Kompleksisen luokitellun aineiston pelkistäminen: sovellus ylioppilaskirjoitusaineistoon

Satu Helske

Jyväskylän yliopisto
Matematiikan ja tilastotieteen laitos

JYVÄSKYLÄN YLIOPISTO
Matematiikan ja tilastotieteen laitos

HELSKE, SATU: Kompleksisen luokitellun aineiston riippuvuusrakenteen pelkistäminen: sovellus ylioppilaskirjoitusaineistoon

Pro gradu -tutkielma, 73 sivua, kaksi liitettä (5 sivua)

Tilastotiede

Kesäkuu 2009

Tiivistelmä

Ylioppilaskokeella on merkittävä asema lukiolaisen äidinkielen osaamisen arvioinnissa. Tässä tutkielmassa on tarkasteltu kahden kevään äidinkielen kokeita ja pyritty selvittämään, millaisiin oppilasryhmiin eri tehtävänannot vetoavat ja millaisia arvosanoja ne tuottavat. Kiinnostavaa on tarkastella myös suomen- ja ruotsinkielisten tehtävänantojen eroja ja yhtäläisyyksiä.

Pyrkimyksenä on pelkistää laaja moniulotteinen ristiintaulukko useammaksi pieniulotteiseksi tauluksi, jotka on helppo esittää ja tulkita. Pelkistäminen tehdään log-lineaaristen mallien ja Schwarzin informaatiokriteerin avulla.

Kun tehtävänantomuuttujan luokkia yhdistetään klusteroimalla, saadaan aineiston yksinkertaistumisen lisäksi tulos, että harvemmin valituista aiheista saadaan parempia arvosanoja kuin suosituimmista tehtävänannoista. Tämän uskotaan johtuvan siitä, että paremmat opiskelijat valitsevat haastavampia aiheita. Valintamekanismiin ei kuitenkaan päästä käsiksi, koska opiskelijan aiemmasta osaamistasosta ei ole tietoa.

Tehtävänantoja on luokiteltu myös laadullisin menetelmin seitsemällä tavalla ja näin luotu uudet muuttujat, jotka kertovat tehtävänannon muotoilusta (aihepiiri, otsikko, teema ja toivottu genre) sekä tehtävän virityksestä (annettu tekstiaineisto, vaadittu intertekstuaalinen varanto ja moodi). Kun tehtävänannon muotoiluun liittyviä muuttujia tarkastellaan yhdessä kokelaan sukupuolen, arvosanan ja kokeen kielen kanssa, saadaan aineisto pelkistettyä kolmeen kaksiulotteiseen ja neljään kolmiulotteiseen ristiintaulukkoon, joista myös kukin kolmiulotteisista tauluista saadaan esitettyä kahtena kaksiulotteisena. Kahden muuttujan yhdysvaikutus löytyy sukupuolen ja arvosanan, otsikon ja arvosanan sekä aihepiirin ja genren väliltä. Kolmen muuttujan yhdysvaikutus on sukupuolella, aihepiirillä ja teemalla; kielellä, aihepiirillä ja teemalla; sukupuolella, aihepiirillä ja otsikolla sekä sukupuolella, teemalla ja genrellä.

Tehtävän viritykseen liittyviä muuttujia sekä sukupuolta, arvosanaa ja kieltä tarkasteltaessa päädytään kahteen kaksiulotteiseen ja kolmeen kolmiulotteiseen ristiintaulukkoon, joista jälkimmäiset saadaan jälleen esitettyä kahtena kaksiulotteisena tauluna. Kahden muuttujan yhdysvaikutus on sukupuolen ja arvosanan sekä pohjatekstin ja vaaditun intertekstuaalisen varannon välillä ja kolmen muuttujan yhdysvaikutus sukupuolen, pohjatekstin ja moodin; sukupuolen, intertekstuaalisen varannon ja moodin sekä kielen, intertekstuaalisen varannon ja moodin välillä.

Avainsanat: redusointi, moniulotteinen ristiintaulukko, loglineaarinen malli, informaatiokriteeri, mallinvalinta, ylioppilastutkinto, äidinkielen koe

Sisältö

1 Johdanto **6**

2 Tutkimusaineiston kuvaus **8**

3 Loglineaarinen malli moniulotteiselle ristiintaulukolle **10**

3.1 Loglineaarinen malli kaksiulotteiselle taululle 10

3.1.1 Riippumattomuusmalli 11

3.1.2 Kyllästetty malli 12

3.1.3 Parametrien tulkinta 13

3.2 Kolmen muuttujan parittaiset vaikutukset 13

3.2.1 Marginaali- ja osataulut 14

3.2.2 Riippumattomuuden tyypit 14

3.2.3 Marginaalinen ja ehdollinen riippumattomuus 16

3.2.4 Parittaiset yhdysvaikutukset 16

3.3 Loglineaarinen malli kolmiulotteiselle taululle 17

3.3.1 Hierarkkiset mallit 17

3.3.2 Parametrien tulkinta 18

3.3.3 Marginaalinen ja osittainen riippuvuus 19

3.3.4 Mallien tulkinta 19

3.4 Loglineaarinen malli useammalle muuttujalle 23

3.4.1 Neliulotteinen taulu 23

3.4.2 Vakiointi 23

3.4.3 Loglineaarisen ja logit-mallin välinen yhteys 24

3.5 Kvasiloglineaariset mallit 26

3.5.1 Epätäydellinen kaksisuuntainen taulu 26

3.5.2 Lineaariset kontrastit 27

3.5.3 Epäinteraktiiviset solut 27

3.5.4 Liittyneisyys ja eroteltavuus 28

3.5.5 Eroteltavien taulujen analysointi 28

3.6 Yhteensopivuuden testaaminen 30

4 Mallien vertailu suurten aineistojen tapauksessa **31**

4.1 Informaatiokriteerit 32

4.2 Informaatiokriteereihin perustuva mallinvalinta 34

4.3 Esimerkki ylioppilasaineistosta 34

5 Malliin perustuva pelkistäminen **37**

5.1 Riippuvuusrakenteet 37

5.2 Yleinen menetelmä 39

5.3 Esimerkki ylioppilasaineistosta 40

6 Luokkien yhdistäminen **42**

6.1 Luokkien yhdistäminen klusterianalyysilla 42

7 Tutkimustulokset **47**
7.1 Tehtävänannon muotoilu 47
7.2 Tehtävän viritys 66

8 Diskussio **79**

Kiitosmaininnat **80**

Liitteet **82**

A Tehtävänantojen koodaus **82**

B Tehtävänantojen luokitukset **84**

1 Johdanto

Vuonna 2006 alkoi Jyväskylän ja Helsingin yliopistojen yhteinen tutkimushanke nimeltä Kielitaidon kirjo, jossa on tarkoituksena selvittää, mitä tarkoitetaan suomea ja ruotsia äidinkielenään puhuvien lukiolaisten kielitaidolla ja miten sitä voidaan kehittää ja arvioida. Hankkeen tavoitteena on tukea äidinkielen ja kirjallisuuden opetuksen kehittämistä tilanteessa, jossa viestintäympäristön muutos asettaa äidinkielen taidoille uudenlaisia haasteita ja jossa kielitaitoa mittaavia ylioppilastutkinnon koetyyppejä kehitetään.

Aiempi selvitys osoitti, ettei lukiolaisten äidinkielen osaamisesta ole olemassa tutkimustietoa juuri lainkaan. Ylioppilaskokeella on tunnetusti merkittävä asema lukiolaisten äidinkielen osaamisen arvioinnissa, joten hankkeessa tarkastellaan sitä sekä kvalitatiivisin että kvantitatiivisin menetelmin. Näitä menetelmiä sovelletaan Ylioppilastutkintolautakunnalta (YTL) saatuun laajaan rekisteriaineistoon. Erityisesti ollaan kiinnostuneita selvittämään, millaisiin oppilasryhmiin eri tehtävänannot vetoavat ja minkälaisia arvosanoja ne tuottavat. Lisäksi halutaan selvittää suomen- ja ruotsinkielisten tehtävänantojen eroja ja yhtäläisyyksiä.

YTL:n aineisto sisältää paljon luokiteltuja muuttujia, joista tässä tutkimuksessa päädyttiin käsitelemään vain muutamaa: äidinkielen kokeen arvosanaa, kieltä ja tehtävänantoa sekä kokelaan sukupuolta. Lisäksi asiantuntijat ovat luokitelleet tehtävänantoja useiden kriteerien perusteella ja muodostaneet näin uusia tehtävänantojen ominaisuuksiin perustuvia laadullisia muuttujia. Analyysissa ei pyritä testaamaan ennalta asetettuja hypoteeseja vaan lähestymistapa on data-analyyttinen. Tavoitteena on esittää moniulotteisia ristiintaulukoita useamman yksinkertaisen taulun avulla, joiden esitys on alkuperäistä selkeämpi mutta jotka sisältävät kuitenkin kaiken kiinnostavan informaation. Nämä taulut ovat helposti tulkittavissa, jotta lukija voi itse tehdä johtopäätökset myös ilman tilastotieteen tuntemusta.

Analyysissa käytetään loglineaarisia malleja riippuvuussuhteiden tarkasteluun sekä logit-malleja silloin, kun halutaan selvittää syy-seuraussuhteita. Perinteisesti sopivin malli etsitään uskottavuusosamäärätestin avulla, mutta näin suuren aineiston tapauksessa pienetkin erot tulevat merkitseviksi eikä malli pelkisty riittävästi. Koska kiinnostaviin eroihin ei päästä käsiksi testauksella, käytetään mallien vertailuun informaatiokriteereitä. Ne ottavat huomioon uskottavuuden ja sakottavat mallia sen parametrien määrän mukaan. Tässä tutkimuksessa näistä on käytetty Akaiken ja Schwarzin kriteereitä (AIC ja BIC).

Taulun tulkitseminen voi olla hankalaa, jos muuttujalla on useita luokkia. Niitä voidaan kuitenkin yhdistellä joko tietokoneella tilastollisin menetelmin tai laadullisen tutkimuksen avulla. Edellinen voidaan tehdä esimerkiksi klusterointimenetelmillä, joilla saadaan yhdistettyä luokkia helpommin hahmotettaviksi kokonaisuuksiksi. Jälkimmäinen tapa johtaa helppoon tulkintaan, mutta saattaa vaatia vahvaa sovellusalan asiantuntemusta. Tässä aineistossa tapa tuottaa lisäksi runsaasti rakenteellisia nollia, jotka tulee ottaa huomioon analyysissa. Siksi loglineaaristen mallien sijasta käytetään kvasiloglineaarisia malleja, joissa rakenteelliset nollat kiinnitetään.

Esimerkkinä sovelluksesta käsitellään kuusi- ja seitsemänulotteisia ristiintaulukoita, jotka sisältävät kahdentyyppisiä muuttujia. Esimerkiksi näistä muuttujista suku-

puoli ja kokeen kieli ovat yksikäsitteisiä. Tehtävänannon aihepiiri ja teema on sen sijaan saatu sovellusalan asiantuntijoilta laadullisen tutkimuksen perusteella tehdyistä luokitteluista. Tässä yhdistetään siis laadullinen ja määrällinen lähestymistapa. Aineiston sisältämät oleelliset riippuvuudet saadaan esitettyä kaksi- ja kolmiulotteisina tauluina, joiden sisältämä informaatio on nyt helposti tulkittavissa prosenttijakaumien avulla.

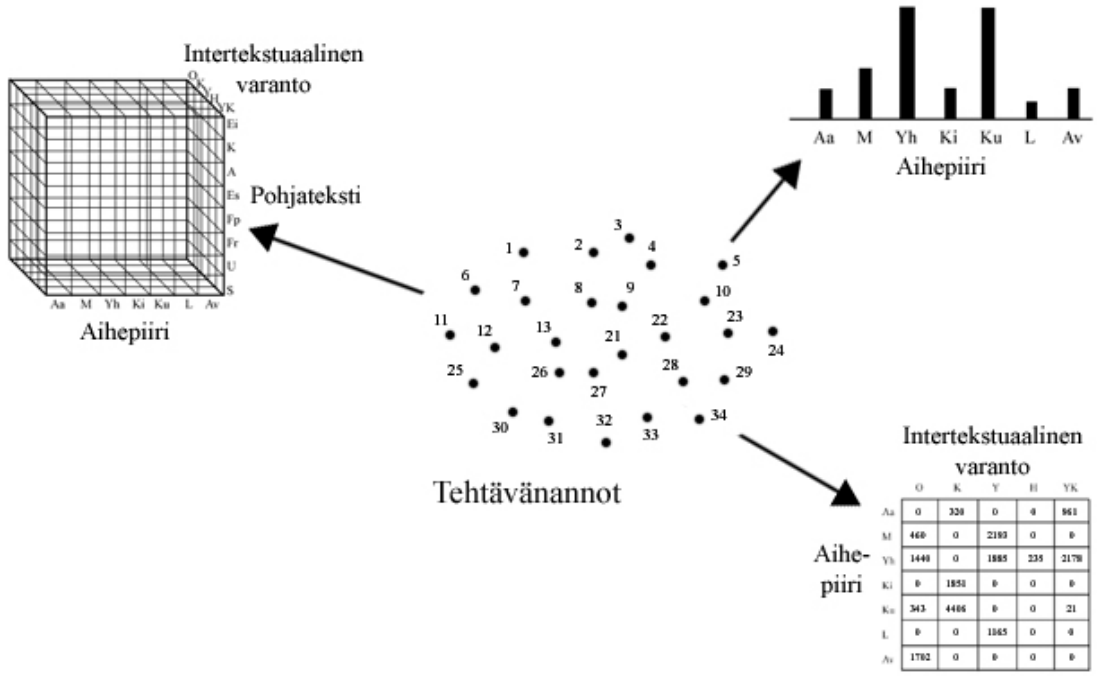
Luvussa 2 tutustutaan aluksi tutkimusaineistoon ja luvussa 3 esitellään loglineaarisen mallin muodostaminen moniulotteiselle kontingenssitaululle. Luvussa 4 käsitellään mallien vertailua suurten aineistojen tapauksessa: informaatiokriteereitä, mallinvalinnan periaatteita ja informaatiokriteereihin perustuvaa mallinvalintaa. Luvussa 5 tarkastellaan aineiston pelkistämistä ja luvussa 6 luokkien yhdistämistä klusterianalyysin avulla. Lopuksi tutkimustuloksia esitellään luvussa 7.

2 Tutkimusaineiston kuvaus

Ylioppilastutkintolautakunnalta saatu aineisto sisältää tiedot kaikkien keväällä 2006 ja 2007 suomen, ruotsin tai saamen äidinkielen kirjoittaneiden pakollisista kokeista. Tämän tutkimuksen osalta kiinnostavia muuttujia ovat äidinkielen kokeesta saatu arvosana, kokeen kieli, kokeen tehtävänanto ja kokelaan sukupuoli, joka saadaan käyttöön henkilötunnuksen avulla. Näiden kolmen muuttujan lisäksi aineistosta löytyy myös muita kokelaaseen ja kokeeseen liittyviä muuttujia, jotka päätettiin kuitenkin rajata tästä tutkimuksesta pois. Kokeen kielenä on otettu huomioon ainoastaan suomi ja ruotsi, koska näiden puhujien vertailu on yksi tutkimuksen painopisteistä ja lisäksi saamen äidinkielenään kirjoittaneiden määrä on hyvin pieni.

Huomionarvoista on, että äidinkielen kokeen rakennetta muutettiin keväällä 2007. Aiemmassa koetyypissä, siis myös keväällä 2006, kirjoitettiin aineisto- ja otsikkokoe. Tällöin kokelaan arvosana määräytyi paremman suorituksen mukaan ja huonompi jätettiin huomiotta. Vuoden 2007 uudistuksessa ensimmäisenä päivänä kirjoitetaan uusi tekstitaidon koe ja toisena perinteinen esseekoe. Arvosana määräytyy nyt näiden yhteispisteiden mukaan.

