuista löytyy mm. Hyperstat Online palvelusta. Siellä kerrotaan mm. sellaisista jakauman tunnusluvuista, joita ei tässä yhteyssä käsitelty. Hyperstat Onlinen osoite on:

- http://davidmlane.com/hyperstat/index.html
- ja keskiluvuista kerrotaan erityisesti sivulla:
  http://davidmlane.com/hyperstat/desc_univ.html

Toinen hyvä verkkoresurssi on Gene V. Glassiin pitämän "Intro to Quant Methods" -kurssin sivut osoitteessa:

- http://glass.ed.asu.edu/stats/

jakauman tunnuslukuja käsitellään erityisesti sivulla:

- http://glass.ed.asu.edu/stats/lesson2/

Hajontaluvut

Keskeisimmät hajontaluvut on esitelty kaikissa tilastotieteiden ja kvantitatiivisten menetelmien perusoppaissa. Hyvä suomenkielinen opastus on esimerkiksi:

- Heikkilä, Juha (1993): Tilastotieteen ABC-kirja 1. Yliopistopaino, Jyväskylä.

Englannin kielellä hajontaluvuista ja niiden sovelluksista yhteiskuntatieteellisessä tutkimuksessa voi lukea esimerkiksi seuraavista teoksista:

- Jones, Laurence F. & Olson, Edward C. (1996): Political Science Research. A Handbook of Scope and Method. Addison Wesley Longman, New York.
- De Vaus, D.A. (1994): Surveys in Social Research. Third edition. UCL Press, Guildford.

Verkosta lisätietoa keskiluvuista löytyy mm. Hyperstat Online -palvelusta. Siellä kerrotaan mm. sellaisista jakauman tunnusluvuista, joita ei tässä yhteydessä käsitelty. Hyperstat Onlinen osoite on:

- http://davidmlane.com/hyperstat/index.html
- Hajontaluvuista kerrotaan erityisesti sivulla: http://davidmlane.com/hyperstat/desc_univ.html

Toinen hyvä verkkoresurssi on Gene V. Glassiin pitämän "Intro to Quant Methods" -kurssin sivut osoitteessa:

- http://glass.ed.asu.edu/stats/
- Jakauman tunnuslukuja käsitellään erityisesti sivulla: http://glass.ed.asu.edu/stats/lesson2/

Ristiintaulukointi

Suomeksi ristiintaulukoista voi lukea lisää esimerkiksi kirjasta

- Alkula, Tapani & Pöntinen, Seppo & Ylöstalo, Pekka (1994): Sosiaalitutkimuksen kvantitatiiviset menetelmät. WSOY, Juva.

Englanninkielellä tietoja ristiintaulukoinnista löytyy lähes jokaisesta yhteiskuntatieteellisestä kvantitatiivisten menetelmien oppaasta. Seuraavassa muutama hyvä esimerkki:

- Bohrnstedt, George W. & Knoke, David (1988): Statistics for Social Data Analysis. F.E. Peacock, Itasca.
- Moore, David S. (1995): The Basic Practice of Statistics. W.H. Freeman & co, New York.
- De Vaus, D.A. (1994): Surveys in Social Research. Third edition. UCL Press, Guildford.

Verkosta löytyy lisätietoja ristiintaulukoinnista esimerkiksi Marion Joppen "The Research Process"-sivustosta valitsemalla sieltä kohdat "Cross tabulations" ja "Calculating the chi-square". Sivuston osoite on:

- http://www.ryerson.ca/~mjoppe/research/index.html

Toinen hyvä verkkolähde on "Statistics resource center", josta valitsemalla kohdan "Cross tabulations" saa lisätietoja ristiintaulukoista. Osoite on:

- http://www.millsaps.edu/www/socio/statsresources.htm

Korrelaatio

Kirjallisuutta:

