

Bayesilaisen tilastoanalyysin käyttömahdolli- suudet ammatti- kasvatuksen tutkimuksessa

Petri Nokelainen

Dosentti, FT

Tampereen yliopisto, Ammattikasvatuksen tutkimus- ja koulutuskeskus

petri.nokelainen@uta.fi

Abstrakti

Pohdin artikkelin alussa tyypillisiä ongelmia, joihin kvantitatiivisia tutkimusmenetelmiä käyttävä tutkija törmää, tai joista hänen ainakin tulisi olla tietoinen. Seuraavaksi kuvaan bayesilaisen tilastollisen mallintamisen peruspiirteitä ja sen tarjoamia mahdollisuuksia kasvatustieteellisessä tutkimuksessa. Vertaan artikkelissa bayesilaista ja perinteistä frekventististä lähestymistapaa kvantitatiivisessa tutkimuksessa yleisesti käytössä olevaan nollahypoteesintestaukseen. Lopuksi esittelen kaksi bayesilaisen mallintamisen mahdollistavaa tietokonesovellusta.

Asiasanat: kvantitatiiviset tutkimusmenetelmät, bayesilainen mallinnus, lo-maketutkimus, tilastollinen merkitsevyys, nollahypoteesin testaus.

Abstract

This article discusses some basic assumptions of quantitative statistical analysis and how they may become problematic to educational researchers. A Bayesian way of doing statistical analysis is presented and compared to traditional frequentistic methods. Null hypothesis significance testing is discussed from both frequentistic and Bayesian viewpoints. Finally, I present two Bayesian statistical computer applications.

Descriptors: quantitative methods, bayesian modeling, survey research, statistical significance, null hypothesis testing.

.....

Johdanto

Kvalitatiivisen tutkimusmetodologian alati vahvistuvasta asemasta huolimatta osa ammattikasvatuksen kentän tutkimusaineistoista hankitaan yhä kvantitatiivisilla (ns. määrällisillä) menetelmillä, esimerkiksi paperi- tai tietokonekyselylomakkeen välityksellä. Motivaatio kvantitatiivisten menetelmien käyttöön juontaa juurensa vuosikymmenten taakse: Tarvitaan vertailuaineistoja aiemmin kerätyille aineistoille, halutaan tutkia tietyn ilmiön vaikuttavuutta mahdollisimman yleistämiskelpoisilla (generalizability) menetelmillä, tai kehitetään barometreja teoreettisilta taustoiltaan tunnettujen ilmiöiden jatkuvaan tarkkailuun.

Kasvatustieteellisen tutkimuksen kohteena olevista ilmiöistä merkittävä osa (esimerkiksi ammatillinen kasvu, organisaation työilmapiiri, koulukiusaaminen jne.) on perusluonteeltaan sellaisia, että ne eivät sovellu yksinomaan kvantitatiivisella tai kvalitatiivisella lähestymistavalla tutkittaviksi, vaan edellyttävät laajempaa, monimetodista lähestymistapaa (multi/mixed method).

Tällöin kvantitatiivinen ja kvalitatiivinen tutkimusote vuorottelevat tutkimusongelmien vaatimusten mukaisesti. On harhaanjohtavaa leimata yhtä tutkimusmenetelmää ”laadulliseksi” ja toista ”määrälliseksi” muista kuin käytännön syystä, jollaisesta esimerkkinä mainittakoon kvantitatiivisen (tai numeerisen empiirisen aineiston) analyysin kurssi (Töttö 2004).

Jotta monimetodinen lähestymistapa olisi käytännössä toteutettavissa, jokaisen kasvatustieteen opiskelijan menetelmälliseen koulutukseen tulee kuulua riittävästi opintoja kummankin lähestymistavan hallitsemiseen, ainakin peruseriaatteiden tasolla. Opiskelijoilla tulee olla valmius lukea ja ymmärtää tieteellisiä artikkeleita (Rautopuro & Malin 2008). Myöhemmin työelämään siirtyessään heillä tulee olla kyky valita oikeat ”työvälineet” mitä moninaisimpien tutkimusongelmien ratkaisuun (Silén 2008), puhumattakaan akateemisen uran valinneilla väistämättä eteen tulevasta opiskelijoiden menetelmällisestä ohjauksesta.

Seuraavaksi kuvaan kasvatustieteiden, erityisesti ammattikasvatuksen, tutkijoiden käyttämien tilastollisten työvälineiden keskeisiä ominaisuuksia sekä

ongelmia, jotka seuraavat, kun näitä työvälineitä käytetään monivalintakyselylomakkeella hankitun numeerisen empiirisen aineiston analyysiin.

Kvantitatiivisten menetelmien käytön haasteita kasvatustieteellisessä tutkimuksessa

Yleensä kasvatustieteiden kvantitatiivisella tutkimuksella tarkoitetaan monivalintaväittämiä sisältävällä kyselylomakkeella kerätyn empiirisen numeerisen aineiston analyysia. On totta, että numeerista aineistoa voidaan kerätä tai tuottaa myös muilla tavoin, esimerkiksi kategorioimalla haastatteluaineisto numeerisiin luokkiin tai keräämällä tietokonesovelluksen tuottamaa lokidataa (ks. esim. Kurhila, Miettinen, Nokelainen & Tirri 2007).

Epäjatkuvan (discrete) numeerisen tutkimusaineiston analyysi on jo vuosikymmeniä tunnustettu haastavaksi tehtäväksi laajalti käytössä oleville, ns. 'perinteisiksi' kutsutuille, parametrisille lineaarisille frekventistisille menetelmille. Tällaisia keskiarvon laskentaan perustuvia menetelmiä ovat esimerkiksi t-testi, regressioanalyysi ja eksploratiivinen faktorianalyysi (Marini, Li & Fan 1996). Ongelmat johtuvat pääasiassa siitä, että edellä mainitut analyysimenetelmät asetavat tutkimusaineistolle tiettyjä oletuksia.