Aineistossa valmiina olevien muuttujien lisäksi on kolme kielitaidon, ylioppilastutkinnon ja opetuksen asiantuntijaa analysoinut tehtävänantoja ja luokitellut niitä erilaisiin ryhmiin useiden kriteerien suhteen. Jokainen tehtävänanto kuuluu kussakin uudessa muuttujassa yhteen luokkaan ja näitä voidaan tarkastella myös useita kerrallaan (Kuva 2.1).



Kuva 2.1: Tehtävänantoja voidaan luokitella useilla tavoilla.

Tehtävän *aihepiirillä* (ap) määritellään aihe kokonaisuutena eli niin kuin se on

määrätty tehtävänannon ja aineiston perusteella. Se voi olla myös niiden yhteisvaikutuksesta syntyvä, tarjottu tarkastelutapa. Esimerkkejä mahdollisista aihepiireistä ovat media, talous, yhteiskunta ja kulttuuri. Muuttuja *otsikko* (o) kertoo, onko aineelle annettu valmis otsikko. *Pohjateksti* (pt) on tehtävässä mainitun ja koetilanteessa annetun aineiston tai aineistojen tyyppi. Aineisto voi olla esimerkiksi kuva, essee tai runo. Useita erityyppisiä aineistoja sisältävät tehtävänannot muodostavat oman luokkansa. *Teema* (t) kertoo, onko tehtävänannossa määritelty teema vai onko se jätetty avoimeksi tai määritelty vain väljästi. Muuttuja *inter* (i) määrittelee millaista intertekstuaalista varantoa opiskelijalta vähintään edellytetään. Kokelaalta voidaan vaatia kulttuuri-, yhteiskunta- tai historiatietoa. *Moodi* (m) kuvaa, millaista ääntä tehtävänanto kutsuu. Kokelaan voidaan olettaa kirjoittavan esimerkiksi tehtävänannosta, joka on ilmaistu korkeakulttuurisella, yleiskielisellä tai interpersoonallisella eli kirjoittajan kokemukseen vetoavalla tyylillä. *Genre* (g) kertoo, onko pyydetyn kirjoitelman tekstilaji määritelty tai onko siihen annettu viittauksia. Tarkemmat tiedot tehtävänantojen luokitteluista ja luokista löytyvät liitteistä A ja B.

3 Loglineaarinen malli moniulotteiselle ristiintaulukolle

Tässä luvussa tarkastellaan moniulotteisen luokitellun aineiston riippuvuusrakenteen mallintamisen periaatteita. Koska aineisto on luokiteltu, voidaan se kuvata ristiintaulukkoina. Tutkittaessa pelkkiä riippuvuussuhteita tehdään tämä loglineaarisin mallein, vasteen ja selittäjien tapauksessa logit-malleilla. Tässä esitelty loglineaaristen mallien teoria pohjautuu Alan Agrestin oppikirjaan *Categorical Data Analysis* (1990).

3.1 Loglineaarinen malli kaksiulotteiselle taululle

Olkoot X ja Y luokiteltuja satunnaismuuttujia siten, että muuttujalla X on I tasoa ja muuttujalla Y J tasoa. Kun luokitellaan havainnot kummankin muuttujan suhteen, on mahdollisia luokkien yhdistelmiä IJ kappaletta. Muuttujilla X ja Y on todennäköisyysjakauma, jota voi kuvata I:n rivin ja J:n sarakkeen muodostamassa taulussa. Taulun IJ solua kuvaavat mahdollisia tuloksia, joiden todennäköisyydet ovat $\{\pi_{ij}\}$, missä π_{ij} kuvaa todennäköisyyttä, että satunnaisesti poimittu havainto osuu riville i ja sarakkeeseen j. Ristiintaulukoksi kutsutaan taulua, jonka soluissa kuvataan havaintojen frekvenssit $\{n_{ij}\}$ (Taulukko 1). Nämä noudattavat multinomijakaumaa, ts.

$$n_{ij} \sim \mathrm{Mult}(n, \{\pi_{ij}\}), \quad \text{missä } i = 1, \ldots, I,\ j = 1, \ldots, J.$$

Poisson-malli johtaa samoihin estimaatteihin ja siten myös johtopäätöksiin, joten analyysissa voitaisiin yhtä lailla käyttää sitä. Havaintojen odotusarvoille eli odotetuille frekvensseille pätee $\mathbb{E}n_{ij} = n\pi_{ij} = m_{ij}$.

Taulukko 1: Kaksiulotteinen ristiintaulukko

	1	2	...	J	Yht.
1	n_{11}	n_{12}	...	n_{1J}	n_{1+}
2	n_{21}	n_{22}	...	n_{2J}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	n_{I1}	n_{I2}	...	n_{IJ}	n_{I+}
Yht.	n_{+1}	n_{+2}	...	n_{+J}	n

Todennäköisyysjakauma $\{\pi_{ij}\}$ on muuttujien X ja Y yhteisjakauma (Taulukko 2). Marginaali- eli reunajakaumat saadaan summaamalla yli yhteisjakauman rivien tai sarakkeiden. Näitä merkitään $\{\pi_{i+}\}$ ja $\{\pi_{+j}\}$, missä plusmerkki kuvaa summaa yli sen indeksin, jonka se korvaa. Siispä

$$\pi_{i+} = \sum_j \pi_{ij} \quad \text{ja} \quad \pi_{+j} = \sum_i \pi_{ij},$$

ja näille pätee $\sum_i \pi_{i+} = \sum_j \pi_{+j} = \sum_i \sum_j \pi_{ij} = 1$. Reunajakaumat eivät sisällä tietoa muuttujien välisestä assosiaatiosta.

Taulukko 2: Kaksiulotteinen todennäköisyysjakauma

	1	2	...	J	Yht.
1	π_{11}	π_{12}	...	π_{1J}	π_{1+}
2	π_{21}	π_{22}	...	π_{2J}	π_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	π_{I1}	π_{I2}	...	π_{IJ}	π_{I+}
Yht.	π_{+1}	π_{+2}	...	π_{+J}	1

3.1.1 Riippumattomuusmalli

Yhteisjakauman muodostavien muuttujien sanotaan olevan tilastollisesti riippumattomia, kun $\pi_{ij} = \pi_{i+}\pi_{+j}$, jolloin marginaalijakaumat $\{\pi_{i+}\}$ ja $\{\pi_{+j}\}$ määräävät siis koko taulun $\{\pi_{ij}\}$. Tällöin odotetut frekvenssit ovat $m_{ij} = n\pi_{i+}\pi_{+j}$ ja odotetun frekvenssin logaritmi solulle (i, j) saa muodon

$$\log m_{ij} = \log n + \log \pi_{i+} + \log \pi_{+j}, \tag{3.1}$$

missä π_{i+} kuvaa rivin i vaikutusta ja π_{+j} sarakkeen j vaikutusta odotettuun frekvenssiin.

Kun kuvataan rivimuuttujaa X:llä ja sarakemuuttujaa Y:llä, voidaan yhtälö (3.1) ilmaista muodossa

$$\log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y, \tag{3.2}$$

missä

$$\lambda_i^X = \log \pi_{i+} - \left(\sum_h \log \pi_{h+} \right) / I$$

$$\lambda_j^Y = \log \pi_{+j} - \left(\sum_h \log \pi_{+h} \right) / J$$

$$\mu = \log n + \left(\sum_h \log \pi_{h+} \right) / I + \left(\sum_h \log \pi_{+h} \right) / J.$$

Ylläolevista parametrien määrittelyistä seuraavat nollasummarajoitteet

$$\sum_i \lambda_i^X = \sum_j \lambda_j^Y = 0.$$

Nämä ovat yleinen tapa saada mallin parametreista identifioituvia, mutta parametrit on mahdollista määritellä myös toisin. Kun lisätään vakio yhteen termiin ja vähennetään se toisesta, saadaan malli, joka tuottaa samat odotetut frekvenssit. Tällöin esimerkiksi $\lambda_i^X = \log \pi_{i+} - \log \pi_{I+}$ ja $\lambda_i^Y = \log \pi_{+j} - \log \pi_{+J}$, jolloin $\lambda_I^X = 0$ ja

$\lambda_J^Y = 0$. Kummassakin tapauksessa parametrit voidaan tulkita seuraavalla tavalla: kun λ_i^X kasvaa, lisääntyy alkioiden määrä muuttujan X tasolla i.

Rajoitteet vähentävät mallin vapausasteita. Siinä olevat $(I-1)$ riviparametria ja $(J-1)$ sarakeparametria voidaan määrätä vapaasti, mutta loput määräytyvät nollasummarajoitteiden perusteella. Yhteensä riippumattomuusmallissa on vapausasteita $1+(I-1)+(J-1) = I+J-1$ kappaletta.

3.1.2 Kyllästetty malli

Oletetaan, että muuttujien välillä on riippuvuutta ja kaikki odotetut frekvenssit $m_{ij} > 0$. Olkoon

$$\eta_{ij} = \log m_{ij},$$

$$\eta_{i.} = \left(\sum_j \eta_{ij}\right) / J = \frac{\eta_{i+}}{J},$$

$$\eta_{.j} = \left(\sum_i \eta_{ij}\right) / I = \frac{\eta_{+j}}{I}$$

ja

$$\mu = \eta_{..} = \left(\sum_i \sum_j \eta_{ij}\right) / IJ = \frac{\eta_{++}}{IJ}.$$

Kun nyt valitaan $\lambda_i^X = \eta_{i.} - \eta_{..}$, $\lambda_j^Y = \eta_{.j} - \eta_{..}$ ja $\lambda_{ij}^{XY} = \eta_{ij} - \eta_{i.} - \eta_{.j} + \eta_{..}$, saadaan odotetun frekvenssin logaritmille esitys

$$\log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}. \tag{3.3}$$

Tällaista muuttujien välisen yhdysvaikutustermin sisältävää mallia kutsutaan saturoiduksi eli kyllästetyksi. Parametrit $\{\lambda_i^X\}$ ja $\{\lambda_j^Y\}$ ovat poikkeamia keskiarvosta ja edelleen niille ovat voimassa aiemmat nollasummarajoitteet

$$\sum_i \lambda_i^X = \sum_j \lambda_j^Y = 0.$$

Myös parametrille λ_{ij}^{XY} asetetaan vastaavat rajoitteet

$$\sum_i \lambda_{ij}^{XY} = \sum_j \lambda_{ij}^{XY} = 0.$$

Nyt $I-1$ riviparametria ja $J-1$ sarakeparametria voidaan valita vapaasti, mutta rajoitteet määräävät parametrit viimeisellä rivillä ja sarakkeella. Yhteensä kyllästetyssä mallissa on nyt $1+(I-1)+(J-1)+(I-1)(J-1) = IJ$ vapausastetta, missä alkuosa $1+(I-1)+(J-1)$ kuvaa riippumattomuutta ja loppuosa $(I-1)(J-1)$ muuttujien X ja Y välistä vuorovaikutusta. Mallin (3.3) oikea puoli muistuttaakin yhdysvaikutuksen sallivaa kaksisuuntaista ANOVAa.

3.1.3 Parametrien tulkinta

Loglineaarinen malli ei erottele vaste- eikä selittäviä muuttujia toisistaan. Monissa tutkimuksissa tämä ulottuvuus on kuitenkin mukana, mikä saattaa vaikuttaa sopivimman mallin valintaan ja parametrien tulkintaan. Selittämiseen pyrkiviä ehdollisia malleja kutsutaan logit-malleiksi.

Tulkinta on yksinkertaisin kaksiluokkaisen vasteen tapauksessa, joten tarkastellaan $I \times 2$-taulua. Kaksiluokkaisen muuttujan Y logit-malli saadaan, kun ehdollistetaan I-luokkaisella muuttujalla X:

$$\begin{aligned} \log(\pi_{1|i}/\pi_{2|i}) &= \log(m_{i1}/m_{i2}) = \log(m_{i1}) - \log(m_{i2}) \\ &= (\mu + \lambda_i^X + \lambda_1^Y) - (\mu + \lambda_i^X + \lambda_2^Y) = \lambda_1^Y - \lambda_2^Y, \end{aligned}$$

missä $\{\pi_{j|i},\ j = 1, 2\}$ kuvaa muuttujan Y todennäköisyysjakaumaa muuttujan X tasolla i ja $\sum_j \pi_{j|i} = 1$. Suhteen logaritmi on siis sama joka rivillä. Koska parametreille on asetettu nollasummarajoitteet, on $\lambda_1^Y = -\lambda_2^Y$ ja tällöin

$$\log(\pi_{1|i}/\pi_{2|i}) = 2\lambda_1^Y.$$

Jokaisella rivillä on näin ollen $\exp(2\lambda_1^Y)$ kertaa todennäköisempää, että sarakkeen luokka on 1 kuin 2.

Ristitulosuhteen ja yhdysvaikutustermien välillä on suora yhteys loglineaaristen mallien tapauksessa. Yhteys on yksinkertaisin 2×2 -taululle. Tällöin kyllästetylle mallille ristitulosuhteen

$$\theta = \frac{m_{11}m_{22}}{m_{12}m_{21}}$$

logaritmi on

$$\begin{aligned} \log\theta &= \log((m_{11}m_{22})/(m_{12}m_{21})) = \log(m_{11}) + \log(m_{22}) - \log(m_{12}) - \log(m_{21}) \\ &= (\mu + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY}) + (\mu + \lambda_2^X + \lambda_2^Y + \lambda_{22}^{XY}) \\ &\quad -(\mu + \lambda_1^X + \lambda_2^Y + \lambda_{12}^{XY}) - (\mu + \lambda_2^X + \lambda_1^Y + \lambda_{21}^{XY}) \\ &= \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}. \end{aligned}$$

Koska nollasummarajoitteet ovat edelleen voimassa, on $\lambda_{11}^{XY} = \lambda_{22}^{XY} = -\lambda_{12}^{XY} = -\lambda_{21}^{XY}$ ja ristitulosuhteen logaritmi saa muodon

$$\log\theta = 4\lambda_{11}^{XY}. \tag{3.4}$$

Logit-malli yleistyy moniluokkaisen vasteen tapauksessa $I \times J$-taululle. Nyt ristitulosuhteita saadaan $J - 1$ kappaletta.

3.2 Kolmen muuttujan parittaiset vaikutukset

Laajennetaan tarkastelua koskemaan kolmen muuttujan yhteisjakaumaa. Havaintojen frekvenssit $\{n_{ij}\}$ noudattavat edelleen multinomijakaumaa, aivan kuten kaksiulotteisessa tapauksessa. Taulukossa 3 on esitetty kolmiulotteinen taulu, jossa plusmerkki

kuvaa jälleen summaa yli indeksin, jonka se korvaa. Kuten huomataan, käy jo kolmiulotteisen taulun esittäminen graafisesti melko haastavaksi. Kaikkien marginaalijakaumien esittäminen ei ole havainnollista ja lisäksi taulusta tulee hyvin kookas, jos muuttujilla on useita luokkia. Tässä osiossa tutustutaan käsitteisiin, joiden avulla voidaan tarkastella moniulotteisia ilmiöitä.

Taulukko 3: Kolmiulotteinen ristiintaulukko.