- Alkula, Pöntinen, Ylöstalo (1994). Sosiaalitutkimuksen kvantitatiiviset menetelmät. Helsinki: WSOY.
- Vasama, Vartia (1980). Johdatus Tilastotieteeseen. Helsinki: Gaudeamus.
- Manninen, Pentti (1996). Johdatus tilastolliseen data-analyysiin sovellus- ja atk-keskeinen näkökulma. Tampere: Tampereen yliopisto.
- Agresti, Alan (1996). Introduction to categorical data analysis. NY: John Wiley and Sons.
- Liebetrau, Albert M. (1983). Measures of association. Newbury Park, CA: Sage Publications. Quantitative Applications in the Social Sciences Series No. 32.
- Cohen, Jacob and Patricia Cohen (1983). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, Second Edition. Hillsdale, NJ: Lawrence Erlbaum Assoc; ISBN: 0898592682.
- Kendall, Maurice and Jean Dickinson Gibbons (1990). Rank Correlation Methods, Fifth Edition. NY: Oxford Univ Press; ISBN: 0195208374.
- Blalock, Hubert. (1961). Causal inferences in nonexperimental research. Chapel Hill, NC: UNC Press.
- Davis, James A. (1985). The logic of causal order. Quantitative applications in the social sciences series, no. 55. Thousand Oaks, CA: Sage Publications. *Pp. 38 - 44 provide a non-technical introduction to partial correlation inferences.*

Verkossa:

Pertti Vuorisen SPSS-opas:

- http://www.evamk.fi/koti/~vuorinen/Tilasto/SPSSopas/korrelaa.htm

Pentti Roution Taideteollisessa korkeakoulussa kokoama tutkimusmetodiopas 'Taito-oppi' käsittelee mm. korrelaatioita:

- http://www.uiah.fi/projects/metodi/080.htm#korr

Tero Erkkilän metodi-kurssilta:

- http://www.valt.helsinki.fi/staff/terkkila/metodi1/esitys5.htm

Hypoteesien testaus

Englanninkielellä hypoteesien päättelyn perusteita voi opiskella esimerkiksi seuraavista kirjoista:

- Bohrnstedt, George W. & Knoke, David (1988): Statistics for Social Data Analysis. Toinen pianos. F.E. Peacock Publishers, Itasca.
- Cohen, Louis & Holliday, Michael (1996): Practical Statistics for Students. Paul Chapman Publishing, Lontoo.
- Moore, David S. (1995): The Basic Practice of Statistics. W.H. Freeman and Company, New York.

Kanjin teos on hyvä käsikirja erilaisiin tilastollisiin testeihin. Se ei kuitenkaan välttämättä sovi aloittelijan tarpeisiin:

- Kanji, Gopal K. (1999): 100 Statistical Tests. Sage, London.

Keskustelua tilastollisten testien mielekkyydestä ja hyödyllisyydestä yhteiskuntatieteissä löytyy mm. seuraavista teoksista:

- Henkel, Ramon E. (1976): Tests of Significance. Sage, Beverly Hills.
- Mäkelä, Jukka (1991): Sunnuntaina sataa aina - tutkimus tilastollisen ajattelun siirtymisestä osaksi empiiristä sosiaalitutkimusta. Lapin yliopiston yhteiskuntatieteellisiä julkaisuja 13.

Verkosta lisätietoa tilastollisesta päättelystä löytyy mm. Hyperstat Online -palvelun kohdasta "The Logic of Hypothesis Testing". Osoite on:

- http://davidmlane.com/hyperstat/logic_hypothesis.html

Toinen hyvä verkkoresurssi on Valerie J. Eastonin ja John H. McCollin "Statistics Glossary" ja sen alakohta "Hypothesis Testing" osoitteessa:

- http://www.cas.lancs.ac.uk/glossary_v1.1/hyptest.html

Bill L. Thompson on kerännyt verkkoon laajan kokoelman lähdeviitteitä, joissa kritisoidaan hypoteesien testauksen menetelmää. Osoite on:

- http://www.cnr.colostate.edu/~anderson/thompson1.html

Koska Thompson haluaa olla tasapuolinen ja esittää myös toisen "kiistapuolen" näkemyksen, on hän myös kerännyt viitelistan artikkeleihin ja kirjoihin, joissa tähän kritiikkiin vastataan. Osoite on:

- http://www.cnr.colostate.edu/~anderson/thompson2.html

Varianssianalyysi

Suomeksi varianssianalyysin perusteista voi lukea lisää esimerkiksi seuraavista teoksista:

- Alkula, Tapani & Pöntinen, Seppo & Ylöstalo, Pekka (1994): Sosiaalitutkimuksen kvantitatiiviset menetelmät. WSOY, Juva.
- Toivonen, Timo (1999): Empiirinen sosiaalitutkimus: filosofia ja metodologia. WSOY, Porvoo.