Ensimmäisenä oletuksena on se, että tutkittavan ilmiön ja sitä kuvaavan otoksen tulisi olla jatkuvia (continuous). Jatkuvan mittaustason oletus perustuu siihen, että parametriset menetelmät perustuvat keskiarvon laskemiselle. Keskiarvon voi laskea vain mittaustasoltaan välimatka- ja suhdeasteikkolisista (ns. kvantitatiivisista) muuttujista. Kyselylo-

*Otos voi edustaa
populaatiota vain,
jos jokaisella
populaation jäsenellä
on ollut yhtä suuri
todennäköisyys tulla
valituksi otokseen*

makeissa tyypillisesti käytettävä Rensis Likertin (1932) kehittämä seitsemän (tai viisi) portainen asteikko on järjestysasteikko, joka tuottaa ns. kvalitatiivisia muuttujia. Toinen tyypillinen kvalitatiivinen muuttuja on mittaustasoltaan nominaali, esimerkiksi sukupuoli tai tehtävä organisaatiossa. Toki yksittäisistä kvalitatiivisista, epäjatkuvista (discrete), muuttujistakin voi keskiarvon teknisesti laskea, mutta sitä ei käytännössä kannata tehdä. Ensinnäkin, analyysissa ja myöhemmin tulosten tulkinnassa ei tällöin käytetä alkuperäisiä vastausarvoja (esim. 1 = mies, 2 = nainen; 1 = täysin eri mieltä, ..., 5 = täysin samaa mieltä), vaan niiden sijaan käsitellään arvoja, joita kyselylomakkeeseen vastanneet henkilöt eivät todellisuudessa ole voineet valita. On mielekkäämpää ilmoittaa sukupuolesta moodi tai mediaani kuin keskiarvo ja keskihajonta. Toiseksi, vaikka kuvailuvien tunnuslukujen yhteydessä Likertasteikkolisista monivalintaväittämiä voidaan esittää tulkinnallisesti mielek-

käitä keskiarvoja ja hajontoja, näin voidaan tehdä vain niiden muuttujien kohdalla, jotka ovat otoksessa normaalijakautuneita.

Toinen oletus on se, että sekä tutkittava ilmiö että sitä kuvaava otos ovat normaalijakautuneita (normal distribution). Keskiarvoperustaiset monimuuttujanalyysit, esimerkiksi kahden muuttujan välinen tulomomenttikorrelaatiokerroin (tai useamman muuttujan välisen osittaiskorrelaatiokertoimen laskeminen), perustuvat siihen, että kaikkien analyysiin valittujen muuttujien jakaumat ovat normaaleja. Normaalijakauman käytölle on luonteva historiallinen peruste: 1950-luvulle saakka tieteelliset laskutoimitukset suoritettiin pääasiassa käsin, jolloin parametriperustaisten jakaumien käyttö vähensi olennaisesti työmäärää.

Kolmantena oletuksena on tarkasteltavien riippuvuussuhteiden lineaarisuus. Vertailtavien muuttujien arvojen tulee joko kasvaa tai vähentyä monotonisesti, toisin sanoen muuttujan x arvojen kasvaessa, muuttujan y arvojen tulee joko kasvaa tai vähentyä siten, että kullekin arvolle on vain yksi vastinpari. Jos esimerkiksi viisi vastaajaa vastaa kysymyksiin x ja y viisiportaisella asteikolla arvopareilla $\{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}$, tuloksena on aidosti vähenevä monotoninen funktio ja lineaarinen yhteys. Jos vastaajat vastaavat arvopareilla $\{(1, 5), (2, 4), (3, 3), (4, 4), (5, 5)\}$, tuloksena on epälineaarinen yhteys muuttujien välillä.

Neljäntenä oletuksena on se, että (käytettävästä analyysimenetelmästä riippuen) havaintoaineiston koon tulisi olla vähintään sadan, mielellään kahdensadan paremmalla puolella. Tämä oletus

juontaa juurensa normaalijakauman ehdosta, koska ko. jakauman muodostuminen edellyttää tietynsuuruista frekvenssikertymää.

Viidentenä on oletus satunnaisotannalla hankitusta otoksesta. Tämä oletus on perusteltavissa sillä, että koska tutkittavan ilmiön on oltava normaalijakautunut sekä populaatiossa että sitä edustavassa otoksessa, otos voi edustaa (normaalijakaumallaan) populaatiota vain, jos jokaisella populaation jäsenellä on ollut yhtä suuri todennäköisyys tulla valituksi otokseen. On myös syytä muistaa, että erityisesti vastaajaryhmien vertailuun suunnitellut analyysit, kuten laajalti käytetty varianssianalyysi (Analysis of Variance, ANOVA), perustuvat satunnaisotantaa edellyttävään tutkimusasetelmaan (koe vs. kontrolliryhmä).

Kuudes oletus saattaa kuulostaa hieinan kummalliselta, mutta on olennainen tämän artikkelin kannalta: Parametrisilla menetelmillä analysoitavan empiirisen frekventistisen aineiston tulee sisältää *numeerista* informaatiota, koska muutoin keskiarvon laskeminen olisi mahdotonta.

Ensimmäisen oletuksen tilastotieteilijät ovat viimeisen kymmenen vuoden aikana pyrkineet kiertämään osoittamalla simulointiaineistoihin perustuvien tutkimusten avulla (ks. esim. Johnson & Creech 1983), että useissa tapauksissa epäjatkuva ns. kvalitatiivisia muuttujia sisältävä aineisto voidaan analysoida riittävän luotettavasti (robust) kvantitatiivisille muuttujille suunnitelluilla analyysimenetelmillä. Tämän luonteeltaan *teknisen* ongelman ohittaminen 'aineiston riittävällä mittatarkkuudella' ei kuitenkaan poista tieteellisen tutkimuksen laadun kannalta tärke-

ää, edellä mainittua toista, *teoreettista* ongelmaa, joka koskee sitä, ovatko operationalisoitujen muuttujien taustalla olevat *ilmiöt* luonteeltaan jatkuvia vai epä-jatkuvia, normaaleja vai epänormaaleja! Tilanne helpottuu hieman, jos teoreettinen malli olettaa tiettyjen normaalijakautuneiden, mutta yksittäisellä tasolla epäjatkuvien (esim. Likert-asteikolla mitattujen), muuttujien kuvaavan samaa ilmiötä: Näistä muuttujista lasketun summamuuttujan oletetaan käyttäytyvän jatkuvan muuttujan tavoin. Psykometriikan (ks. esim. Nunnally 1978) perusohje on, että summamuuttujan luomiseen tarvitaan vähintään kolme yksittäistä muuttujaa. Jos tarkasteltavan ilmiön ja sitä operationalisoivan otosjakauman ei voida olettaa olevan normaaleja, ei normaalijakaumaan perustuvia parametrisia menetelmiä tulisi käyttää aineiston analysoinnissa.