		1	2	...	J	Yht.
1	1	n_{111}	n_{121}	...	n_{1J1}	n_{1+1}
	2	n_{211}	n_{221}	...	n_{2J1}	n_{2+1}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	I	n_{I11}	n_{I21}	...	n_{IJ1}	n_{I+1}
2	1	n_{112}	n_{122}	...	n_{1J2}	n_{1+2}
	2	n_{212}	n_{222}	...	n_{2J2}	n_{2+2}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	I	n_{I12}	n_{I22}	...	n_{IJ2}	n_{I+2}
$\vdots$	$\vdots$					$\vdots$
K	1	n_{11K}	n_{12K}	...	n_{1JK}	n_{1+K}
	2	n_{21K}	n_{22K}	...	n_{2JK}	n_{2+K}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	I	n_{I1K}	n_{I2K}	...	n_{IJK}	n_{I+K}
	Yht.	n_{+1+}	n_{+2+}	...	n_{+J+}	n

3.2.1 Marginaali- ja osataulut

Olkoot kolme muuttujaa nimeltään X, Y ja Z. Kun esitetään muuttujien X ja Y yhteisjakauma muuttujan Z eri tasoilla, saadaan kokoelma osatauluja. Osataulussa muuttujan Z arvo pidetään vakiona eli sen vaikutus kontrolloidaan. Kun osataulut summataan yhteen, saadaan marginaalitaulu. Se jättää muuttujan Z täysin huomiotta.

Muuttujien väliset yhteydet saattavat näyttää täysin erilaisilta riippuen siitä, tarkastellaanko marginaali- vai osatauluja. Moniulotteisessa tapauksessa ei pitäisikään analysoida pelkkiä marginaalijakaumia, sillä niiden antamat tulokset saattavat olla harhaanjohtavia.

3.2.2 Riippumattomuuden tyypit

Kolmiulotteisessa taulussa muuttujilla X, Y ja Z voi olla useita riippuvuusrakennetyyppejä, jotka määräytyvät solutodennäköisyyksistä $\{\pi_{ijk}\}$. Merkitään solutodennäköisyyksiä $\{\pi_{ijk},\ i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, \ldots, K\}$, joille $\sum_i \sum_j \sum_k \pi_{ijk} = 1$.

Muuttujat X, Y ja Z ovat (keskenään) riippumattomia, kun

$$\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k} \quad \text{kaikille } i, j \text{ ja } k. \tag{3.5}$$

Tätä vastaa loglineaarinen malli

$$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z. \tag{3.6}$$

Muuttuja Y on riippumaton parista (X, Z), kun

$$\pi_{ijk} = \pi_{i+k}\pi_{+j+} \quad \text{kaikille } i, j \text{ ja } k. \tag{3.7}$$

Tätä vastaava loglineaarinen malli on

$$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}. \tag{3.8}$$

Samoin tietysti X voi olla riippumaton parista (Y, Z) tai Z parista (X, Y). Kun yhtälön (3.5) ehto on voimassa, on $\pi_{i+k} = \pi_{i++}\pi_{++k}$, joten myös yhtälö (3.7) pätee. Keskinäisestä riippumattomuudesta seuraa siis myös yhden muuttujan riippumattomuus kahden muun muodostamasta parista.

Tarkastellaan seuraavaksi muuttujien X ja Y välistä yhteyttä, kun muuttujan Z vaikutus kontrolloidaan. Jos X ja Y ovat riippumattomia muuttujan Z tasolla k muodostetussa osataulussa, sanotaan niiden olevan riippumattomia Z:n tasolla k. Kun yhtälö $\pi_{ij|k} = \pi_{ijk}/\pi_{++k}$ kuvaa muuttujien X ja Y yhteisjakaumaa Z:n tasolla k, on riippumattomuus muuttujan Z tasolla k

$$\pi_{ij|k} = \pi_{i+|k}\pi_{+j|k} \quad \text{kaikille } i \text{ ja } j. \tag{3.9}$$

Muuttujat X ja Y ovat ehdollisesti riippumattomia, kun Z on annettu, jos ne ovat riippumattomia muuttujan Z jokaisessa osataulussa. Tällöin siis yhtälö (3.9) pätee kaikille k ja yhtäpitävästi

$$\pi_{ijk} = \pi_{i+k}\pi_{+jk}/\pi_{++k} \quad \text{kaikille } i, j \text{ ja } k. \tag{3.10}$$

Muuttujien X ja Y ehdollista riippumattomuutta vastaava loglineaarinen malli on

$$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}. \tag{3.11}$$

Oletetaan, että Y on kokonaan riippumaton parista (X, Z), jolloin $\pi_{ijk} = \pi_{i+k}\pi_{+j+}$. Tällöin $\pi_{ij|k} = \pi_{ijk}/\pi_{++k} = \pi_{i+k}\pi_{+j+}/\pi_{++k}$. Kun summataan tässä yli i:n kummallakin puolella, saadaan $\pi_{+j|k} = \pi_{+j+}$. Siksi $\pi_{ij|k} = (\pi_{i+k}/\pi_{++k})\pi_{+j+} = \pi_{i+|k}\pi_{+j|k}$, joten muuttujat X ja Y ovat myös ehdollisesti riippumattomat ehdolla Z. Samoin Y ja Z ovat ehdollisesti riippumattomat ehdolla X. Kaiken kaikkiaan siis muuttujien keskinäisestä riippumattomuudesta seuraa muuttujan Y yhteinen riippumattomuus muuttujista X ja Z, josta taas seuraa muuttujien X ja Y ehdollinen riippumattomuus ehdolla Z.

3.2.3 Marginaalinen ja ehdollinen riippumattomuus

Oletetaan, että Y on kokonaan riippumaton parista (X, Z) eli $\pi_{ijk} = \pi_{i+k}\pi_{+j+}$. Edellä todistettiin, että tästä seuraa muuttujien X ja Y välinen ehdollinen riippumattomuus ehdolla Z. Kun summataan kummallakin puolella yli k:n, saadaan $\pi_{ij+} = \pi_{i++}\pi_{+j+}$. Näin ollen yhteisestä riippumattomuudesta seuraa marginaalinen riippumattomuus. Koska kolmen muuttujan keskinäisestä riippumattomuudesta seuraa kahden muuttujan yhteinen riippumattomuus kolmannesta, seuraa tästä siis myös marginaalinen ja ehdollinen riippumattomuus. Ehdollisesta riippumattomuudesta ei sen sijaan seuraa marginaalinen riippumattomuus, kuten seuraavasta esimerkistä ilmenee.

Esimerkki 3.1. *Olkoot X, Y ja Z kaksiluokkaisia satunnaismuuttujia. Tarkastellaan muuttujien X ja Y yhteisjakaumaa muuttujan Z kummassakin luokassa erikseen. Nyt X ja Y ovat riippumattomia muuttujan Z osatauluissa, sillä $m_{ij|k} = m_{i+|k}m_{+j|k}/m_{++|k}$ kaikille i ja j, esimerkiksi $m_{1+|1}m_{+2|1}/m_{++|1} = 700 \cdot 900/1000 = 630 = m_{12|1}$. Kuitenkaan X ja Y eivät ole riippumattomat muuttujan Z marginaalitaulussa, sillä esimerkiksi $m_{1+|+}m_{+2|+}/m = 800 \cdot 1000/1500 = 533.33 \neq 650 = m_{12|+}$.*

Z = 1

		Y		
		1	2	Yht.
X	1	70	630	700
	2	30	270	300
	Yht.	100	900	1000

Z = 2

		Y		
		1	2	Yht.
X	1	80	20	100
	2	320	80	400
	Yht.	400	100	500

Z:n marginaalijakauma

		Y		
		1	2	Yht.
X	1	150	650	800
	2	350	350	700
	Yht.	500	1000	1500

3.2.4 Parittaiset yhdysvaikutukset

Kun kaikki kolme muuttujaa X, Y ja Z ovat pareittain ehdollisesti riippuvia, päädytään parittaisten vaikutusten malliin

$$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}. \tag{3.12}$$

Tämä voidaan esittää yhtäpitävästi todennäköisyyksien avulla, jolloin solutodennäköisyydet saavat muodon

$$\pi_{ijk} = \psi_{ij}\phi_{jk}\omega_{ik}.$$

Nyt parametreille ei ole olemassa esitystä suljetussa muodossa joitain erikoistapauksia lukuunottamatta.

Kuten seuraavassa luvussa esitetään, kuvaa malli (3.12) tapausta, jossa minkä tahansa kahden muuttujan ristitulosuhteet ovat samat kaikilla kolmannen muuttujan tasoilla. Kun vasteen ja selittäjien tapauksessa sanotaan, ettei yhdysvaikutusta ole, tarkoitetaan tällä sitä, että selittävän muuttujan vaikutus vasteeseen pysyy samana kaikilla muiden muuttujien tasoilla.

3.3 Loglineaarinen malli kolmiulotteiselle taululle

Tässä osiossa tarkastellaan tarkemmin loglineaarisia malleja kolmiulotteiselle ristiintaulukolle.

3.3.1 Hierarkkiset mallit

Oletetaan, että $m_{ijk} > 0$ ja $\eta_{ijk} = \log m_{ijk}$. Piste indeksissä tarkoittaa jälleen keskiarvoa kyseisen indeksin suhteen eli esimerkiksi $\eta_{.jk} = (\sum_i \eta_{ijk})/I$. Asetetaan

$$\begin{aligned}
&\mu = \eta_{...},\\
&\lambda_i^X = \eta_{i..} - \eta_{...}, \quad \lambda_j^Y = \eta_{.j.} - \eta_{...}, \quad \lambda_k^Z = \eta_{..k} - \eta_{...},\\
&\lambda_{ij}^{XY} = \eta_{ij.} - \eta_{i..} - \eta_{.j.} + \eta_{...}, \quad \lambda_{ik}^{XZ} = \eta_{i.k} - \eta_{i..} - \eta_{..k} + \eta_{...},\\
&\lambda_{jk}^{YZ} = \eta_{.jk} - \eta_{.j.} - \eta_{..k} + \eta_{...},\\
&\lambda_{ijk}^{XYZ} = \eta_{ijk} - \eta_{ij.} - \eta_{i.k} - \eta_{.jk} + \eta_{i..} + \eta_{.j.} + \eta_{..k} - \eta_{...}.
\end{aligned}$$

Parametrien summa yli minkä tahansa indeksin on nolla eli

$$\sum_i \lambda_i^X = \sum_j \lambda_j^Y = \sum_k \lambda_k^Z = \sum_i \lambda_{ij}^{XY} = \sum_j \lambda_{ij}^{XY} = \cdots = \sum_k \lambda_{ijk}^{XYZ} = 0.$$

Nyt yleinen loglineaarinen malli kolmiulotteiselle ristiintaulukolle on

$$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}. \tag{3.13}$$

Kun asetetaan yhtälössä (3.13) joitain parametreja nollaksi, päädytään edellisessä luvussa esitettyihin malleihin. Taulukossa 4 on listattu joitain näistä malleista. Ne ovat hierarkkisia: jos mukana on korkeamman asteen termejä, sisältyvät mukaan myös näitä vastaavat alemman asteen termit. Jos malli siis sisältää esimerkiksi termin λ_{ij}^{XY}, tulee mukaan ottaa myös termit λ_i^X ja λ_j^Y. Hierarkkisia malleja käytetään, koska todennäköisyyksien avulla määritellyt riippuvuudet ovat hierarkkisia. Esimerkki epähierarkkisesta mallista on

$$\log m_{ij} = \mu + \lambda_i^X + \lambda_{ij}^{XY}.$$

Tämä hyväksyy yhdysvaikutuksen muuttujien X ja Y välillä, mutta pakottaa keskimääräisen odotetun frekvenssin logaritmin riveillä olemaan sama joka sarakkeessa. Epähierarkkisia malleja käytetään joissain erikoistapauksissa.

Taulukko 4: Joitakin loglineaarisia malleja kolmiulotteiselle taululle

Loglineaarinen malli	Merkintä
$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$	$X + Y + Z$
$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}$	$X * Y + Z$
$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}$	$X * Y + Y * Z$
$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ}$	$X * Y + Y * Z + X * Z$
$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ} + \lambda_{ijk}^{XYZ}$	$X * Y * Z$

3.3.2 Parametrien tulkinta

Tulkittaessa loglineaarisia malleja kuvaillaan niiden marginaalisia ja osittaisia assosiaatioita ristitulosuhteiden avulla. Muuttujien X ja Y välisessä marginaalitaulussa $\{\pi_{ij+}\}$ marginaalisen riippuvuuden kuvaamiseen tarvitaan $(I-1)(J-1)$ ristitulosuhdetta, joista esimerkkinä

$$\theta_{ij}^{XY} = \frac{\pi_{ij+}\pi_{i+1,j+1,+}}{\pi_{i+1,j,+}\pi_{i,j+1,+}}, \quad 1 \leq i \leq I-1,\ 1 \leq j \leq J-1.$$

Kun muuttujat X ja Y ovat riippumattomia, pätee $\theta_{ij}^{XY} = 1$ kaikilla $1 \leq i \leq I$ ja $1 \leq j \leq J$. Kun muuttujan Z taso k kiinnitetään, kuvaa vastaava ristitulosuhde

$$\theta_{ij(k)}^{XY} = \frac{\pi_{ijk}\pi_{i+1,j+1,k}}{\pi_{i+1,j,k}\pi_{i,j+1,k}}, \quad 1 \leq i \leq I-1,\ 1 \leq j \leq J-1,$$

muuttujien X ja Y välistä ehdollista riippuvuutta ehdolla $Z = k$. Vastaavasti muuttujien X ja Z välistä ehdollista riippuvuutta ehdolla $Y = j$ kuvaa $(I-1)(K-1)$ ristitulosuhdetta $\{\theta_{i(j)k}\}$ muuttujan Y J:llä tasolla ja muuttujien Y ja Z välistä ehdollista riippuvuutta ehdolla $X = i$ kuvaa $(J-1)(K-1)$ ristitulosuhdetta $\{\theta_{(i)jk}\}$ muuttujan X I:llä tasolla.

Loglineaarisen mallin parametrit ovat ehdollisten ristitulosuhteiden funktioita. Yhteys on yksinkertaisin kaksiluokkaisille muuttujille. Kun $2\times2\times2$-kokoisessa taulussa sijoitetaan yhtälö (3.13) $\log m_{ijk}$:n paikalle ehdollisissa ristitulosuhteissa, saadaan

$$\lambda_{111}^{XYZ} = \frac{1}{8}\log\left(\frac{\theta_{11(1)}}{\theta_{11(2)}}\right) = \frac{1}{8}\log\left(\frac{\theta_{1(1)1}}{\theta_{1(2)1}}\right) = \frac{1}{8}\log\left(\frac{\theta_{(1)11}}{\theta_{(2)11}}\right), \tag{3.14}$$

kun parametrille $\{\lambda_{ijk}^{XYZ}\}$ on voimassa nollasummarajoitteet. Jokainen λ_{ijk}^{XYZ}-termi on nolla, kun kahden muuttujan välinen ristitulosuhde on sama jokaisella kolmannen muuttujan tasolla. Yleisemmin $\lambda_{ijk}^{XYZ} = 0$ kaikille i, j ja k $2 \times 2 \times K$-taulussa, kun $\theta_{11(1)} = \cdots = \theta_{11(K)}$, jolloin

$$\lambda_{11}^{XY} = \frac{1}{4}\log\theta_{11(k)}, \quad \text{kun } k = 1, \ldots, K. \tag{3.15}$$

Kuten kaksiulotteisessa tapauksessa (ks. (3.4)), yhdysvaikutusparametri on suoraan verrannollinen ristitulosuhteen logaritmiin.

Loglineaarisia malleja voidaan kuvata käyttämällä ehdollisia ristitulosuhteita. Esimerkiksi muuttujien X ja Y välistä ehdollista riippumattomuutta vastaa $\theta_{ij(k)} = 1$ kaikilla $i = 1, \ldots, I,\ j = 1, \ldots, J$ ja $k = 1, \ldots, K$.