Laajemmin varianssi-ja kovarianssianalyysiin sekä MANOVAan voi tutustua esimerkiksi seuraavien kirjojen avulla:

- Bray, James H. & Maxwell, Scott E. (1985): Multivariate Analysis of Variance. Sage, Beverly Hills.
- Iversen, Gudmund R. & Norpoth, Helmut (1987): Analysis of Variance. Sage Newbury Park.
- Tabachnick, Barbara G. & Fidell, Linda S. (1996): Using Multivariate Statistics. Harper Collins, New York.
- Wildt, Albert R. & Ahtola, Olli T. (1978): Analysis of Covariance. Sage, Beverly Hills.

Verkosta löytyy suhteellisen paljon varianssianalyysia ja sen laajennuksia käsittelevää materiaalia. Varianssianalyysin perusteita käsitellään mm. seuraavilla sivuilla:

- http://www2.chass.ncsu.edu/garson/pa765/anova.htm
- http://davidmlane.com/hyperstat/intro_ANOVA.html
- http://www.psychstat.smsu.edu/introbook/SBK27.htm

Kovarianssianalyysista lisätietoa löytyy seuraavilta sivuilta:

- http://www.basic.nwu.edu/statguidefiles/ancova.html
- http://www.cogs.susx.ac.uk/users/andyf/teaching/rm2/ancova_files/frame.htm

Erityisesti MANOVAa käsitteleviä sivustoja ovat:

- http://www.statsoftinc.com/textbook/stanman.html
- http://www.richmond.edu/~pli/psy538/MANOVA/manova1%20copy/
- http://www.richmond.edu/~pli/psy538/MANOVA/index.html

Regressioanalyysi

Yhteiskuntatieteellisten tutkimusalojen opiskelijat ja tutkijat voivat perehtyä suomeksi regressioanalyysin perusteisiin muun muassa seuraavissa kirjoissa:

- Alkula, Tapani & Pöntinen, Seppo & Ylöstalo, Pekka (1994): Sosiaalitutkimuksen kvantitatiiviset menetelmät. WSOY, Juva.
- Karma, Kai & Komulainen, Erkki (1992): Käyttäytymistieteiden tilastomenetelmien jatkokurssi. Yliopistopaino, Helsinki.

- Nummenmaa, Tapio & Konttinen, Raimo & Kuusinen, Jorma & Leskinen, Esko (1996): Tutkimusaineiston analyysi. WSOY, Porvoo.
- Toivonen, Timo (1999): Empiirinen sosiaalitutkimus: filosofia ja metodologia. WSOY, Porvoo.

Englanniksi regressioanalyysin perusteista voi lukea mm. seuraavista kirjoista. Näistä De Vausin kirja sisältää vain regressioanalyysin perusteet, mutta toisaalta se on vasta-alkajalle erittäin helppolukuinen. Tabachnickin ja Fidellin kirjassa on huomattavasti kattavampi regressioanalyysin esittely.

- De Vaus, D.A. (1994): Surveys in Social Research. Third edition. UCL Press, Guildford.
- Tabachnick, Barbara G. & Fidell, Linda S. (1996): Using Multivariate Statistics. Harper Collins, New York.

Sagen julkaisemassa määrällisten menetelmien opassarjassa on useita selkeitä regressioanalyysikirjoja. Näistä Lewis-Beckin kirja on helppolukuisin.