Kolmantena mainittu oletus muuttujien välisten riippuvuussuhteiden lineaarisuudesta on kasvatustieteellisessä tutkimuksessa muutamaa poikkeusta lukuunottamatta jätetty huomiotta (Nokelainen 2008).

Neljänteen oletukseen voidaan monissa tapauksissa vastata kasvattamalla otoskoko. Traditionaalisia lineaarisia frekventistisiä menetelmiä käytettäessä havaintoaineiston koon kasvattaminen aiheuttaa kuitenkin ongelmia tulosten *tilastollisen merkitsevyyden* tulkinnaissa, koska otoskoko on yksi tilastollisen merkitsevyyden laskennan komponentti. Vaikka nollahypoteesiin perustuvan tutkimuksen tulosten merkitsevyydestä on keskusteltu aktiivisesti vuosikymmenen ajan (ks. esim. Thompson 1994), silti harva sosiaalitieteilijä liittyy vielä tänään päivänä raporttiinsa lukijaa palvelevan, otoskoon huomioivan analyysin

saatujen tulosten *tieteellisestä merkitsevyydestä* (effect size analysis). On luonnollista, että frekventistiset menetelmät edellyttävät frekvenssiä toimiakseen luotettavasti. On yhtä luonnollista, että kasvatustieteilijät haluavat tutkia ilmiöitä, joiden luotettava ja edustava ('representativeness', Gobo 2004) mittaaminen ei edellytä frekventististä otosta. Olisi naivia väittää, että muulla kuin frekventistisellä menetelmällä vastatut tutkimuskysymykset olisivat mielettömiä.

Viidennen oletuksen ehdon täyttäminen on toisaalta kiinni rahasta (satunnaisotoksen tilaaminen edellyttää usein osoitetietojen hankkimista esim. Tilastokeskukselta tai suoraan kohdejoukkoa edustavilta jäsenjärjestöiltä maksua vastaan) ja toisaalta tutkimustehtävän luonteesta. On ymmärrettävää, että joissakin tapauksissa kokonaisotanta (ns. tapaustutkimus, case study) on ainoa mahdollisuus tutkimusaineiston keruustrategiaksi. Jokaisen kvantitatiivisen aineiston keräämistä suunnittelevan tutkijan tulisi toisaalta vakavasti pohtia, ovatko edellä kuvatut selitykset päteviä syitä olla hankkimatta satunnaisotantaa, koska tällöin vaarannetaan kvantitatiivisen tutkimuksen merkittävä lisäarvo kvalitatiiviseen lähestymistapaan verrattuna, yleistettävyyden populaatioon (Jackson 2006).

Kuudetta oletusta numeerisesta aineistosta voidaan kiertää luokittelemalla nominaalinen tai järjestysasteikollinen tieto numeerisiin luokkiin (esim. 1 = johtaja, 2 = opettaja, 3 = muu). On kuitenkin muistettava, että numeroiksi koodattu tekstimuotoinen kvalitatiivinen tieto on analyysissa edelleenkin kvalitatiivista tietoa, mittaustasoltaan nominaali- tai järjestysasteikollista.

Bayesilainen todennäköisyysteoria (ks. esim. Bernardo & Smith 2000; Myllymäki & Tirri 1998) käsittelee sitä, kuinka varmoja olemme tietyn väittämän paikkansapitävyyden todennäköisyydestä.

Bayes-laskenta sopii epäjatkuvan aineiston analyysiin hyvin, koska otoskoolle ja mittaustasolle ei aseteta muita kuin subjektiivisia oletuksia. Koska Bayes-laskenta operoi todennäköisyyksien, ei frekvenssijakaumien parissa, sen avulla ei selvitetä tilastollisia merkitsevyyksiä. Käytännössä tämä tarkoittaa sitä, että bayesilaisella laskennalla ei ole muita kuin tieteellisessä mielessä optimaalisia otoskokoja: Jos on tieteellisesti mielekästä tarkastella tyhjää havaintojoukkoa, se voidaan teknisessä mielessä tehdä pelkän etukäteistiedon (a priori -information) pohjata. Toisaalta, jos tutkimusongelman ratkaiseminen edellyttää satoja, tuhansia tai miljoonia havaintoja, laskenta-algoritmit eivät frekventististen parametrusten menetelmien tapaan tuota otoskoon suuruudesta johdettavaa tilastollisesti merkitsevää tulosta (ks. esim. Murphy & Myors 1998, 5-6). Bayes-laskenta soveltuukin erityisen hyvin ihmistieteiden ja yritysmaailman tilastoaineistojen tulkintaan, koska se mahdollistaa tarvittaessa faktuaaliseen tai asiantuntijatietämykseen perustuvan *a priori* -informaation sisällyttämisen tutkimuksen kohteena olevaan malliin.