3.3.3 Marginaalinen ja osittainen riippuvuus

Tarkastellaan seuraavaksi ehtoja, joiden voimassa ollessa XY-taulun ristitulosuhteet ovat samat marginaali- ja osatauluissa. Niiden yhtäläisyydestä on hyötyä, sillä tällöin voidaan tutkia muuttujien X ja Y assosiaatiota yksinkertaisemmalla tavalla, jossa vakioidaan yli muuttujan Z, joka voi olla yksi- tai moniulotteinen. Ristitulosuhteen vakiointiehtoja (collapsibility conditions)

$$\theta_{ij}^{XY} = \theta_{ij(1)} = \cdots = \theta_{ij(K)}, \quad 1 \le i \le I-1,\ 1 \le j \le J-1 \tag{3.16}$$

voidaan käyttää, jos toinen tai kummatkin seuraavista ehdoista ovat voimassa:

$$\theta_{i(j)k} = 1, \quad 1 \le i \le I-1,\ 1 \le j \le J,\ 1 \le k \le K-1,$$

$$\theta_{(i)jk} = 1, \quad 1 \le i \le I,\ 1 \le j \le J-1,\ 1 \le k \le K-1.$$

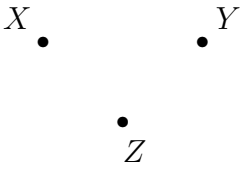
Toisin sanoen muuttujien X ja Y marginaaliset ja osittaiset assosiaatiot ovat identtiset, mikäli joko Z ja X ovat ehdollisesti riippumattomat tai Z ja Y ovat ehdollisesti riippumattomat.

Ristitulosuhteen vakiointiehtojen todistus käy suoraviivaisesti (Bishop, 1971). Siinä näytetään, että kun malli $X * Y + Y * Z$ tai $X * Y + X * Z$ pätee, on odotettujen frekvenssien $\{m_{ij+}\}$ loglineaarisella mallilla samat yhdysvaikutusparametrit kuin mitä parametri $\{\lambda_{ij}^{XY}\}$ on edellä mainituissa malleissa, ja siten näillä on myös samat ristitulosuhteet.

3.3.4 Mallien tulkinta

Tarkastellaan nyt Taulukossa 4 listattuja loglineaarisia malleja kuvien ja ristitulosuhteen vakiointiehtojen avulla sekä tulkitaan mallit käyttämällä vastaavia logit-malleja.

X + Y + Z



Kaikki kolme muuttujaa X, Y ja Z ovat riippumattomat mallissa

$$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z. \tag{3.17}$$

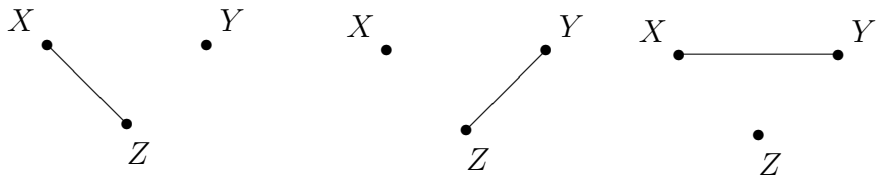
Luvussa 3.2 osoitettiin, että jokainen muuttujapari on myös ehdollisesti ja marginaalisesti riippumaton. Esimerkiksi muuttujien X ja Y marginaalinen assosiaatio on sama kuin niiden osittainen assosiaatio (kun muuttujan Z vaikutus on vakioitu), koska Z on ehdollisesti riippumaton muuttujasta X (kun muuttujan Y vaikutus on vakioitu) tai myös koska Z on ehdollisesti riippumaton muuttujasta Y (kun muuttujan X vaikutus on vakioitu).

Oletetaan nyt, että Y on kaksiluokkainen muuttuja. Kun muuttuja Y ajatellaan vasteeksi, on mallia (3.17) vastaava logit-malli

$$\log\left(m_{i1k}/m_{i2k}\right) = \lambda_1^Y - \lambda_2^Y.$$

Nollasummarajoitteiden ollessa voimassa saadaan tästä $2\lambda_1^Y$, ja tällöin $\exp\left(2\lambda_1^Y\right)$ on veto m_{i1k}/m_{i2k} jokaiselle muuttujien X ja Z tasojen yhdistelmälle. Kun muuttujalla Y on enemmän kuin kaksi tasoa, on $\log\left(m_{iak}/m_{ibk}\right) = \lambda_a^Y - \lambda_b^Y$. Mitä suurempi λ_j^Y:n arvo on, sitä enemmän havaintoja Y:n j:nnellä tasolla suhteessa on.

$\mathbf{X * Z + Y}$ tai $\mathbf{Y * Z + X}$ tai $\mathbf{X * Y + Z}$



Kolmen muuttujan tapauksessa on olemassa kolme mallia, joissa vain yksi muuttujapari on ehdollisesti riippuva. Merkintä $X * Z + Y$ tarkoittaa mallia

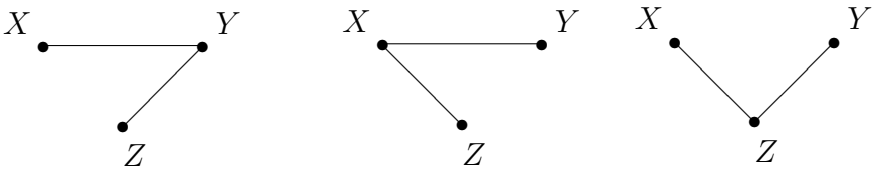
$$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}, \tag{3.18}$$

missä Y on yhteisesti riippumaton muuttujista X ja Z. Silloin luvun 3.2 perusteella Y ja X ovat ehdollisesti riippumattomia, kun muuttujan Z vaikutus on vakioitu, ja Y ja Z ovat ehdollisesti riippumattomia, kun muuttujan X vaikutus on vakioitu. Mallin merkintätapa kuvaa muuttujien X ja Z välistä ehdollista riippuvuutta, ja samaa riippuvuutta kuvaa myös parametri $\{\lambda_{ik}^{XZ}\}$.

Luvun 3.2.3 perusteella myös Y on riippumaton muuttujista X ja Z vastaavissa marginaalitauluissa XY ja YZ. Lisäksi ristitulosuhteen vakiointiehdoista seuraa, että XZ-marginaalitaulun vetosuhteet ovat samat kuin vastaavan XZ-osataulun vetosuhteet, koska Y on riippumaton muuttujasta X (kun Z:n vaikutus on vakioitu).

Kun muuttuja Y ajatellaan vasteeksi, vastaa loglineaarista mallia $X * Z + Y$ logit-malli, jossa vaste ei riipu kummastakaan selittäjästä.

X ∗ Y + Y ∗ Z tai X ∗ Y + X ∗ Z tai X ∗ Z + Y ∗ Z



Kolmen muuttujan tapauksessa on olemassa kolme mallia, joissa vain yksi muuttujapari on ehdollisesti riippumaton. Merkintä $X * Z + Y * Z$ tarkoittaa mallia, jossa X ja Y ovat ehdollisesti riippumattomia, kun muuttujan Z vaikutus on vakioitu, eli

$$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}. \tag{3.19}$$

Parametrit $\{\lambda_{ik}^{XZ}\}$ ja $\{\lambda_{jk}^{YZ}\}$ liittyvät vastaaviin X:n ja Z:n sekä Y:n ja Z:n osittaisiin yhdysvaikutuksiin.

Luvussa 3.3.3 osoitettiin, että X ja Y voivat olla marginaalisesti riippuvia, vaikka ne olisivatkin ehdollisesti riippumattomia. Tämäkin seuraa ristitulosuhteen vakiointiehdoista, koska Z on ehdollisesti riippuva kummastakin muuttujasta X ja Y. Samojen ehtojen perusteella kuitenkin XZ- ja YZ-marginaalitaulut antavat samat ristitulosuhteet kuin vastaavat osataulut, koska X ja Y ovat ehdollisesti riippumattomia.

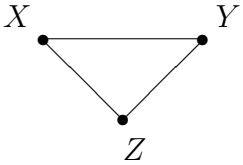
Kun vaste Y on kaksiluokkainen, vastaa loglineaarista mallia $X * Z + Y * Z$ logitmalli Z, joka on muotoa

$$\log (m_{i1k}/m_{i2k}) = (\lambda_1^Y - \lambda_2^Y) + (\lambda_{1k}^{YZ} - \lambda_{2k}^{YZ}).$$

Nollasummarajoitteiden ollessa voimassa saadaan $\log (m_{i1k}/m_{i2k}) = 2\lambda_1^Y + 2\lambda_{1k}^{YZ}$. Logit-malli voidaan esittää muodossa $\alpha + \beta_k^Z$, kun $2\lambda_{1k}^{YZ}$ on muuttujan Z k:nnen tason vaikutus (eli β_k^Z) ja $2\lambda_1^Y = \alpha$. Logit-malli riippuu Z:n tasosta k muuttujien Y ja Z yhdysvaikutustermin kautta, mutta ei riipu X:n tasosta i.

Malli $X * Z + Y * Z$ on erittäin tärkeä kausaalitulkinnan kannalta. Muuttujien X ja Y välillä havaittu yhdysvaikutus saattaa olla valheellinen, mikäli löydetään kolmas muuttuja Z siten, että yhdysvaikutus häviää, kun sen vaikutus vakioidaan.

X ∗ Y + X ∗ Z + Y ∗ Z



Mallissa

$$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \tag{3.20}$$

on yhdysvaikutustermit jokaiselle muuttujaparille, joten yksikään pari ei ole ehdollisesti riippumaton. Tätä kutsutaan parittaisten vaikutusten malliksi, eikä sillä ole

tulkintaa riippumattomuuden tai ehdollisen riippumattomuuden avulla. Ristitulosuhteen vakiointiehdot eivät ole voimassa, joten marginaaliset ristitulosuhteet saattavat erota osittaisista ristitulosuhteista kaikilla muuttujapareilla.

Sijoittamalla (3.20) ehdollisen ristitulosuhteen logaritmin $\log \theta_{ij(k)}$ lausekkeeseen saadaan

$$\log \theta_{ij(k)} = \lambda_{ij}^{XY} + \lambda_{i+1,j+1}^{XY} - \lambda_{i,j+1}^{XY} - \lambda_{i+1,j}^{XY}.$$

Koska yhtälön oikea puoli on sama kaikille k, kolmen muuttujan yhdysvaikutustermin puuttumisesta seuraa

$$\theta_{ij(1)} = \theta_{ij(2)} = \cdots = \theta_{ij(K)} \quad \text{kaikille } i \text{ ja } j.$$

Samoin perustein toisista osittaisista ristitulosuhteista saadaan, että kolmen muuttujan yhdysvaikutustermin puuttumisesta seuraa

$$\theta_{i(1)k} = \theta_{i(2)k} = \cdots = \theta_{i(J)k} \quad \text{kaikille } i \text{ ja } k$$

ja

$$\theta_{(1)jk} = \theta_{(2)jk} = \cdots = \theta_{(I)jk} \quad \text{kaikille } j \text{ ja } k.$$

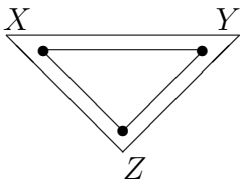
Kun kolmen muuttujan yhdysvaikutustermi puuttuu, on kahden muuttujan välinen assosiaatio siis sama kaikilla kolmannen muuttujan tasoilla.

Kun vaste Y on kaksiluokkainen ja nollasummarajoitteet ovat voimassa, yksinkertaistuu loglineaarista mallia $X*Y+X*Z+Y*Z$ vastaava logit-malli $X+Z$ luvunn 3.4.3 perusteella muotoon

$$\log (m_{i1k}/m_{i2k}) = 2\lambda_1^Y + 2\lambda_{i1}^{XY} + 2\lambda_{1k}^{YZ}.$$

Logit-malli voidaan esittää myös muodossa $\alpha + \beta_i^X + \beta_k^Z$, jolloin vaste riippuu sekä X:n että Z:n tasosta, mutta additiivisesti. Muuttujan X vaikutus on sama kaikilla Z:n tasoilla ja Z:n vaikutus on sama kaikilla X:n tasoilla.

X ∗ Y ∗ Z



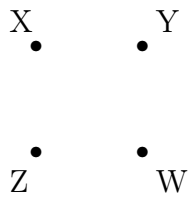
Yleinen malli kolmelle muuttujalle (ks. (3.13)) sallii kolmen muuttujan yhdysvaikutuksen. Jokainen muuttujapari saa olla ehdollisesti riippuva ja minkä tahansa parin ristitulosuhde saa vaihdella kolmannen muuttujan tasoilla. Tämä malli kuvaa koko joukon $\{m_{ijk}\}$, joka saa positiivisia arvoja. Koska $\sum_i \lambda_i^X = 0$, on lineaarisesti riippumattomia $\{\lambda_i^X\}$-parametreja $I-1$ kappaletta, ja koska $\sum_i \lambda_{ij}^{XY} = \sum_j \lambda_{ij}^{XY} = 0$, on lineaarisesti riippumattomia $\{\lambda_{ij}^{XY}\}$-parametreja $(I-1)(J-1)$ kappaletta. Tämä pätee yhtäpitävästi muille parametreille. Yhteensä lineaarisesti riippumattomia parametreja on (μ mukaanlukien) $1+(I-1)+(J-1)+(K-1)+(I-1)(J-1)+(I-1)(K-1)+(J-1)(K-1)+(I-1)(J-1)(K-1) = IJK$, mikä vastaa solujen määrää ristiintaulukossa. Malli on kyllästetty.

3.4 Loglineaarinen malli useammalle muuttujalle

Mallit kolmiulotteisille tauluille ovat monimutkaisempia kuin kaksiulotteisille, koska mahdollisten yhdysvaikutusten määrä on suurempi. Ulottuvuuksien määrän kasvu tuokin mukanaan joitain ongelmia erityisesti siinä tapauksessa, kun muuttujilla on useita luokkia. Ensinnäkin mahdollisten yhdysvaikutusten yhdistelmien määrä kasvaa valtavasti ja toisekseen analyysia saattaa vaikeuttaa solujen määrän suuri lisääntyminen. Jollei otoskoko ole suuri, voi taulussa olla paljon havaittuja nollafrekvenssejä, mistä saattaa seurata ongelmia asymptotiikan kannalta.

3.4.1 Neliulotteinen taulu

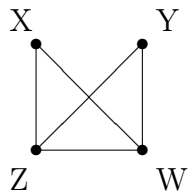
Tarkastellaan esimerkkinä moniulotteisesta taulusta neliulotteista tapausta, jossa on muuttujat W, X, Y ja Z. Tulkinnan helpottamiseksi on toivottavaa, että mallissa olisi mahdollisimman vähän korkeamman asteen termejä. Lopulliseksi malliksi pyritään siis valitsemaan mahdollisimman vähäparametrinen malli.



Kuva 3.1: Kaikki muuttujat riippumattomia

Riippumattomuusmalli $W+X+Y+Z$ (kuva 3.1) on niin yksinkertainen, että sillä on harvoin paljonkaan merkitystä käytännössä. Kaikki mallit, joilla ei ole kolmen muuttujan yhdysvaikutusta, ovat sisäkkäisiä mallille

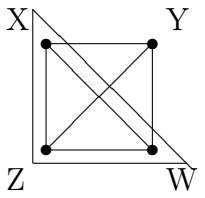
$$\begin{aligned}\log m_{ijk} &= \mu + \lambda_h^W + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{hi}^{WX} + \lambda_{hj}^{WY} \\ &\quad + \lambda_{hk}^{WZ} + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}.\end{aligned}$$



Kuva 3.2: X ja Y ehdollisesti riippumattomia

Sitä merkitään $W*X+W*Y+W*Z+X*Y+X*Z+Y*Z$. Tässä mallissa ehdollinen riippumattomuus vastaa tiettyjen kahden muuttujan yhdysvaikutusten puuttumista. Jos esimerkiksi $\lambda_{ij}^{XY}=0$ kaikille i ja j, ovat muuttujat X ja Y ehdollisesti riippumattomia kaikilla muuttujien W ja Z tasojen yhdistelmillä eli ehdollisesti riippumattomia ehdolla Z ja W (Kuva 3.2).