- Berry, William D. (1993): Understanding Regression Assumptions. Sage, Newbury Park.
- Berry, William D. & Feldman, Stanley (1985): Multiple Regression in Practice. Sage, Beverly Hills.
- Fox, John (1991): Regression Diagnostics. Sage, Newbury Park.
- Hardy, Melissa A. (1993): Regression with Dummy Variables. Sage, Newbury Park.
- Lewis-Beck, Michael S. (1980): Applied Regression: An Introduction. Sage, Beverly Hills.
- Ostrom, Charles W. jr (1990): Time Series Analysis. Sage, Beverly Hills.
- Sayrs, Lois W. (1989): Pooled Time Series Analysis. Sage, Newbury Park.

Tilastotieteelliseltä kannalta regressioanalyysia käsitellään seuraavissa teoksissa:

- Bohrnstedt, George W. & Knoke, David (1988): Statistics for Social Data Analysis. Toinen painos. F.E. Peacock Publishers, Itasca.
- Moore, David S. (1995): The Basic Practice of Statistics. W.H. Freeman and Company, New York.
- Moore, David S. & McCabe, George P. (1999): Introduction to the Practice of Statistics. W.H. Freeman and Company, New York.

Kaikkein kattavimmin regressioanalyysista kerrotaan kansantaloustieteen ekonometrian oppikirjoissa. Kirjat voivat pikaisen silmäyksen perusteella vaikuttaa vaikeilta. Niihin kannattaa silti tutustua, jos haluaa oppia syvällisesti erilaisista regressioanalyysin käyttömahdollisuuksista. Verrattain helppolukuisia, mutta siitä huolimatta kattavia ekonometrian oppikirjoja ovat mm.:

- Gujarati, Damodar N. (1988): Basic Econometrics. McGraw-Hill, New York.
- Kennedy, Peter (1998): A Guide to Econometrics. MIT Press, Boston.
- Kmenta, Jan (1986): Elements of Econometrics. MacMillan, New York.
- Pindyck, Robert S. & Rubinfeld, Daniel L. (1997): Econometric Models and Economic Forecasts. Irwin, Boston.

Verkosta löytyy runsaasti regressioanalyysiin liittyvää materiaalia. Katso esimerkiksi David Garsonin Statnotes: an Online Textbook -sivujen regressioanalyysia käsittelevä osuus osoitteessa:

- http://www2.chass.ncsu.edu/garson/pa765/regress.htm

Myös "Statistics Resource Centre" -sivustolla käsitellään lyhyesti regressioanalyysia osoitteessa:

- http://www.millsaps.edu/www/socio/regression.htm

Seuraavasta osoitteesta löytyy pieni java-applet, jonka avulla voi interaktiivisesti testata regressioanalyysin perusperiaatteita:

- http://www.stattucino.com/berrie/dsl/regression/regression.html

TV:stä tutut animaatiohahmot Ren ja Stimpy opettavat hauskasti regressioanalyysin perusteita osoitteessa:

- http://www-psych.nmsu.edu/regression/home.html

Faktorianalyysi

Faktorianalyysista on saatavilla runsaasti suomenkielistä materiaalia. Seuraavassa listassa Nummenmaan ym. kirja käsittelee faktorianalyysia kaikkein perusteellisimmin. Lisäksi siinä esitellään myös konfirmatorista faktorianalyysia ja rakenneyhtälömalleja.

- Alkula, Tapani & Pöntinen, Seppo & Ylöstalo, Pekka (1994): Sosiaalitutkimuksen kvantitatiiviset menetelmät. WSOY, Juva.
- Karma, Kai & Komulainen, Erkki (1992): Käyttäytymistieteiden tilastomenetelmien jatkokurssi. Yliopistopaino, Helsinki.
- Nummenmaa, Tapio & Konttinen, Raimo & Kuusinen, Jorma & Leskinen, Esko (1996): Tutkimusaineiston analyysi. WSOY, Porvoo.
- Toivonen, Timo (1999): Empiirinen sosiaalitutkimus: filosofia ja metodologia. WSOY, Porvoo.