Edellä mainittu etukäteistiedon käyttö herättää helposti ns. objektiivisen tilastotieteen edustajan mielessä seuraavan kysymyksen: Kuinka a priori -arvot asetetaan, jotta tutkimus palvelisi tiedeyhteisöä mahdollisimman hyvin? Jos etukäteistietoa on saatavilla, esimerkiksi haastatteluihin, aiempaan kokemuk-

seen tai tilastoihin pohjautuen, a priori -arvot voidaan asettaa sen perusteella. Jos etukäteistietoa ei ole saatavilla, tai sitä ei ”objektiivisuuden” nimissä haluta käyttää, a priori -arvot voidaan asettaa ns. ESS-kaavan (equivalent sample size, ks. Heckerman, Geiger & Chickering 1995) perusteella:

$$ESS = \frac{|V_1| + |V_2| + \dots + |V_n|}{2N} \quad (1)$$

Kaavassa 1 $|V_i|$ edustaa kunkin muuttujan V_i saamien arvojen lukumäärää (esim. tyypillisesti kyselylomakkeessa on viisi tai seitsemän vastausvaihtoehtoa) ja N kuvaa muuttujien V_i (esim. kyselylomakkeen väittämä) lukumäärää (Myllymäki, Silander, Tirri & Uronen 2002). Jos analysoitavassa mallissa on kolme viisiarvoista muuttujaa, kunkin muuttujan a priori -arvoksi asetetaan analyysissa 2.5 $((5+5+5)/6)$.

Seuraavaksi vertaan frekventististä ja bayesilaista lähestymistapaa hypoteesintestaukseen.

Frekventistisen ja bayesilaisen hypoteesintestauksen vertailua

Yksi tärkeimmistä sosiaalitieteiden keskeisten tutkimusjulkaisujen julkaistavaksi hyväksymien kvantitatiivisten artikkeleiden tulosten merkitsevyyskriteereistä on viimeiset viisikymmentä vuotta ollut ns. merkitsevyytestaus (significance testing), joka perustuu nollahypoteesin H_0 hylkäämiseen (Gigerenzer 2000). Tutkimusraportoinnin perusteella ensimmäinen ihmistieteiden tutkijoiden keskuudessa yleinen p -arvoon liittyvä väärinymmärrys kytkee em. arvoja 0..1 välillä saavan tunnusluvun saatujen tulosten toistettavuuteen, eli

todennäköisyyteen sille, että sama tulos saavutetaan toistotutkimuksessa (Thompson 2006). Toinen, sekä ihmistieteiden tutkijoiden että opiskelijoiden jakama väärinymmärrys on se, että p -arvo voisi vahvistaa tai hylätä nollahypoteesin (Haller & Krauss 2002). Toiveajattelusta huolimatta, $P(D|H_0)$ ei ole sama asia kuin $P(H_0|D)$, joten merkitsevyytestaus ei voi tarjota todennäköisyyttä nollahypoteesille – toisin kuin myöhemmin esiteltävä Bayes-mallinnus. Tilastollisen merkitsevyyden testaus kertoo todennäköisyyden sille, että nollahypoteesi on täydellisesti paikkansa pitävä tietyssä populaatiossa. Näin ollen on vaikeaa osoittaa, ainakaan empiirisen kasvatustieteen ongelmakentässä, sellaista tutkimusongelmaa, jossa nollahypoteesi voisi olla muuta kuin epätosi. Universumissamme vain äärimmäisen harva asia on täysin itsenäinen – siis vähäisimmässäkään määrin muista tekijöistä riippumaton. Koska nollahypoteesi ei voi koskaan olla täysin tosi, on se aina hylättävissä otoskoko kasvatamalla.

Vaikka suurten havaintoaineistojen edut ovat kiistattomat tavoiteltaessa suurinta mahdollista varmuutta nollahypoteesin hylkäämispäätökselle, on muistettava, että kasvatustieteen klassikot kuten Piaget, Pavlov, Bartlett ja Skinner kehittivät teorianensa yksittäisiä koehenkilöitä tarkkailemalla ilman nollahypoteesin hylkäämistä. Sosiaalitieteen tutkimuksessa opittiin 1940 -luvulla tarkastelemaan tutkimusongelmia nollahypoteesin hylkäämisen avulla pitkälti Sir Ronald Aylmer Fisherin (1890-1962) työn ansiosta. Fisheriä pidetään yhtenä tämän vuosisadan merkittävimmistä tilastotieteilijöistä, koska hän muutti pitkän uransa aikana siihenastiset hajanaiset ad hoc -tekniikat systemaattisiksi teo-

reettisiksi käsitteiksi ja käytännöllisiksi menetelmiksi.

Fisher vakuutti suurimman osan korkeellista tutkimusta tekevästä tilastotieteilijöistä pääasiassa ennen toista maailmansotaa julkaisemillaan suurta julki-suutta saaneilla tutkimuksillaan, joissa esitettiin mm. ajatus siitä, että nollahypoteesi voidaan vain hylätä, ei vahvistaa. Fisherin mukaan todennäköisyys voidaan laskea aineistolle annettuna hypoteesi $P(D|H_0)$, kun taas bayesilaisen ajattelun mukaan myös muille kuin nollahypoteeseille voidaan laskea todennäköisyys annettuna aineistona $P(H|D)$. Fisherin (1935) mukaan nollahypoteesi hylätään, kun esim. t - tai F -testin antama tunnusluku on tilastollisesti merkitsevä. Tilastolliselle merkitsevyydelle tulee asettaa tutkijan kannalta tarkoituksenmukainen raja-arvo (yleensä 5 % tai 1 %) ennen tilastollisen merkitsevyyden laskemista. Tällä Fisher halusi korostaa sitä, että raja-arvon tuli riippua tutkimustehtävän luonteesta. Nokelainen (2008) esittää em. lähestymistapojen tarkemman laskennallisen vertailun.

Koska Fisherin ajatukset oli esitetty matemaattisessa muodossa, soveltajille kirjoitettiin niistä erikseen tulkintoja 1950 ja 60 -luvuilla. Myöhemmin, kun Jerzy Neymanin (1894-1981) ja Egon Pearsonin (1885-1980) osin ristiriitaiset ajatukset tulivat julki, ne lisättiin lähteitä mainitsematta fisheriläisten mallien päälle. Myöhemmin Fisher muutti ajatuksiaan hypoteesin testauksesta ja astui askeleen kohti bayesilaisuutta myöntämällä, että subjektiivisella todennäköisyydellä (tai kuten hän asian ilmaisi, ”käänteisellä todennäköisyyslaskennalla”, inverse probability) on meriittinsä (1956, 8-17).