Neljän muuttujan tapauksessa on suuri määrä malleja, joissa on jonkinlainen kolmen muuttujan yhdysvaikutus. Esimerkiksi mallissa $W*Y+X*Y+Y*Z+W*X*Z$ (Kuva 3.3) jokaisella muuttujan Y tasolla assosiaatio W:n ja X:n, W:n ja Z:n tai X:n ja Z:n välillä vaihtelee jäljelle jäävän muuttujan eri tasoilla. Osittainen assosiaatio Z:n ja jonkun toisen muuttujan välillä on sama jokaisella kahden muun muuttujan tason yhdistelmällä.



Kuva 3.3: Kolmen muuttujan yhdysvaikutus

3.4.2 Vakiointi

Pelkistettäessä moniulotteisia tauluja on vakiointi oleellinen menetelmä, sillä vakioitaessa yli yhden tai useamman muuttujan helpottuu tällaisen taulun analysoiminen.

Edellä annettiin ehdot, joiden voimassa ollessa osittaiset ja marginaaliset assosiaatiot ovat yhtäsuuret. Bishopin, Fienbergin ja Hollandin (1975) mukaan teoreema antaa ehdot vakioinnille moniulotteisen taulun tapauksessa:

"Oletetaan, että T-ulotteisessa taulussa on jaettu muuttujat kolmeen toisensa poissulkevaan ryhmään. Jos kaikki mallin termit, jotka yhdistävät ensimmäisen ryhmän muuttujat toisen ryhmän muuttujiin, ovat nollia, eivät mallin ensimmäisen ryhmän termit muutu vakioitaessa yli toisen ryhmän."

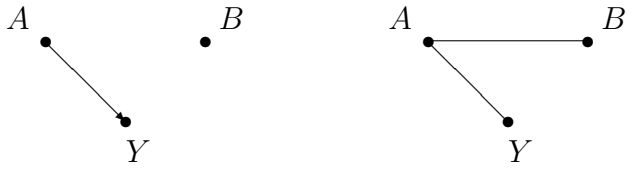
Havainnollistetaan tätä tulosta käyttämällä neliulotteista taulua, jossa W ja X muodostavat ensimmäisen muuttujaryhmän, Y toisen ryhmän ja Z kolmannen ryhmän. Jos parametrit λ_{hj}^{WY} ja λ_{ij}^{XY} ovat nollia, pysyvät muuttujien W ja X väliset osittaiset assosiaatiot samoina, kun vakioidaan taulussa yli muuttujan Y. Tämä tapahtuu mallille $Y * Z + W * X * Y$ tai yksinkertaisemmille malleille. Oletetaan seuraavaksi, että W ja X ovat ensimmäisessä ryhmässä, Y ja Z toisessa ryhmässä ja että kolmas ryhmä on tyhjä. Mallissa $W * X + Y * Z$ W:n ja X:n välinen osittainen assosiaatio on sama kuin W:n ja X:n välinen marginaalinen assosiaatio.

Tämä tulos tarkoittaa myös sitä, että kun joku muuttuja on riippumaton kaikista muista muuttujista, voidaan taulussa vakioida sen yli vaikuttamatta muihin mallin parametreihin. Esimerkiksi assosiaatiot W:n, X:n ja Y:n välillä ovat samat mallissa $W * X + W * Y + X * Y + Z$ ja mallissa $W * X + W * Y + X * Y$.

Kun kaikki kahden muuttujan väliset yhdysvaikutukset ovat mukana mallissa, saattaa minkä tahansa muuttujan yli vakioiminen aiheuttaa näiden muuttumisen. Mallissa $W * X + W * Y + W * Z + X * Y + X * Z + Y * Z$ saattavat jokaisen muuttujaparin osittaiset ja marginaaliset assosiaatiot erota.

3.4.3 Loglineaarisen ja logit-mallin välinen yhteys

Monissa sovelluksissa mahdollisesti hyödyllisten mallien joukko on suppea. Useat tutkimukset esimerkiksi tekevät eron selittävien ja selitettävien muuttujien välille, jolloin on tärkeämpää tutkia selittävien muuttujien vaikutusta vasteeseen kuin tarkastella näiden keskinäisiä riippuvuuksia. Logit-malleiksi kutsuttu loglineaaristen mallien osajoukko palvelee tätä tarkoitusta. Kun logit-malli halutaan ilmaista loglineaarisena, tulee siihen samat vasteen ja selittäjien yhdysvaikutukset kuin logit-mallissakin, ja lisäksi kaikki selittävien muuttujien väliset yhdysvaikutukset. Tästä esimerkkinä on havainnollistettu graafien avulla toisiaan vastaavat logit-malli A ja loglineaarinen malli $A * B + A * Y$ kuvassa 3.4.



Kuva 3.4: Toisiaan vastaavat logit-malli A ja loglineaarinen malli $A * B + A * Y$

Havainnollistukseksi tarkastellaan $I \times J \times 2$ -kokoista taulua. Logit-mallia $A + B$

vastaava loglineaarinen malli on

$$\log\left(\frac{m_{ij1}}{m_{ij2}}\right) = \alpha + \beta_i^A + \beta_j^B. \tag{3.21}$$

Tässä mallissa vasteella Y on yhdysvaikutus selittävien muuttujien A ja B kanssa, mutta kummankin muuttujan vaikutus on sama kaikilla toisen muuttujan tasoilla. Loglineaarinen malli sisältää yhdysvaikutustermit λ_{ik}^{AY} ja λ_{jk}^{BY} ja selittäjien yhdysvaikutustermin λ_{ij}^{AB}. Malli on siis $A*B+A*Y+B*Y$.

Jotta voitaisiin osoittaa, että loglineaarinen malli $A*B+A*Y+B*Y$ on sama kuin logit-malli (3.21), huomataan, että loglineaariselle mallille

$$\begin{aligned}
\log\left(\frac{m_{ij1}}{m_{ij2}}\right) &= \log(m_{ij1}) - \log(m_{ij2}) \\
&= [\mu + \lambda_i^A + \lambda_j^B + \lambda_1^Y + \lambda_{ij}^{AB} + \lambda_{i1}^{AY} + \lambda_{i1}^{BY}] \\
&\quad -[\mu + \lambda_i^A + \lambda_j^B + \lambda_2^Y + \lambda_{ij}^{AB} + \lambda_{i2}^{AY} + \lambda_{i2}^{BY}] \\
&= (\lambda_1^Y - \lambda_2^Y) + (\lambda_{i1}^{AY} - \lambda_{i2}^{AY}) + (\lambda_{j1}^{BY} - \lambda_{j2}^{BY}).
\end{aligned}$$

Nollasummarajoitteiden ollessa voimassa eli kun $\sum_k \lambda_k^Y = \sum_k \lambda_{ik}^{AY} = \sum_k \lambda_{jk}^{BY} = 0$, saadaan $\lambda_1^Y = -\lambda_2^Y$, $\lambda_{i1}^{AY} = -\lambda_{i2}^{AY}$ ja $\lambda_{j1}^{BY} = -\lambda_{j2}^{BY}$, koska vasteella Y on kaksi tasoa. Logit-malli yksinkertaistuu siis muotoon

$$\log\left(\frac{m_{ij1}}{m_{ij2}}\right) = 2\lambda_1^Y + 2\lambda_{i1}^{AY} + 2\lambda_{j1}^{BY}.$$

Tämä on sama malli kuin (3.21), kun $2\lambda_{i1}^{AY}$ on muuttujan A tason i vaikutus vasteen Y logit-muunnokseen (eli β_i^A), $2\lambda_{j1}^{BY}$ on muuttujan B tason j vaikutus vasteen Y logit-muunnokseen (eli β_j^B) ja $2\lambda_1^Y = \alpha$.

Selittävien muuttujien yhdysvaikutustermit λ_{ij}^{AB} supistuvat logaritmien erotuksessa. Logit-malli ei siis sisällä tietoa niiden riippuvuuksista.

Tarkastellaan seuraavaksi neliulotteista tapausta, jossa Y on kaksiluokkainen vaste. Logit-malli $A+B+C$ eli

$$\log\left(\frac{m_{ij1}}{m_{ij2}}\right) = \alpha + \beta_h^A + \beta_i^B + \beta_j^C \tag{3.22}$$

sisältää selittäjien päävaikutustermit, mutta ei näiden yhdysvaikutuksia. Tämä malli vastaa loglineaarista mallia, jossa on kaikkien selittäjien yhdysvaikutus vasteen Y kanssa sekä kaikki näiden keskinäiset yhdysvaikutukset, ts. mallia $A*Y+B*Y+C*Y+A*B*C$. Yhtälön (3.22) erikoistapaus $A+B$, jossa kaikki termit β_j^C ovat nollia, vastaa loglineaarista mallia $A*Y+B*Y+A*B*C$, jossa Y ja C ovat ehdollisesti riippumattomia, kun A ja B on annettu. Yhtälön (3.22) yleistys $C+A*B$, joka sisältää myös termin β_{hi}^{AB}, vastaa loglineaarista mallia $C*Y+A*B*Y+A*B*C$, jossa A:n ja B:n välillä on yhdysvaikutus vasteeseen Y.

3.5 Kvasiloglineaariset mallit

Ristiintaulukoita tarkasteltaessa saattaa vastaan tulla tyhjiä soluja. Niitä voi olla kahta tyyppiä: otannasta johtuvia ja rakenteellisia nollia. Näistä ensimmäisellä on positiivinen odotusarvo, luultavasti suhteellisen pieni sellainen, ja otoskokoa kasvattamalla solun odotetusta nollafrekvenssistä pitäisi periaatteessa päästä eroon. Sen sijaan rakenteellinen nolla on kyseessä siinä tapauksessa, kun solun odotettu frekvenssi on nolla eli siihen ei ole mahdollistakaan saada havaintoja. Rakenteellisia nollia sisältäviä tauluja kutsutaan epätäydellisiksi.

Epätäydellisten taulujen analysoinnissa käytetään kvasiloglineaarisia malleja. Malli on loglineaarinen, mutta se määritellään vain niille soluille, joiden odotusarvo poikkeaa nollasta. Mallin tarkastelu on oleellista ylioppilasaineiston tapauksessa, koska tehtävänantojen luokittelujen takia siitä muodostetut ristiintaulukot saattavat sisältää runsaasti rakenteellisia nollia. Tarkastellaan aluksi epätäydellisiä kaksisuuntaisia tauluja ja kvasiriippumattomuusmalleja. Luvun teoria perustuu Bishopin, Fienbergin ja Hollandin oppikirjaan *Discrete multivariate analysis* (1975).

3.5.1 Epätäydellinen kaksisuuntainen taulu

Tarkastellaan $I \times J$-kokoista taulua, jonka otoskoko on N. Olkoon S niiden solujen joukko, jotka eivät sisällä rakenteellista nollaa. Oletetaan lisäksi, että kaikille $m_{ij} \in S$ pätee riippumattomuusoletus $m_{ij} = a_i b_j$, missä a_i ja b_j ovat positiivisia vakioita. Tässä tapauksessa sanotaan, että rivit ja sarakkeet ovat kvasiriippumattomia. Koska ne solut, jotka eivät kuulu joukkoon S, sisältävät rakenteellisia nollia (eli $m_{ij} = 0$ kaikille $(i, j) \notin S$), ovat epätäydellisen taulun marginaalijakaumat myös joukon S marginaalijakaumia.

Odotettujen solufrekvenssien malleja epätäydellisille tauluille voidaan myös tarkastella loglineaaristen mallien kehikossa. Yleisessä tapauksessa pätee

$$\log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \tag{3.23}$$

kaikille $(i, j) \in S$ ja

$$\sum_{i=1}^{I} \delta_i^{(2)} \lambda_i^X = \sum_{j=1}^{J} \delta_j^{(1)} \lambda_j^Y = 0, \tag{3.24}$$

$$\sum_{i=1}^{I} \delta_{ij} \lambda_{ij}^{XY} = \sum_{j=1}^{J} \delta_{ij} \lambda_{ij}^{XY} = 0, \tag{3.25}$$

missä

$$\delta_{ij} = \begin{cases} 1 & \text{kaikille } (i, j) \in S \\ 0 & \text{muuten,} \end{cases} \tag{3.26}$$

$$\delta_i^{(2)} = \begin{cases} 1 & \text{jos } \delta_{ij} = 1 \text{ jollekin } j \\ 0 & \text{muuten.} \end{cases} \tag{3.27}$$

Samoin määritellään $\delta_j^{(1)}$. Kvasiriippumattomuusmalli määritellään nyt osataululle S asettamalla $\lambda_{ij}^{XY} = 0$ kaikille $(i, j) \in S$, jolloin

$$\log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y \tag{3.28}$$

kaikille $(i, j) \in S$.

3.5.2 Lineaariset kontrastit

Toinen tapa määritellä kvasiriippumattomuus loglineaarisella asteikolla on interaktiokontrastien kautta. Tämä määritellään

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \delta_{ij} \beta_{ij} \log m_{ij} \tag{3.29}$$

($\beta_{ij} \neq 0$ joillekin i, j), missä

$$\sum_{i=1}^{I} \delta_{ij} \beta_{ij} = \sum_{j=1}^{J} \delta_{ij} \beta_{ij} = 0.$$

Termi δ_{ij} rajoittaa yhtälön niihin kontrasteihin, jotka kuuluvat soluihin joukossa S, ja $0 \log 0$ saa aina arvon 0. Nyt voidaan kirjoittaa interaktiokontrastit

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \delta_{ij} \beta_{ij} \lambda_{ij}^{XY} \tag{3.30}$$

($\beta_{ij} \neq 0$ joillekin i, j). Kvasiriippumattomuus on voimassa, kun kaikki interaktiokontrastit ovat nollia.

3.5.3 Epäinteraktiiviset solut

Jotkin solut ovat sellaisella kohtaa taulussa, että ne eivät osallistu interaktion muodostamiseen. Oletetaan, että solulle (k, l) epätäydellisessä osataulussa S ei päde kaavan (3.30) mukainen kontrasti ja $\beta_{ij} \neq 0$. Tällöin parametri, joka vastaa solua λ_{kl}^{XY}, ei ole mukana missään interaktiokontrastissa. Tällöin sanotaan, että solu (k, l) on epäinteraktiivinen, koska se ei kerro mitään interaktiokontrastien arvoista. Tarkastellaan esimerkkinä Taulukkoa 5, missä $m_{13} = m_{23} = 0$ (viivat kuvaavat rakenteellisia nollia). Nyt joukkoon S kuuluvat kaikki solut lukuunottamatta soluja (1,3) ja (2,3).

Taulukko 5:

m_{11}	m_{12}	—
m_{21}	m_{22}	—
m_{31}	m_{32}	m_{33}

Solu (3,3) on nyt epäinteraktiivinen ja interaktiokontrastit ovat kaikki muotoa

$$\sum_{i=1}^{3}\sum_{j=1}^{2}\beta_{ij}\lambda_{ij}^{XY}.$$

Epäinteraktiivisten solujen poistaminen estimoinnin ajaksi yksinkertaistaa laskuja huomattavasti.

3.5.4 Liittyneisyys ja eroteltavuus

Kun tarkastellaan odotusarvojen m_{ij} estimointia kvasiriippumattomuusmallin tapauksessa, tulee vastaan tilanteita, joissa kannattaa käyttää vain tiettyjä osia epätäydellisestä taulusta. Näitä tilanteita on helpoin kuvata liitettävyyden ja eroteltavuuden käsitteillä. Eroteltavuudesta huolehtiminen on oleellista vapausasteita tarkasteltaessa, vaikkakin eroteltavien epätäydellisten taulujen pelkistäminen johtaa usein odotettujen solufrekvenssien estimoimiseen tarkoitettujen laskutoimitusten yksinkertaistumiseen.

Kaksisuuntaisessa taulussa kaksi solua ovat *yhteydessä*, mikäli ne eivät sisällä rakenteellisia nollia ja ne ovat joko samalla rivillä tai samassa sarakkeessa. Sellaisia soluja, jotka eivät sisällä rakenteellisia nollia, sanotaan *liittyneiksi*, jos jokainen solupari voidaan liittää toisiinsa sellaisen ketjun kautta, jossa mitkä tahansa vierekkäiset solut ovat yhteydessä toisiinsa. Epätäydellinen kaksisuuntainen taulu on liittynyt, jos sen rakenteellisia nollia sisältämättömät solut muodostavat liittyneen joukon.