Englanniksi faktorianalyysia esitellään hyvin seuraavissa kirjoissa. De Vausin kirja sopii parhaiten aloittelijalle. Tabachnickin ja Fidellin kirja menee esittelyssään syvemmälle. Siinä käsitellään sekä eksploratiivista että konfirmatorista faktorianalyysia.

- De Vaus, D.A. (1994): Surveys in Social Research. Third edition. UCL Press, Guildford.

- Tabachnick, Barbara G. & Fidell, Linda S. (1996): Using Multivariate Statistics. Harper Collins, New York.

Verkosta löytyy paljon faktorianalyysia käsittelevää materiaalia. Richard B. Darlington Cornellin yliopistosta on kirjoittanut hyvän aloittelijoille sopivan esittelyn faktorianalyysista. Se löytyy osoitteesta:

- http://comp9.psych.cornell.edu/Darlington/factor.htm

Myös Colleen Flynn Thapalia on kirjoittanut lyhyen esittelyn faktorianalyysista. Se löytyy osoitteesta:

- http://trochim.human.cornell.edu/tutorial/flynn/factor.htm

Connie D. Stapleton on tuottanut hyvän esittelyn konfirmatorisesta faktorianalyysista. Osoite on:

- http://ericae.net/ft/tamu/Cfa.HTM

David Garsonin Statnotes: an Online Textbook on oivallinen lähde sekä faktorianalyysiin että rakenneyhtälömalleihin liittyvissä asioissa. Osoitteet ovat:

- Faktorianalyysi: http://www2.chass.ncsu.edu/garson/pa765/factor.htm
- Rakenneyhtälömallit: http://www2.chass.ncsu.edu/garson/pa765/structur.htm

Yleisesti konfirmatorisesta faktorianalyysista ja rakenneyhtälömalleista kerrotaan SEM FAQ sivuilla (SEM = Structural Equation Models) osoitteessa:

- http://www.gsu.edu/~mkteer/semfaq.html

Konfirmatorisessa faktorianalyysissa ja rakenneyhtälömalleissa yleisesti käytettyjä ohjelmia ovat:

- AMOS: http://www.smallwaters.com/
- EQS: http://www.mvsoft.com/
- LISCOMP: http://www.gsu.edu/~mkteer/liscomp.html
- LISREL: http://www.ssicentral.com/lisrel/mainlis.htm

Logistinen regressio

Suomen kielellä logistisesta regressioanalyysista ei toistaiseksi löydy kattavaa yleisesittelyä. Englanniksi menetelmää on käsitelty mm .Tabachnickin ja Fidellin kirjassa:

- Tabachnick, Barbara G. & Fidell, Linda S. (1996): Using Multivariate Statistics. Harper Collins, New York.

Sagen määrällisten menetelmien opassarjassa on useita logistista regressioanalyysia käsitteleviä kirjoja. Alla mainituista teoksista Liaon kirjassa käsitellään myös multinomiaalista logistista regressiota.

- Liao, Tim Futing (1994): Interpreting Probability Models. Logit, Probit, and Other Generalized Linear Models. Sage, Thousand Oaks.
- Menard, Scott (1995): Applied Logistic Regression Analysis. Sage, Thousand Oaks.
- Pampel, Fred C. (2000): Logistic Regression. A Primer. Sage, Thousand Oaks.

Verkossa David Garsonin Statnotes: an Online Textbook on hyvä lähde logistiseen regressiomalliin ja sen tulosten tulkintaan liittyvissä asioissa. Suora osoite on:

- http://www2.chass.ncsu.edu/garson/pa765/logistic.htm

Graafinen esitys

Kirjallisuutta:

- Kuusela, Vesa: Tilastografiikan perusteet. Oy Edita Ab. Helsinki 2000. ISBN 951-37-3116-2
- Hirsjärvi Sirkka, Liikanen Pirkko, Remes Pirkko, Sajavaara Paula: Tutkimus ja sen raportointi. Kirjayhtymä, Jyväskylä 1993.
- Hirsjärvi Sirkka, Remes Pirkko, Sajavaara Paula: Tutki ja kirjoita. Tammi. Helsinki 2001. 7. painos. ISBN 951-26-4618-8