Karl Popper (1902-1994) kannatti Fisherin alkuperäistä ajatusta siitä, että nollahypoteesi voidaan vain hylätä, toisin kuin Harold Jeffreys (1891-1989), joka näki käytännön tulkinnan hyötyjä siitä, että kaikkien tieteellisesti mielenkiintoisten hypoteesien paikkansapitävyyttä voitiin ilmaista todennäköisyydellä, joka saa arvoja välillä 0..1 (Lindley 2001, 403). Fisher taisteli suuren osan aktiivisesta työurastaan sekä toisten frekventistien (mm. Neyman ja Pearson) että bayesilaisittain orientoituneiden tutkijoiden (mm. Jeffreys) ajatuksia vastaan. Fisherin suuren vaikutusvallan vuoksi monet hänen myöhemmin itsekin huonosti toimiviksi myöntämistään ajatuksista, kuten nollahypoteesin testaus, ovat jääneet elämään soveltavan tilastotieteen oppikirjoihin ja tutkimuskäytänteisiin, esimerkiksi tieteellisten aikakauslehtien julkaisukriteereihin.

Seuraavaksi kuvaan bayesilaisen ajattelun, ns. *subjektiivisen todennäköisyyslaskennan*, peruseriaatteita ja vertaan niitä traditionaaliseen frekventistiseen lähestymistapaan.

Objektiivista ja subjektiivista todennäköisyyslaskentaa

1 1750-luvun puolivälissä tiedettiin, että jos jokaisessa n kokeessa onnistumisen mahdollisuudella oli sama arvo θ , niin onnistumisen tarkka todennäköisyys r saadaan binomijakau-
masta

$$P(r | \theta, n) = \binom{n}{r} \theta^r (1 - \theta)^{n-r}. \quad (2)$$

Jakob Bernoulli (1687-1759) oli todistanut suurten lukujen heikon lain ja Abraham de Moivre (1667-1754) oli kehittänyt binomitodennäköisyyden kaavaa arvioimalla binomikertoimessa

esiintyviä kertomatermejä symmetrisen binomijakauman tapauksessa. Myöhemmin 1800-luvun alussa Pierre-Simon de Laplace (1749-1827) esitti todennäköisyyslaskennan keskeisen raja-arvolauseen, jonka mukaan lähes mielivaltaisten muuttujien summa jakautuu yhteenlaskettavien määrän kasvaessa normaalijakauman mukaisesti. Normaalijakauma on frekventistisen tilastotieteen keskeinen käsite, joka perustuu seuraaviin oletuksiin: 1. Aineisto on frekventistinen (frekvenssikertymien analyysi edellyttää yleensä vähintään 30 havaintoa, mielellään yli sataa), 2. Mitta-asteikko on jatkuva, 3. Tutkittavien ilmiöiden jakauma on normaali, 4. Aineiston havaintojen keskiarvojen jakaumat ovat lähellä normaalia. Normaalijakaumaa käytetään tilastollisen merkitsevyyden etsimiseen edellä kuvatussa hypoteesin testauksessa.

Thomas Bayes (1701-1761) tutki käänteistä ongelmaa kysyen mitä aineisto (r, n) kertoo todennäköisyydestä θ ,

$$P(\theta | r, n) \propto P(r | \theta, n)P(\theta | n). \quad (3)$$

Kaavassa 3 $P(\theta | n)$ on a priori -jakauma todennäköisyyksille, ennen kuin tiedetään r . Se edustaa subjektiivista näkemystä siitä, miten asian oletetaan olevan ennen todisteiden huomioimista. Teoreema kuvaa sitä, kuinka näkemykset onnistumisen todennäköisyydestä θ muuttuvat etukäteen asetetusta todennäköisyydestä (a priori) jälkikäteen (a posteriori) asetettuun todennäköisyysarvioon (todistus)aineiston r vaikutuksesta. Bayesin (1763) teoreemaan kiinnitettiin huomiota vasta sen jälkeen, kun Laplace oli kehittänyt siitä vuoden 1774 artikkelissaan yleisen muodon, jonka avulla hän laski todennäköisyyden olevan $1.15 \cdot 10^{-42}$, että tuon ajan Parii-

sisä tyttöjä syntyy poikia enemmän.

Bayesin teoreeman ehdollisen todennäköisyyden käsite perustuu siihen, että todennäköisyys lasketaan käytettävissä olevan informaation avulla. Ehdollista todennäköisyyttä merkitään $P(E|H)$, jossa P = todennäköisyys (probability), E = aineisto (evidence) ja H = hypoteesi (hypothesis). A priori -todennäköisyys $P(H)$ tarkoittaa todennäköisyyttä, joka on voimassa ennen todisteiden (evidence) huomioimista, ja posterioritodennäköisyys $P(H|E)$ on todennäköisyys sille, että tietty hypoteesi on tosi, kun todisteet on otettu huomioon. Edellä kuvatun perusteella Bayesin teoreema voidaan kirjoittaa muotoon

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E|H) \cdot P(H) + P(E|\neg H) \cdot P(\neg H)} \quad (4)$$

Esimerkki Bayesin teoreeman käytöstä

Ammattikasvatuksen tutkimus- ja koulutuskeskus (AKTKK) on onnistunut viime vuosina hankkimaan aiempaa selvästi suuremman määrän ulkopuoliseen rahoitukseen perustuvia projekteja. Projekteihin palkattavilta tutkijoilta edellytetään vuosien saatossa työelämässä hankittua kokemusperäistä tietoa. Aiemmin tutkimuskeskuksen johtaja haastatteli kaikki hakijat henkilökohtaisesti, mutta hakijamäärän kasvun vuoksi se ei enää ole mahdollista. AKTKK:n psykometrikolle annettiin tehtäväksi kehittää WWW-pohjainen itsearviointikyselylomake, jonka perusteella valitaan haastatteluun kutsuttavat henkilöt. Henkilöstöosaston psykometriikko tuntee hyvin kyselylomakemittauksen haitat: Joissakin tapauksissa se saattaa estää potentiaalisen tulevan työntekijän pääsyn haastatteluun asti, tai toisaalta tuhata johtajan haastatteluaikaa työnhakijaan,

joka ei voi tulla valituksi. Psykometriikko arvioi, että lomake toimii ainakin yhdeksänkymmenen hakijan kohdalla sadasta oikein, eli antaa korkeat pisteet henkilölle, joka tulisi haastattelunkin perusteella valituksi. Hakijapopulaatiota kuvaavista tilastoista voidaan arvioida, että hakukriteerit täyttyvät noin yhdellä prosentilla koko populaatiosta. Myös aiempina vuosina suoritettujen haastattelujen tukevat tätä havaintoa. Kysymys kuuluu: Jos henkilö saa haastatteluun tarvittavat pisteet, saako hän haastattelunkin jälkeen työtä?