Epätäydellinen taulu, joka ei ole liittynyt, on *eroteltava*. Sopivien rivien ja sarakkeiden permutaatioiden jälkeen voidaan eroteltavan taulun epätyhjät solut jakaa ainakin kahteen erilliseen osatauluun, jossa millään osataululla ei ole yhtään yhteistä riviä tai saraketta muiden osataulujen kanssa. Eroteltavan taulun rivejä ja sarakkeita voidaan permutoida niin, että epätyhjät solut järjestetään ainakin kahteen erilliseen blokkiin, jotka kulkevat diagonaalilla. Tällaista taulua kutsutaan blokkidiagonaaliseksi. Jokainen blokki saattaa kuitenkin yhä sisältää joitakin rakenteellisia nollia.

Eroteltavuus määriteltiin tässä odotettujen frekvenssien ja rakenteellisten nollien suhteen. Kun tarkastellaan havaittuja frekvenssejä, saatetaan havaita myös satunnaisia nollia. Määriteltäessä, onko havaittujen frekvenssien taulu eroteltava, ei tehdä eroa satunnaisten ja rakenteellisten nollien välille. Havaittu taulu on siis eroteltava, jos positiivisten havaittujen arvojen muodostama epätäydellinen taulu ei ole liittynyt.

3.5.5 Eroteltavien taulujen analysointi

Oletetaan, että odotettujen frekvenssien epätäydellinen taulu on eroteltavissa täsmälleen k:hon suorakulmioon siten, että epätyhjien solujen joukko S koostuu tasan k:sta erillisestä alijoukosta $S_1, S_2, \ldots, S_k$. Näistä jokaisessa on yhden eroteltavan suorakulmion sisältämät epätyhjät solut. Koska alijoukot S_v ja S_w (mille tahansa parille (v, w), $v \neq w$) eivät sisällä samoja rivejä tai sarakkeita, voidaan kvasiriippumattomuusmalli kirjoittaa sellaisessa muodossa, että odotetut solufrekvenssit joukossa S_v eivät sisällä samoja parametreja kuin joukon S_w odotetut frekvenssit. Jokaiselle alijoukolle S_l, $l = 1, \ldots, k$, on siis erillinen kvasiriippumattomuusmalli.

Eroteltavan taulun uskottavuusfunktio voidaan kirjoittaa tulona, jonka osatekijät sisältävät vain yhden osajoukon S_l parametrit. Siten suurimman uskottavuuden estimaatit yhdelle osajoukolle S_l perustuvat ainoastaan sen osajoukon sisältämään aineistoon. Näin ollen estimoinnissa tai testauksessa on mahdollista tarkastella jokaista eroteltavaa osaa erikseen.

Savage (1973) esitti välttämättömät ja riittävät ehdot yksikäsitteisen ja nollasta eroavan suurimman uskottavuuden estimaatin olemassaololle. Nämä ehdot ottavat huomioon havaittujen nollien jakauman. Oletetaan aluksi, että K on alijoukko riveistä $\{1,\ldots,\mathrm{I}\}$ ja K' alijoukko sarakkeista $\{1,\ldots,\mathrm{J}\}$. Lisäksi oletetaan, että S_K ja $S_{K'}$ ovat S:n rivien ja sarakkeiden tekijät ja ne vastaavat tekijöitä joukoissa K ja K'. Lopuksi oletetaan vielä, että jos A on joukon S osajoukko, niin $x(A) = \sum_{(i,j)\in A} x_{ij}$ eli $x(A)$ on osajoukon havaittu kokonaissumma. Nyt ehdot ovat

(i) havaitut marginaalit x_{i+} kaikille i ja x_{+j} kaikille j ovat positiivisia;

(ii) jos $S_K \subseteq S_{K'}$, niin $x(S_K) \leqq x(S_{K'})$, tai jos $S_K \supseteq S_{K'}$, niin $x(S_K) \geqq x(S_{K'})$, missä epäsuuruus on aito, jos osajoukko on aito.

Fienberg (1970) korvasi Savagen toisen ehdon myöhemmin yksinkertaisemmalla ehdolla

(ii′) havaittu taulu ei ole eroteltava.

Kvasiriippumattomuusmallin vapausasteet ovat $IJ - z_e - (I + J - 1) = (I-1)(J-1) - z_e$, missä z_e on rakenteellisten nollien lukumäärä. Odotetut frekvenssit ovat samat kuin havaitut frekvenssit, mikäli riippumattomien parametrien määrä on suurempi tai yhtäsuuri kuin solujen kokonaismäärä. Esimerkiksi Taulukko 6 toteuttaa ehdot (i) ja (ii), mutta siinä vapausasteiden määrä on nolla. Koska odotetut frekvenssit määräytyvät täysin marginaalien mukaan, nähdään helposti, että odotettujen frekvenssien taulu on sama kuin havaittujen. Vapausasteiden määrä on nolla ainoastaan Taulukon 6 kaltaisissa tilanteissa, joissa kaikki solut ovat epäinteraktiivisia.

Taulukko 6: Kaikki solut epäinteraktiivisia.

4	6	—	—	—	10
—	5	6	—	—	12
—	—	6	3	—	9
—	—	—	4	5	9
4	11	13	7	5	40

Yksinkertaisuuden vuoksi tässä on käsitelty ainoastaan kaksiulotteista tapausta. Kvasiloglineaaristen mallien teoria yleistyy kuitenkin myös useampiulotteisiin tapauksiin.

3.6 Yhteensopivuuden testaaminen

Halutun mallin yhteensopivuutta voidaan testata uskottavuusosamäärän testisuureen G^2 avulla. Olkoon $G^2(M)$ uskottavuusosamäärän testisuureen arvo mallille M. Se perustuu sovitettujen arvojen $\hat{\mathbf{m}}$ ja aineiston $\mathbf{n}$ vertailuun:

$$G^2(M) = -2[\log L(\hat{\mathbf{m}}; \mathbf{n}) - \log L(\mathbf{n}; \mathbf{n})],$$

missä $L(\hat{\mathbf{m}}; \mathbf{n})$ on uskottavuusfunktion maksimi mallille M ja $L(\mathbf{n}; \mathbf{n})$ uskottavuusfunktio rajoittamattomalle (kyllästetylle) mallille. Testisuure noudattaa asymptoottisesti χ^2-jakaumaa vapausastein *df*, missä *df* on mallien vapausasteiden erotus.

Tarkastellaan malleja M_1 ja M_2, joista M_2 on mallin M_1 erikoistapaus, ts. yksinkertaisempi malli. Mallien sanotaan olevan sisäkkäisiä. Olkoot ν_1 ja ν_2 vastaavasti mallien M_1 ja M_2 vapausasteet. Koska M_2 on yksinkertaisempi kuin M_1, on $\nu_1 < \nu_2$, ja pienempi määrä parametreja toteuttaa mallin M_2 kuin mallin M_1. Siten $L(\hat{\mathbf{m}}_\mathbf{2}; \mathbf{n})$ ei voi olla suurempi kuin $L(\hat{\mathbf{m}}_\mathbf{1}; \mathbf{n})$, mistä seuraa, että

$$G^2(M_1) \leq G^2(M_2).$$

Olettaen, että malli M_1 pätee, voidaan mallin M_2 sopivuutta testata käyttämällä logaritmisten uskottavuusfunktioiden erotusta. Tällöin

$$\begin{aligned} G^2(M_2|M_1) &= -2[\log L(\hat{\mathbf{m}}_\mathbf{2}; \mathbf{n}) - \log L(\hat{\mathbf{m}}_\mathbf{1}; \mathbf{n})] \\ &= -2[\log L(\hat{\mathbf{m}}_\mathbf{2}; \mathbf{n}) - \log L(\mathbf{n}; \mathbf{n})] - \{-2[\log L(\hat{\mathbf{m}}_\mathbf{1}; \mathbf{n}) - \log L(\mathbf{n}; \mathbf{n})]\} \\ &= G^2(M_2) - G^2(M_1). \end{aligned}$$

Testisuureen arvo on suuri, kun M_2 sopii huonosti verrattuna malliin M_1.

Uskottavuusosamäärän testisuure $G^2(M_2)$ voidaan osittaa kahden muun edellä esitellyn testisuureen avulla, jolloin saadaan

$$G^2(M_2) = G^2(M_1) + G^2(M_2|M_1).$$

Kun malli M_2 pätee, noudattaa $G^2(M_1)$ asymptoottisesti χ^2-jakaumaa vapausastein ν_1, $G^2(M_2)$ asymptoottisesti χ^2-jakaumaa vapausastein ν_2 ja $G^2(M_2|M_1)$ asymptoottisesti χ^2-jakaumaa vapausastein $\nu_2 - \nu_1$. Kun M_1 pätee mutta M_2 ei, noudattaa testisuure $G^2(M_1)$ edelleen χ^2-jakaumaa, mutta kaksi muuta testisuuretta kasvavat rajatta, kun n kasvaa.

Verrattaessa hierarkkisia malleja on edellä esitelty menetelmä yleisesti käytössä. Kuitenkin kun otoskoko n on hyvin suuri, nousevat pienetkin erot merkitseviksi. Siksi tässä esitelty uskottavuusosamäärien suhteisiin perustuva menetelmä ei ole käyttökelpoinen ylioppilasaineistossa.

4 Mallien vertailu suurten aineistojen tapauksessa

Tarkastellaan aluksi yksinkertaista esimerkkiä havainnollistaaksemme otoskokoon liittyvää ongelmaa mallinvalinnassa. Nyt sovellettava informaatiokriteereiden esittely ja niihin liittyvän mallinvalinnan teoria perustuu artikkeliin Burnham ja Anderson (2004).

Esimerkki 4.1. *Oletetaan, että aineisto on seuraava* 2×2*-taulu.*

n_{11}	n_{12}	n_{1+}
n_{21}	n_{22}	n_{2+}
n_{+1}	n_{+2}	n

Tästä voidaan laskea ristitulosuhde

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}},$$

jonka logaritmille nollahypoteesin tapauksessa pätee asymptoottisesti (Bishop, Fienberg ja Holland, 1975)

$$\log \hat{\theta} \sim \mathrm{N}\left(0, \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right).$$

Oletetaan, että tapauksessa $n = 100$ *saadaan ristitulosuhteen logaritmille p-arvo 0.05. Kerrotaan nyt taulun soluja 500:lla, jolloin* $n = 50000$*. Edelleen muuttujien riippuvuutta kuvaava ristitulosuhteen arvo pysyy muuttumattomana, mutta ristitulosuhteen varianssi pienenee 500-osaan aiemmasta. Mikä nyt on p-arvo, joka kertoo saman informaation kuin edellinen?*

Edellä oleva esimerkki kuvaa suurten aineistojen käsittelyyn liittyvää ongelmaa. Kun havaintoja on paljon, on varianssi hyvin pieni ja tällöin pienetkin erot riippumattomuudesta tulevat merkitseviksi. Tavallisella hypoteesintestauksella ei näin ollen päästä käsiksi kiinnostaviin riippuvuussuhteisiin.

Informaatiokriteereihin perustuva mallinvalinta pohjaa testaukseen verrattuna toisenlaiseen ajatukseen: pyrkimyksenä on löytää tasapaino mallin ja aineiston kompleksisuuksien välille. Liian vähäparametriset mallit saattavat olla harhaisia, kun taas yliparametrisoidut mallit saattavat olla liian spesifejä sille aineistolle, josta ne on estimoitu.

Mallinvalinnassa, toisin kuin testauksessa, ei aseteta nollahypoteesia, joka pyrittäisiin kumoamaan. Nollahypoteesin sijaan käytetään useita perusteltuja hypoteeseja, malleja, joita tarkastellaan. Aineisto antaa tukea osalle malleista, kun taas osa katsotaan riittämättömiksi. Yksi lähestymistapa mallien vertailuun ovat informaatiokriteerit, jotka ottavat samanaikaisesti huomioon aineiston ja mallin yhteensopivuuden sekä mallin kompleksisuuden. Lisäksi, toisin kuin edellisessä luvussa esitellyssä uskottavuusosamäärätestissä, informaatiokriteerien avulla voidaan vertailla myös eisisäkkäisiä malleja.

4.1 Informaatiokriteerit

Mallin valintaan ja arviointiin löytyy useita menetelmiä. Näihin tarkoituksiin kehitetyn informaatiokriteerin esitteli ensimmäisenä Akaike (1974) ja sen jälkeen kriteereitä on kehitetty lukuisia muita useista eri lähtökohdista. Schwarzin bayesilainen kriteeri (Schwarz, 1978) on nimensä mukaisesti lähtökohdiltaan bayesilainen ja se perustuu tietynlaisten priorien valintaan. Rissasen kriteeri (Rissanen, 1983) perustuu MDL:ään (Minimum Description Length), joka pyrkii kuvaamaan aineiston mahdollisimman lyhyellä koodauksella. Tässä luvussa näistä käsitellään lähemmin kahta ensimmäistä.

Akaiken informaatiokriteeri perustuu Kullback–Leibler-informaatioon (Kullback ja Leibler, 1951), joten aloitetaan tarkastelemalla sitä. Oletetaan, että jakauma f kuvaa todellisuutta ja jakauma g mallia, joka approksimoi todellisuutta. K–L-informaatio $I(f, g)$ kertoo, kuinka paljon informaatiota menetetään, kun mallia käytetään approksimoimaan todellisuutta. Jatkuville funktioille tämä määritellään

$$I(f, g) = \int f(x) \log \left(\frac{f(x)}{g(x|\theta)} \right) dx.$$

Paras malli menettää vähiten informaatiota muihin malleihin verrattuna. Tätä vastaa K–L-informaation minimoiminen g:n suhteen. Vaihtoehtoisesti K–L-informaatio voidaan ajatella todellisuuden ja mallin väliseksi etäisyydeksi, vaikkakaan se ei toteuta kolmioepäyhtälöä eikä siis ole sellaisenaan metriikka matemaattisessa mielessä.

K–L-informaatiota ei voida käyttää suoraan, koska todellisuutta tai sitä approksimoivien mallien parametreja ei tunneta. Aineistoa analysoitaessa parametrit täytyy estimoida ja tähän liittyy lähes aina epävarmuutta. Ero estimoitujen ja todellisten parametrien muodostamien mallien välillä vaikuttaa siihen, kuinka K–L-informaatiota käytetään pohjana mallinvalinnalle. Siksi mallinvalintakriteerin tulee perustua odotetun K–L-informaation minimoimiseen tunnetun K–L-informaation minimoimisen sijaan.

K–L-informaatio voidaan ilmaista muodossa

$$I(f, g) = \int f(x) \log(f(x)) dx - \int f(x) \log(g(x|\theta)) dx$$

tai

$$I(f, g) = \mathbb{E}_f[\log(f(x))] - \mathbb{E}_f[\log(g(x|\theta))],$$

missä odotusarvot otetaan todellisuuden f suhteen. Odotusarvo $\mathbb{E}_f[\log(f(x))]$ on nyt vakio, joka ei riipu aineistosta tai mallista. Näin ollen jokaiselle mallille täytyy estimoida ainoastaan $\mathbb{E}_f[\log(g(x|\theta))]$.

Akaike näytti, että oleellista K–L-informaatioon liittyvän mallinvalintakriteerin muodostamisessa on estimoida

$$\mathbb{E}_y \mathbb{E}_x[\log(g(x|\hat{\theta}(y)))], \tag{4.1}$$

missä sisin osa on $\mathbb{E}_f[\log(g(x|\theta))]$ ja missä θ on korvattu suurimman uskottavuuden estimaattorillaan perustuen malliin g ja aineistoon y. Vaikka vain y kuvaa tässä aineistoa, voidaan sekä x että y käsittää riippumattomiksi satunnaisotoksiksi samasta

jakaumasta. Kumpikin odotusarvo otetaan todellisuuden f suhteen ja lauseke (4.1) on nyt Akaiken kriteerin ja kaikkien muidenkin K–L-informaatioon perustuvien mallinvalintamenetelmien kohde.

Akaike löysi yhteyden K–L-informaation ja uskottavuusteorian välille. Hän huomasi, että uskottavuusfunktion maksimikohta on harhainen estimaattori lausekkeelle $\mathbb{E}_y\mathbb{E}_x[\log(g(x|\hat{\theta}(y)))]$, mutta tämä harha on likimäärin yhtä suuri kuin p eli estimoitavien parametrien määrä mallissa g. Nyt suurille otoksille ja "hyville" malleille likimäärin harhaton estimaattori lausekkeelle $\mathbb{E}_y\mathbb{E}_x[\log(g(x|\hat{\theta}(y)))]$ on

$$\log(L(\hat{\theta})) - p, \tag{4.2}$$

missä p on asymptoottinen harhankorjaustermi. Toisin kuin kirjallisuudessa joskus erheellisesti todetaan, p ei siis suinkaan ole valittu mielivaltaisesti.

Historiallisista syistä Akaike kertoi lausekkeen (4.2) luvulla -2 ja määritteli informaatiokriteerin

$$\mathrm{AIC} = 2p - 2\log(L), \tag{4.3}$$

missä p on mallin parametrien määrä ja L uskottavuusfunktion maksimi. Akaiken kriteeri on estimaatti Kullback–Leibler-informaatiolle ja sen kohdemalli riippuu otoskoosta n. Kriteeriä on arvosteltu siitä, ettei se ole tarkentuva. Kuitenkin n:n kasvaessa AIC:n valinta lähestyy sitä mallia, jossa informaation menetys on pienintä.

Schwarzin esittelemä bayesilainen informaatiokriteeri,

$$\mathrm{BIC} = p\log(n) - 2\log(L), \tag{4.4}$$

ei nimestään huolimatta pohjaa informaatioteoriaan. Sen sijaan kriteeri perustuu suurten otosten approksimaatioon Bayes-tekijälle ja siihen oletukseen, että mallien priorit ovat yhtä suuret. Akaiken kriteeristä poiketen Schwarzin kriteeri ei lähde siitä ajatuksesta, että on olemassa todellisuus, jota yritetään mallintaa. Perinteisesti oletetaan, että mahdollisten mallien joukosta yksi on sellainen, joka on luonut aineiston, ja tämä malli on Schwarzin kriteerin kohdemalli. Tämä kuitenkin pätee vain, kun n on hyvin suuri. Muissa tapauksissa BIC lähestyy niin kutsuttua puolitodellista mallia eli vähäparametrisinta mallia niiden mallien joukossa, jotka ovat lähinnä totuutta Kullback–Leibler-mielessä. Bayesilaisittain tämä tarkoittaa sitä, että mallia koskeva epävarmuus häviää. Pienillä otoksilla BIC saattaa valita liian pelkistetyn mallin, mutta erityisesti erittäin suurten otosten tapauksessa se valitsee tulkinnan kannalta käyttökelpoisen mallin. Schwarzin kriteerillä ei ole välttämättä tarkoitus mallintaa kompleksista todellisuutta. Sen sijaan pyrkimyksenä on löytää aineistosta ne erot ja riippuvuudet, jotka ovat kiinnostavia.

Vaikka vain toinen kriteereistä on nimetty bayesilaisuuden mukaan, on Akaiken kriteerin mukainen mallinvalinta yhtä lailla bayesilainen prosessi kuin Schwarzin kriteerinkin, ero on vain priorijakaumassa. Toisaalta kumpikin voidaan johtaa myös eibayesilaisista lähtökohdista. Näiden kahden kriteerin välinen oleellinen ero onkin lähinnä kohdemallissa sekä muodollisesti siinä, kuinka ne sakottavat uusista parametreista.

4.2 Informaatiokriteereihin perustuva mallinvalinta

Hyvä malli erottaa toisistaan informaation ja kohinan. Informaatiolla viitataan tässä tapauksessa riippuvuuksien rakenteisiin ja kohina on vaihtelua, joka voidaan jättää mallista pois informaation oleellisesti vähenemättä. Tarkoituksena on löytää malli, joka erottaa rakenteellisen informaation kohinasta. Jos ollaan aivan tarkkoja, ei informaatiokriteerien avulla varsinaisesti pyritä mallintamaan aineistoa vaan aineiston sisältämää informaatiota.

Käytännössä informaatiokriteereihin perustuva mallinvalinta alkaa mallien sovituksella. Näistä lasketaan uskottavuusosamäärän arvot ja niiden avulla Akaiken tai Schwarzin kriteerien arvot. Koska BIC on yleensä parempi suurten otosten tapauksessa, käytetään vertailussa ensisijaisesti sitä. Malleista valitaan se, jolla informaatiokriteerin arvo on pienin.

4.3 Esimerkki ylioppilasaineistosta

Tarkastellaan kolmiulotteista taulua, jossa muuttujina ovat äidinkielen ylioppilaskokeen arvosana (as), kokelaan sukupuoli (sp) sekä kokeen kieli (k) (Taulukko 7). Tarkoituksena on saada kolmiulotteinen taulu pelkistettyä useammaksi pienempiulotteiseksi prosenttitauluksi siten, että oleellinen informaatio säilyy ja on helposti luettavissa.

Taulukko 7: Arvosanajakauma luokiteltuna äidinkielen ja sukupuolen mukaan, kevät 2006 ja kevät 2007.

		I	A	B	C	M	E	L	Yhteensä
Suomi	Nainen	334	1611	4124	9805	8286	4216	1677	30053
		1.11	5.36	13.72	32.63	27.57	14.03	5.58	100 (%)
	Mies	565	2663	4665	7397	4279	1745	731	22045
		2.56	12.08	21.16	33.55	19.41	7.92	3.32	100 (%)
	Yhteensä	899	4274	8789	17202	12565	5961	2408	52098
		1.73	8.20	16.87	33.02	24.12	11.44	4.62	100 (%)
Ruotsi	Nainen	22	99	311	761	623	372	142	2330
		0.94	4.25	13.35	32.66	26.74	15.97	6.09	100 (%)
	Mies	70	187	400	531	302	140	41	1671
		4.19	11.19	23.94	31.78	18.07	8.38	2.45	100 (%)
	Yhteensä	92	286	711	1292	925	512	183	4001
		2.30	7.15	17.77	32.29	23.12	12.80	4.57	100 (%)

Loglineaarinen malli

Tutkitaan ensin kolmen muuttujan yhteisjakaumaa log-lineaaristen mallien kehikossa. Näiden vertailussa käytetään Akaiken ja Schwarzin informaatiokriteereitä, joista jälkimmäisen käyttö on perustellumpaa, koska kyseessä on suuri otos. Taulukossa 8 nähdään logaritmisen uskottavuussuhteen arvo (−2·loglik) sekä mallin parametrien määrä (p), joiden avulla on laskettu AIC ja BIC. Malleissa on hierarkkisuusperiaatteen nojalla (luku 3.3.1) mukana myös kaikki alemman asteen termit, jolloin malli `sp*as*k` on lyhenne mallille `vakio+sp+as+k+sp:as+sp:k+as:k+sp:as:k`. Kaksoispiste merkitsee tässä muuttujien välistä yhdysvaikutusta.

Taulukko 8: Loglineaariset mallit

Malli	−2·loglik	p	AIC	BIC
1. `sp*as*k`	290785.57	27	**290839.57**	291081.38
2. `sp*as+sp*k+as*k`	290804.55	21	290846.55	291034.63
3. `sp*as+sp*k`	290825.75	15	290855.75	290990.0
4. `sp*as+as*k`	290804.9	20	290844.90	291024.02
5. `sp*k+as*k`	293259.99	15	293289.99	293424.33
6. `sp*as+k`	290826.21	14	290854.21	**290979.60**
7. `sp*k+as`	293281.3	9	293299.30	293379.91
8. `as*k+sp`	293260.45	14	293288.45	293413.83
9. `sp*as`	342212.01	7	342226.01	342288.70
10. `sp*k`	323592.05	2	323596.05	323613.96
11. `as*k`	294626.15	7	294640.15	294702.84
12. `sp+as+k`	293281.76	8	293297.76	293369.41
13. `sp+as`	342212.01	7	342226.01	342288.70
14. `sp+k`	323592.05	2	323596.05	323613.96
15. `as+k`	294626.15	7	294640.15	294702.84

Akaiken informaatiokriteerin arvo on pienin kyllästetyllä mallilla (malli 1). Taulukko ei siis pelkisty lainkaan. BIC sen sijaan saa pienimmän arvon mallilla 6, jossa on mukana kielen päävaikutus, sukupuolen ja arvosanan yhdysvaikutus sekä hierarkkisuusperiaatteen nojalla myös näiden päävaikutukset. Tässä mallissa siis sukupuoli ja arvosana riippuvat toisistaan, mutta kieli ei näistä kummastakaan. Kokeen kielellä on kuitenkin tasovaikutus, joten sen on oltava mallissa mukana.

Logit-malli

Jos jokin muuttujista tulkitaan vasteeksi, jota selitetään jäljellä olevien muuttujien avulla, otetaan käyttöön logit-mallit. Tässä esimerkissä äidinkielen kokeesta saatu arvosana valitaan vasteeksi, jolloin sukupuoli ja kieli ovat sen selittäjiä. Vertaillaan taas malleja informaatiokriteerien avulla (Taulukko 9).

Tällä kertaa AIC saa pienimmän arvon mallilla 1, jonka mukaan sukupuolen ja kielen yhdysvaikutus sekä niiden päävaikutukset selittävät äidinkielen arvosanaa. BIC

Taulukko 9: Logit-mallit

Malli	−2·loglik	p	AIC	BIC
1. `sp*k`	185521.21	24	**185569.21**	185784.15
2. `sp+k`	185540.19	18	185576.19	185737.39
3. `sp`	185571.4	12	185595.40	**185702.87**
4. `k`	187995.63	12	188019.63	188127.10

sen sijaan saa pienimmän arvon mallilla 3, jonka mukaan ainoastaan sukupuoli selittää äidinkielen arvosanoja.

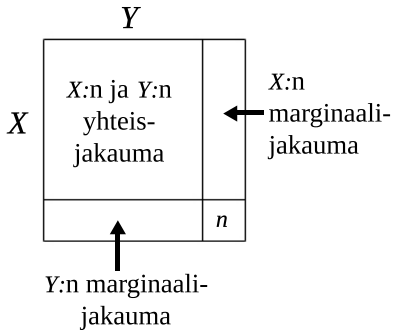
Logit-malli voidaan aina ilmaista loglineaarisena ottamalla mukaan kaikki selittävien muuttujien väliset yhdysvaikutukset, joissa vaste ei ole mukana. Logit-mallia `sp` vastaa siis loglineaarinen malli `sp*as+sp*k`, eli tässä saatu malli ei ole sama kuin aiempi `sp*as+k`.

5 Malliin perustuva pelkistäminen

Usein T-ulotteisen ($T \geq 2$) frekvenssitaulun riippuvuusrakenne voi sallia taulun informaation esittämisen yksinkertaisemmin. Tätä kutsutaan pelkistämiseksi. Pelkistettäessä jätetään esittämättä sellaiset interaktiot, jotka eivät ole oleellisia aineiston sisältämän informaation kannalta. Pelkistämiseen pyritään, koska aineiston tulkitseminen helpottuu usein huomattavasti, jos moniulotteinen taulu voidaan esittää useana pienempiulotteisena tauluna.

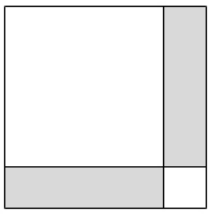
5.1 Riippuvuusrakenteet

Tarkastellaan erilaisia riippuvuusrakenteita kaksiulotteisen $I \times J$-taulun avulla. Esimerkeissä muuttujien X ja Y muodostama ristiintaulukko on jaettu neljään osaan, jotka ovat oleellisia riippuvuusrakennetta tarkasteltaessa: havaintojen kokonaismäärään n, muuttujan X marginaalijakaumaan, muuttujan Y marginaalijakaumaan sekä muuttujien X ja Y yhteisjakaumaan (Kuva 5.1). Tässä pelkistäminen tapahtuu ainoastaan komponenteittain, mutta mahdollista saattaisi olla myös jatkaa pidemmälle ja yhdistää muuttujien luokkia.



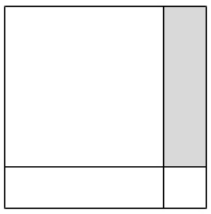
Kuva 5.1: Kaksiulotteinen taulu

Esimerkki 5.1. *Oletetaan, että muuttujat X ja Y ovat riippumattomia. Tällöin loglineaarinen malli on muotoa $X + Y$ ja kaikki aineiston informaatio sisältyy muuttujien X ja Y marginaalijakaumiin. Aineistosta riittää esitellä kummankin muuttujan marginaali- ja prosenttijakaumat (Kuva 5.2).*



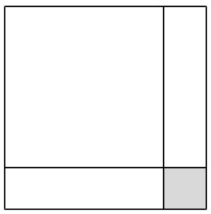
Kuva 5.2: X ja Y riippumattomat

Esimerkki 5.2. *Oletetaan, että muuttuja Y on vakio. Tällöin loglineaarinen malli on muotoa X ja kaikki aineiston informaatio sisältyy muuttujan X marginaalijakaumaan. Aineistosta riittää esitellä X:n marginaali- ja prosenttijakauma (Kuva 5.3).*



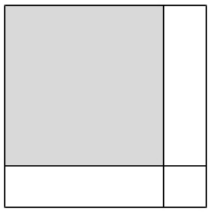
Kuva 5.3: Y vakio

Esimerkki 5.3. *Oletetaan, että sekä X että Y ovat vakioita. Tällöin koko aineiston sisältämä informaatio voidaan tiivistää yhteen lukuun n eli havaintojen määrä (Kuva 5.4).*



Kuva 5.4: X ja Y vakioita

Esimerkki 5.4. *Oletetaan, että muuttujat X ja Y ovat riippuvia. Tällöin loglineaarinen malli on muotoa $X * Y$ ja aineiston esittämiseen tarvitaan koko X:n ja Y:n yhteisjakauma (Kuva 5.5). Prosenttijakaumat voidaan esittää riveittään tai sarakkeittain tarpeen mukaan.*



Kuva 5.5: X ja Y riippuvia

Kun muuttujat X ja Y riippuvat toisistaan, on käytettävä niiden empiiristä yhteisjakaumaa. Riippuvuusrakennetta voi kuitenkin olla mahdollista tutkia esimerkiksi rajoitteiden avulla tai yksinkertaisesti tulkitsemalla prosenttijakaumia.

Erikoistapauksena voidaan käsitellä tilannetta, jossa riippuvuuden aiheuttaa poikkeava solu, jonka frekvenssi on erikoisen pieni tai suuri. Kun kyseinen solu tulkitaan rakenteelliseksi nollaksi, voidaan käyttää kvasiriippumattomuusmallia. Nyt riittää esittää muuttujien X ja Y marginaalijakaumat sekä maininta yhden solun poikkeavan suuresta tai pienestä arvosta.

Kaksisuuntaisessa XY-taulussa esiintyy siis jokin seuraavista riippuvuusrakenteista:

- Kumpikin muuttuja vakio
- Malli X: muuttujalla X päävaikutus, Y vakio
- Malli Y: muuttujalla Y päävaikutus, X vakio
- Malli $X + Y$: muuttujat riippumattomia
- Malli $X * Y$: muuttujat riippuvia.