A priori -todennäköisyyttä $P(H)$ kuvaa työtehtäviä vastaavien henkilöiden lukumäärä kohdepopulaatiossa.

Sitä ilmaistaan tilastojen ko. populaatiosta antaman etukäteistiedon pe-

rusteella luvulla .01. A priori -oletuksen vasta-arvoksi $P(\neg H)$ asetetaan tässä esimerkissä $1-P(H)$ eli .99. Psykometrikon uskomusta kyselylomakkeen erottelevuudesta kutsutaan ehdolliseksi todennäköisyydeksi $P(E|H)$. Se saa tässä esimerkissä arvon .9. Vastaavasti $P(E|\neg H) = .1$ kuvaa niitä hakijoita, jotka kyselylomake lähettää tuloksettomaan haastatteluun. Sijoittamalla luvut Bayesin teoreemaan saadaan posterioritodennäköisyydeksi .083. Tämä tarkoittaa sitä, että vaikka kohdepopulaatiosta vain yksi sadasta (ehdoton todennäköisyys) täyttää haettavan henkilön kriteerit, noin kahdeksan sadasta (ehdollinen todennäköisyys) kyselylomakkeen perusteella haastatteluun valituista saa työpaikan.

Vuosien kuluessa itsearviointikyselylomaketta kehitetään tarkentamalla ja lisäämällä kysyttäviä asioita haastatteluista ja projektityöstä saatujen kokemusten perusteella. Psykometriikko arvioi päivi-

tetyt esivalintainstrumentin ehdolliseksi todennäköisyydeksi .99 ja sen vasta-arvoksi .05. Laskutoimitusten jälkeen uudeksi posterioritodennäköisyydeksi saadaan .167 (hakijapopulaation parametrien säilyessä ennallaan), jolloin noin seitsemäntoista hakijan sadasta ennustetaan saavan projektitutkijan työn tutkimuskeskuksessa haastattelun jälkeen.

Annan lukijalle pohdittavaksi, mikä posterioritodennäköisyys olisi, jos psykometrikko pystyisi kehittämään lomaketta vielä siten, että ehdollisen todennäköisyyden vasta-arvo pienenesi lukuun .01? Oikea vastaus on .50, eli puolet esivalintakyselylomakkeeseen rehellisesti vastanneista olisi haastattelunkin jälkeen potentiaalisia tulevia projektitutkijoita. Tästä huomaamme, että uskomusten päivittäminen on vaikeaa jopa yksinkertaisissakin päättelytehtävissä tilanteiden muuttuessa dynaamisesti (Anderson 1995, 326).

Bayesilaisen tilastollisen analyysin sovelluksia

Edellä kuvattu Bayesin teoreeman sovellusesimerkki antaa vain kalpean kuvan tämän lähestymistavan mahdollisuuksista käytännön tutkimustyössä. Helsingin yliopiston Complex Systems Computation Group (<http://cosco.hiit.fi>) on kehittänyt yhdessä Tampereen yliopiston Ammattikasvatuksen tutkimus- ja koulutuskeskuksen (<http://www.uta.fi/aktkk>) kanssa lukuisia bayesilaiseen laskentaan perustuvia käytännön sovelluksia, joista tässä mainittakoon bayesilainen luokitteluanalyysi (BCM, ks. Silander & Tirri 1999) ja riippuvuussuhdemallinnus (BDM, ks. Heckerman, Geiger & Chickering 1995).

BCM:n avulla voidaan ennustaa, mitkä aineistossa olevat havaitut muuttajat (esim. kyselylomakkeen kysymykset, laadullisen analyysin tuloksena syntyneet luokat jne.) kuvaavat parhaiten tiettyä ryhmäjäsenyyttä (esim. organisaatio, sukupuoli, akateeminen tuottavuus jne.). Luokittelussa automaattinen hakualgoritmi etsii sellaista muuttujaryhmää, joka on paras ennustaja luokamuuttujalle annettuna yksittäisen vastaajan antamat vastaukset. BCM vastaa toimintaperiaatteeltaan lineaarista erotteluanalyysia (LDA, ks. Huberty 1994) esitellen luokittelutarkkuuden (classification accuracy) prosentteina. BCM käyttää kuitenkin huomattavasti LDA:n askeltavaa valintaa (stepwise selection procedure) sofistikoituneempaa geneetisiin algoritmeihin perustuvaa menetelmää muuttujien valinnassa. Lisäksi BCM tarjoaa tutkijalle mahdollisuuden tarkastella valitun luokittelumallin ominaisuuksia prediktiivisesti, esimerkiksi kiinnittämällä luokamuuttujan arvoja ja tarkastelemalla kiinnitysten vaikutusta muiden malliin valittujen muuttujien arvojakaumiin. Koska BCM on asiantuntijajärjestelmä, erityisesti sen prediktiivisten ominaisuuksien tieteellisesti arvokas hyödyntäminen edellyttää tutkijalta mittausmallin taustalla olevien ilmiöiden syvällistä ymmärtämystä.

BDM puolestaan tarjoaa tutkijalle mahdollisuuden tuottaa aineistossa olevien muuttujien todennäköisyysjakamia kuvaavan todennäköisimmän mallin. Sen avulla voidaan tutkia tilastollisia riippuvuuksia muuttujien välillä tarkastelemalla Bayes-verkon rakennetta ja riippuvuussuhteiden voimakkuuksia. Mallin avulla voi myös ennustaa tietyn muuttujan arvojen muutosten vaikutusta toisten muuttujien todennäköisyysjakauksissa. Bayes-verkon graafinen visu-

alisointi (Myllymäki, Silander, Tirri & Uronen 2002) koostuu mitattuja muuttujia (tai summamuuttujaa) kuvaavista solmuista (ellipsi) ja niitä yhdistävistä (ei-rekursiivisista) kaarista. Kahden solmun välillä oleva kaari ilmaisee tilastollista riippuvuussuhdetta (statistical dependency) jonka suuntaa ei tiedetä.

Aikaisemmat tutkimukset ovat osoittaneet molempien edellä kuvattujen työvälineiden hyödyllisyyden tieteelliselle tutkimukselle (Nokelainen 2008; Nokelainen & Tirri 2004; Nokelainen, Silander, Ruohotie & Tirri 2007).

Lopuksi

Artikkelin alussa keskustelin lyhyesti kvalitatiivisten ja kvantitatiivisten menetelmien erilaisista tutkimusteoreettisista orientaatioista. Ne johtavat helposti tilanteeseen, jossa kvalitatiivisesti orientoitunut tutkija hämmästelee kvantitatiivisen tutkijan tulosten pinnallisuutta, ja vastaavasti kvantitatiivisilla menetelmillä operoiva tutkija kiinnittää huomionsa kvalitatiivisella menetelmällä saatujen tulosten heikkoon yleistettävyyteen. Tältä tarkastelutasolta tulisi nousta astetta ylemmälle tasolla ja pohtia, onko valittu tutkimusmetodi tutkimuskysymysten tiedeyhteisölle tuottaman tiedon laadun kannalta optimaalinen.

Esittelin seuraavaksi Bayes-laskennan peruseriaatteet ja vertasin niitä traditionaaliseen frekventistiseen lähestymistapaan. Tarkastelin myös kahden empirisen kasvatustieteen kannalta hyödyllisen bayesilaisen mallinnussovelluksen, bayesilaisen luokittelu- ja riippuvuusuhdeanalyysin, toimintaperiaatteita sekä niiden soveltuvuutta kasvatustieteiden tutkijan työkaluiksi.

Artikkelin keskeinen tehtävä oli antaa vastaus seuraavaan kysymykseen: Tarjoaako bayesilainen tilastollinen mallintaminen kasvatustieteiden tutkijalle olennaista lisäarvoa verrattuna yhä edelleen laajalti käytössä olevaan traditionaaliseen parametriseen lineaariseen frekventistiseen tilastolliseen mallintamiseen? Pyrin joitakin traditionaalisen kvantitatiivisen tutkimuksen ongelmakohtia analysoimalla antamaan lukijalle käsityksen Bayes-laskennan tarjoamista uusista mahdollisuuksista erityisesti silloin, kun tutkimusaineisto on frekvenssiltään pieni, sisältää epälineaarisia vaikutussuhteita tai tutkija haluaa tehdä aineiston perusteella ennusteita.

Bayesilaiselle lähestymistavalle on myös olemassa varteenotettavia vaihtoehtoja, esimerkiksi Bengt Muthénin (1983) kehittämä teoreettinen malli, jonka pohjalta LISREL -ohjelmassa voidaan analysoida tetrakorisia (binäärimuuttujat) ja polykorisia (järjestysasteikolliset muuttujat) korrelaatioita. Kyseessä olevat korrelaatiot mahdollistavat kategorisen aineiston mallintamisen Brownen (1984) kehittämän ADF-mallin (asymptotically distribution-free) avulla. AMOS -ohjelmisto käyttää vastaavasta sovelluksesta Brownen alkupeiräistä ADF -nimeä, mutta EQS -ohjelmistossa se on muutettu muotoon AGLS, ja MPLUS sekä LISREL -ohjelmistoissa nimenä on WLS. ADF -malliin perustuvan analyysin hyvä puoli on se, että tarkasteltavien arvojen ei tarvitse olla normaalijakautuneita. Tämän lähestymistavan puutteina voidaan pitää analysoitavien muuttujien pienehköä maksimimäärää (suositus on alle 20) ja suurta otoskokoja: Yung ja Bentler (1994) suosittelivat yli 2000 havaintoa. Lisäksi Olsson ja kumppanit (2000) havaitsivat että ADF -malliin perustuvat

analyysit eivät saavuttaneet hyviä mallinnustuloksia, jos mallinmäärittelyssä oli puutteita.

Varteenotettavin ADF -perustainen sovellus on Muthénin (1993) MPLUS -ohjelmistossaan implementoima CVM-malli (categorical variable model), josta edellä kuvatut rajoitukset puuttuvat. Viimeaikaisista kehitelmistä haluan lisäksi mainita Anders Skrondalin GLLAMM-mallin (generalized linear latent and mixed model), joka soveltuu erityisen hyvin pitkittäisaineistojen analyysiin (Rabe-Hesketh, Skrondal & Pickles 2004).

Lopuksi haluan korostaa, että kaikista kvantitatiivisista tekniikoista saadaan irti tieteellisesti arvokkain hyöty, kun muistetaan, että erityisesti korrelatiivisissa tutkimusasetelmissä havaittujen muuttujien väliset riippuvuussuhteet kuvaavat tilastollisia, eivät kausaalisia, riippuvuuksia yhdessä mahdollisessa populaatiota edustavassa otoksessa (ks. esim. Pearl 2000). Tässä artikkelissa kuvatut BCM ja BDM -sovellukset ovat asiantuntijajärjestelmiä, ja kuten pitää paikkansa kaikkien asiantuntijajärjestelmien kohdalla niiden antama informaatio jalostuu tieteellisesti arvokkaaksi tiedoksi vasta täydentäessään tutkijan tutkimusongelman ratkaisemiseen valitsemalla, mahdollisesti monimetodista strategiaa, jossa teoreettinen malli ja mitausmalli liittyvät kiinteästi toisiinsa.

Lähteet

Anderson, J. 1995. *Cognitive psychology and its implications*. Freeman: New York.

Bayes, T. 1763. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society* 53, 370-418.

Bernardo, J. & Smith, A. 2000. *Bayesian theory*. New York: Wiley.

Browne, M. W. 1984. Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology* 37, 1-21.

Fisher, R. A. 1935. *The design of experiments*. Edinburgh: Oliver & Boyd.

Fisher, R. A. 1956. *Statistical methods and scientific inference*. Third edition. Hafner: New York.

Gigerenzer, G. 2000. *Adaptive thinking*. New York: Oxford University Press.

Gobo, G. 2004. Sampling, representativeness and generalizability. In C. Seale, J. F. Gubrium, G. Gobo & D. Silverman (Eds.) *Qualitative research practice*. London: Sage, 435-456.

Haller, H. & Krauss, S. 2002. Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research Online* 7 (1), 1-20.

Heckerman, D., Geiger, D. & Chickering, D. 1995. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning* 20 (3), 197-243.

Huberty, C. 1994. *Applied discriminant analysis*. New York: John Wiley & Sons.

Jackson, S. 2006. *Research methods and statistics. A critical thinking approach*. Second edition. Belmont, CS: Thomson.

Johnson, D. R. & Creech, J. C. 1983. Ordinal measures in multiple indicator models: A simulation study of categorization error. *American Sociological Review* 48, 398-407.

Kurhila, J., Miettinen, M., Nokelainen, P. & Tirri, H. 2007. EDUCO: Social navigation and group formation in student-centred e-learning. *Journal of Interactive Learning Research* 18 (1), 65-83.

Likert, R. 1932. A technique for the measurement of attitudes. *Archives of Psychology* 140, 44-53.

Lindley, D.V. 2001. Harold Jeffreys. In C.C. Heyde & E. Seneta (Eds.) *Statisticians of the centuries*. New York: Springer, 402-405.

Marini, M., Li, X. & Fan, P. 1996. Characterizing latent structure: Factor analytic and grade of membership models. *Sociological Methodology* 1, 133-164.

Murphy, K. R. & Myers, B. 1998. *Statistical power analysis. A simple and general model for traditional and modern hypothesis tests*. Mahwah, NJ: Lawrence Erlbaum Associates.

Muthén, B. O. 1983. Latent variable structural equation modeling with categorical data. *Journal of Econometrics* 22, 48-65.

Muthén, B. O. 1993. Goodness of fit with categorical and other non-normal variables. In K. A. Bollen & J. S. Long (Eds.) *Testing structural equation models*. Newbury Park, CA: Sage, 205-243.

Myllymäki, P., Silander, T., Tirri, H. & Uronen, P. 2002. B-Course: A web-based tool for Bayesian and causal data analysis. *International Journal on Artificial Intelligence Tools* 11 (3), 369-387.

Myllymäki, P. & Tirri, H. 1998. Bayes-verkkojen mahdollisuudet. *Teknologiakatsaus* 58/98. Helsinki: TEKES.

Nokelainen, P. 2008. Modeling professional growth and learning: Bayesian approach. Tampere: Tampere University Press.

Nokelainen, P., Silander, T., Ruohotie, P. & Tirri, H. 2007. Investigating the number of non-linear and multi-modal relationships between observed variables measuring a growth-oriented atmosphere. *Quality & Quantity* 41 (6), 869-890.

Nokelainen, P. & Tirri, H. 2004. Bayesian methods that optimize cross-cultural data analysis. In J. R. Campbell, K. Tirri, P. Ruohotie & H. Walberg (Eds.) *Cross-cultural research: Basic issues, dilemmas, and strategies*. Hämeenlinna, Finland: RCVE, 141-158.

Nunnally, J. C. 1978. *Psychometric theory*. New York: McGraw-Hill.

Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. 2000. The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling* 7 (4), 557-595.

Pearl, J. 2000. *Causality. Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Rabe-Hesketh, S., Skrondal, A. & Pickles, A. 2004. Generalized multilevel structural equation modelling. *Psychometrika* 69, 183-206.

Rautopuro, J. & Malin, A. 2008. Miksei asioista puhuta niiden oikeilla nimillä? Tilastollisten menetelmien tarpeellisuudesta kasvatustieteissä. *Kasvatus* 39 (2), 108-118.

Silander, T. & Tirri, H. 1999. Bayesian classification. In P. Ruohotie, H. Tirri, P. Nokelainen & T. Silander (Eds.) *Modern modeling of professional growth*. Vol. 1. Hämeenlinna, Finland: RCVE, 61-84.

Silén, M. 2008. Tilastotiede tutkimuksessa ja opetuksessa. *Kasvatus* 39 (2), 145-154.

Thompson, B. 1994. Guidelines for authors. *Educational and Psychological Measurement* 54 (4), 837-847.

Thompson, B. 2006. *Foundations of behavioral statistics. An insight-based approach*. New York: Guilford Press.

Töttö, P. 2004. *Syvällistä ja pinnallista*. Tampere: Vastapaino.

Yung, Y. F., & Bentler, P. M. 1994. Bootstrap corrected ADF test statistics in covariance structure analysis. *British Journal of Mathematical and Statistical Psychology* 47, 63-84.

9TH CONGRESS OF METHODOLOGY:

Practice Based Inquiry
18. -19.5.2010

Tervetuloa käytäntölähtöisen tutkimuksen
kansainväliseen kongressiin HAMK
Visamäkeen Hämeenlinnaan.

Ohjelmassa kansainvälisiä
avainluentoisijoita.

Lisätietoja:
www.hamk.fi/methodology

Ilmoittautuminen päättyy 30.4.2010.

Järjestäjät:
Käytäntölähtöisen tutkimuksen yhdistys ry. PraBa.
Ammattikorkeakoulujen kehittäjäverkosto KeVer
ja Korkeakoulututkimuksen seura ry. CHERIF