5.2 Yleinen menetelmä

Kun moniulotteista taulua lähdetään pelkistämään ja tarkastellaan muuttujien välisiä riippuvuussuhteita, etsitään ensin sopiva loglineaarinen malli, joka koostuu yhdestä tai useammasta erillisestä komponentista. Tässä työssä käytettyä merkintätapaa noudattaen mallin komponentteja ovat plusmerkillä erotetut osat. Ne muodostuvat joko yhden muuttujan päävaikutustermistä tai vähintään kahden muuttujan välisestä yhdysvaikutustermistä.

Päävaikutustermin tapauksessa komponenttia vastaava esitys on suora jakauma, yhdysvaikutustermin tapauksessa osataulu tai useita osatauluja. Kolmen muuttujan yhdysvaikutuksen tapauksessa voidaan esittää kahden muuttujan yhteisjakaumat kolmannen muuttujan eri tasoilla.

Vastaava menettely sopii myös silloin, kun yksi muuttujista on vaste ja muut selittäjiä. Etsitään siis sopiva logit-malli ja hajotetaan erilliset komponentit osatauluiksi, joissa kaikissa on mukana vaste.

Jos loglineaarisessa tai logit-mallissa on mukana usean muuttujan välisiä yhdysvaikutuksia, joudutaan tekemään ratkaisuja esitystavan suhteen. Esimerkiksi kolmen muuttujan yhdysvaikutuksen tapauksessa esitetään kahden muuttujan yhteisjakauma kolmannen muuttujan eri arvoilla. Jos kolmannella muuttujalla on useita luokkia, ei tämä esitystapa ole kuitenkaan kovin havainnollinen. Tällaisessa tapauksessa voidaan esimerkiksi jäännöksiä tarkastelemalla ja kvasiriippumattomuusmallin sovituksella tutkia, johtuuko korkean asteen yhdysvaikutus yhdestä tai muutamasta solusta. Jos yhteisjakaumat ovat samanlaiset joillakin kolmannen muuttujan tasoilla, voidaan nämä luokat yhdistää, jolloin päästään suppeampaan esitykseen.

Jos aineisto sisältää rakenteellisia nollia, tulisi tämä ottaa huomioon ja sovittaa kvasiloglineaarinen malli.

5.3 Esimerkki ylioppilasaineistosta

Palataan esimerkkiin luvussa 4.3, jossa sovitettiin loglineaarisia malleja muuttujille sukupuoli, kieli ja arvosana. Schwarzin informaatiokriteerin perusteella parhaaksi malliksi valikoitui `sp*as+k`, jossa on sukupuolen ja kielen yhdysvaikutus sekä kielen päävaikutus.

Loglineaarisen mallin perusteella riittää nyt kertoa arvosanat miesten ja naisten suhteen yhdistetyssä aineistossa sekä suomen- ja ruotsinkielisen kokeen kirjoittaneiden määrät. Kolmiulotteinen taulu saadaan näin esitettyä yhden suoran jakauman (Taulukko 10) ja yhden ristiintaulukon (Taulukko 11) avulla eli huomattavasti yksinkertaisemmin kuin alkuperäinen taulu. Näistä saadaan myös helposti tehtyä graafisia esityksiä, kuten Kuva 5.6 ja Kuva 5.7 osoittavat.

Taulukko 10: Kielten jakauma, kevät 2006 ja kevät 2007.

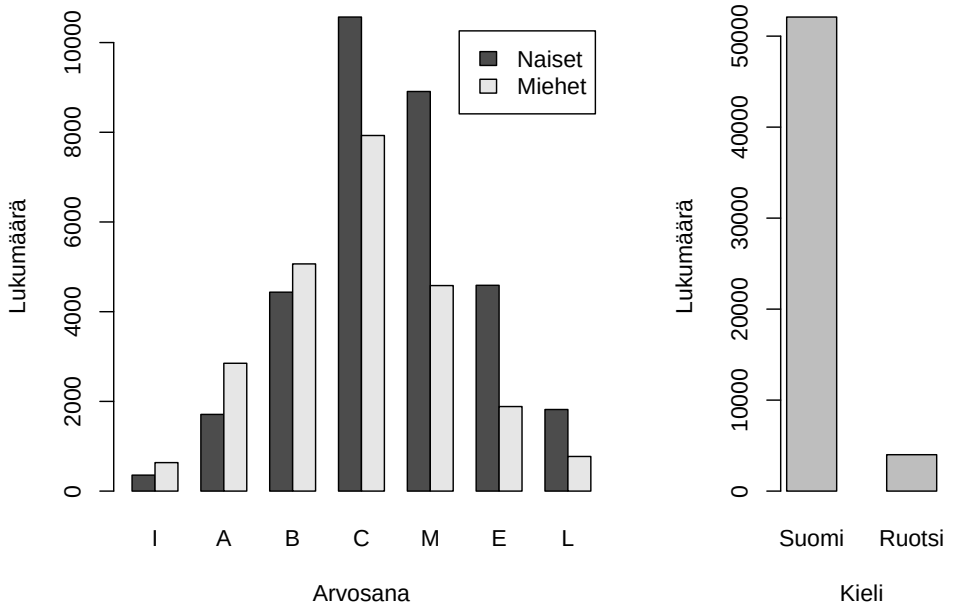
Suomi	Ruotsi
52098	4001
92.87 %	7.13 %

Taulukko 11: Arvosanajakauma sukupuolen suhteen, kevät 2006 ja kevät 2007.

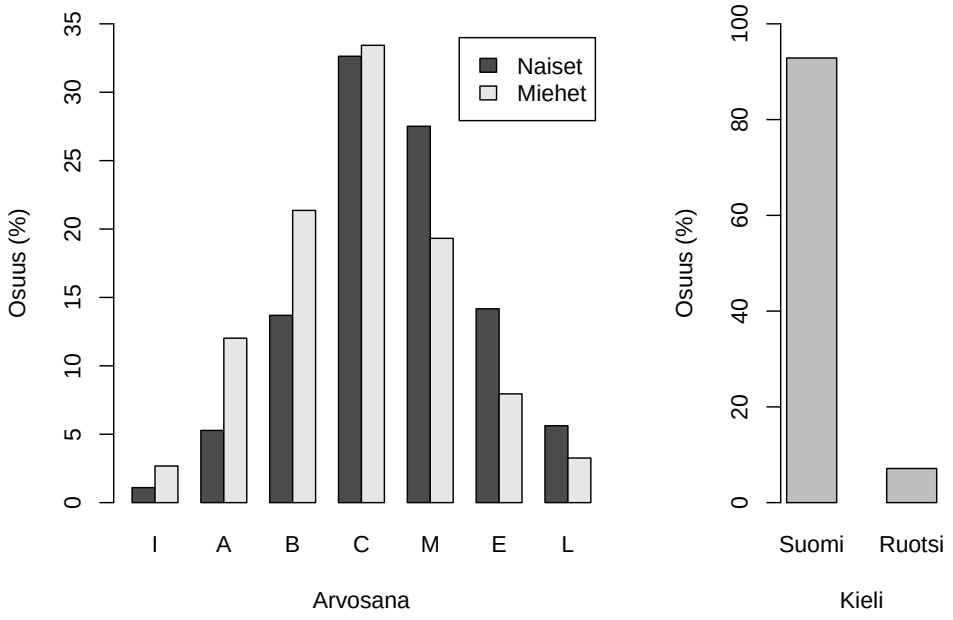
	I	A	B	C	M	E	L	Yhteensä
Nainen	356	1710	4435	10569	8909	4588	1819	32386
	1.10	5.28	13.69	32.63	27.51	14.17	5.62	100 (%)
Mies	635	2850	5065	7928	4582	1885	772	23717
	2.68	12.02	21.36	33.43	19.32	7.95	3.26	100 (%)
Yhteensä	991	4560	9500	18497	13491	6473	2591	56103
	1.77	8.13	16.93	32.97	24.05	11.54	4.62	100 (%)

Kuten Taulukosta 10 nähdään, on suomenkielisen kokeen kirjoittaneita 52098 ja ruotsikielisen kokeen tehneitä 4001. Taulukosta 11 löytyy kunkin arvosanan kirjoittaneiden miesten ja naisten lukumäärät sekä riviprosentit. Taulukossa 11 on huomattavissa, että naisten ja miesten jakaumat poikkeavat toisistaan. Miehiä on huomattavasti naisia runsaammin alempia arvosanoja kirjoittaneissa kun taas naisia löytyy suhteellisesti enemmän hyvin suoriutuneista. Ainoastaan C:n eli arvosanoista keskimmäisen kirjoittaneita on suunnilleen yhtä suuri osuus.

Ehdollisia malleja sovitettaessa parhaaksi logit-malliksi valikoitui `sp`. Tässä tapauksessa riittää esittää sukupuolen ja arvosanan yhteisjakauma.



Kuva 5.6: Arvosanajakaumat sukupuolittain sekä kokeen kielen jakauma



Kuva 5.7: Suhteelliset arvosanajakaumat sukupuolittain sekä kokeen kielen jakauma

6 Luokkien yhdistäminen

Jos jollakin muuttujalla on runsaasti luokkia, voi taulujen tulkitseminen muodostua hankalaksi. Tätä voidaan helpottaa yhdistelemällä luokkia. Toisaalta luokittelemalla voidaan saada myös tutkimuksen kannalta tarpeellista tietoa. Yksi tässä tutkimuksessa kiinnostavista kysymyksistä on selvittää, millaisista aiheista kirjoitetaan keskimäärin parempia arvosanoja ja millaisista huonoja.

Luokittelu voidaan tehdä joko käyttäen aineiston informaatiota tai sisällön perusteella. Edellinen luokitteluvaihtoehto nojaa pelkästään aineistoon ja se tehdään tilastollisin menetelmin. Eräästä soveltuvasta menetelmästä kerrotaan lisää seuraavassa luvussa.

Jälkimmäinen vaihtoehto johtaa helppoon tulkintaan, mutta se saattaa vaatia vahvaa sovellusalan osaamista. Tässä tutkimuksessa on mukana kahdentyyppisiä kvalitatiivisia muuttujia. Muuttujista sukupuoli, kokeen kieli ja arvosana määräytyvät suoraan aineiston perusteella ja ovat luonteeltaan yksiselitteisiä. Näiden lisäksi sovellusalan asiantuntijat ovat luokitelleet kokeiden tehtävänantoja laadullisin perustein (ks. luku 2). Haasteeksi luokittelussa muodostuvat rakenteelliset nollat, joita useamman uuden muuttujan muodostamissa ristiintaulukoissa esiintyy runsaasti.

6.1 Luokkien yhdistäminen klusterianalyysilla

Klusteroiminen on käyttökelpoinen tapa yhdistää luokkia tilastollisesti. Valittavana on useita menetelmiä kuten lähimmän naapurin menetelmä ja naapurikeskiarvomenetelmä. Tässä tutkimuksessa on käytetty k:n keskiarvon menetelmää, jossa tarkoituksena on löytää luonnollisten klustereiden keskikohtia minimoimalla ryhmien sisäinen varianssi tai neliösumma

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2, \tag{6.1}$$

missä S_i on klusteri, $i = 1, 2, \ldots, k$, ja μ_i on kaikkien ryhmään S_i kuuluvien pisteiden x_j keskipiste.

Kun tutkitaan neljän muuttujan, sukupuolen (sp), arvosanan (as), kielen (k) ja tehtävänannon (ta) yhdysvaikutuksia kevään 2006 suomenkielisessä kokeessa, saadaan malli `ta*as+ta*sp+ta*k+as*sp`. Koska tehtävänantoja on 27, tulee tämän muuttujan sisältävistä tauluista hyvin laajoja ja hankalasti tulkittavia. Jotta näistä tauluista saataisiin selkeämpiä, sovelletaan tehtävänantojen ryhmittelyyn klusterointia.

Ennen kuin päästään soveltamaan varsinaisia klusterointimenetelmiä, on aluksi tarpeen laskea etäisyysmatriisi tehtävänantojen arvosanajakaumille (Taulukko 12).

Taulukko 12: Aiheen numero ja arvosana kirjoittajamäärän mukaisessa järjestyksessä, kevään 2006 suomenkielinen koe.

	I	A	B	C	M	E	L	Yhteensä
7.	29	324	643	1162	612	224	55	3049
%	0.95	10.63	21.09	38.11	20.07	7.35	1.80	10.36
11.	44	294	602	956	532	191	44	2663
%	1.65	11.04	22.61	35.90	19.98	7.17	1.65	9.04
34.	21	182	325	718	549	262	112	2169
%	0.97	8.39	14.98	33.10	25.31	12.08	5.16	7.37
26.	28	190	425	715	473	168	38	2037
%	1.37	9.33	20.86	35.10	23.22	8.25	1.87	6.92
30.	22	202	415	690	431	178	53	1991
%	1.10	10.15	20.84	34.66	21.65	8.94	2.66	6.76
6.	24	156	345	609	361	143	48	1686
%	1.42	9.25	20.46	36.12	21.41	8.48	2.85	5.73
8.	16	137	231	530	417	223	77	1631
%	0.98	8.40	14.16	32.50	25.57	13.67	4.72	5.54
33.	13	110	254	548	405	157	60	1547
%	0.84	7.11	16.42	35.42	26.18	10.15	3.88	5.25
32.	15	113	220	403	290	113	83	1237
%	1.21	9.14	17.78	32.58	23.44	9.14	6.71	4.20
27.	18	78	168	375	274	172	111	1196
%	1.51	6.52	14.05	31.35	22.91	14.38	9.28	4.06
5.	22	106	178	336	231	138	51	1062
%	2.07	9.98	16.76	31.64	21.75	12.99	4.80	3.61
21.	17	86	179	359	228	91	31	991
%	1.72	8.68	18.06	36.23	23.01	9.18	3.13	3.37
29.	22	107	182	335	191	86	30	953
%	2.31	11.23	19.10	35.15	20.04	9.02	3.15	3.24
13.	14	91	169	330	194	58	12	868
%	1.61	10.48	19.47	38.02	22.35	6.68	1.38	2.95
24.	2	33	65	239	223	147	81	790
%	0.25	4.18	8.23	30.25	28.23	18.61	10.25	2.68

Jatkuu seuraavalla sivulla

Taulukko 12 – Jatkoa edelliseltä sivulta

	I	A	B	C	M	E	L	Yhteensä
31.	19	127	183	239	127	42	14	751
%	2.53	16.91	24.37	31.82	16.91	5.59	1.86	2.55
1.	12	49	87	195	191	120	59	713
%	1.68	6.87	12.20	27.35	26.79	16.83	8.27	2.42
9.	7	43	66	157	142	92	42	549
%	1.28	7.83	12.02	28.60	25.87	16.76	7.65	1.86
22.	2	24	59	148	146	107	54	540
%	0.37	4.44	10.93	27.41	27.04	19.81	10.00	1.83
4.	10	52	75	182	113	49	27	508
%	1.97	10.24	14.76	35.83	22.24	9.65	5.31	1.73
2.	2	23	47	116	129	86	45	448
%	0.45	5.13	10.49	25.89	28.79	19.20	10.04	1.52
25.	9	66	102	136	89	29	6	437
%	2.06	15.10	23.34	31.12	20.37	6.64	1.37	1.48
3.	3	23	32	104	113	85	74	434
%	0.69	5.30	7.37	23.96	26.04	19.59	17.05	1.47
23.	4	25	44	99	118	61	53	404
%	0.99	6.19	10.89	24.50	29.21	15.10	13.12	1.37
10.	4	23	33	85	91	59	46	341
%	1.17	6.74	9.68	24.93	26.69	17.30	13.49	1.16
28.	4	22	46	89	88	63	18	330
%	1.21	6.67	13.94	26.97	26.67	19.09	5.45	1.12
12.	1	7	11	35	30	21	14	119
%	0.84	5.88	9.24	29.41	25.21	17.65	11.76	0.40

Tämä voidaan tehdä laskemalla maksimietäisyys kaikille tehtävänantopareille ij yli arvosanojen k:

$$d_{ij} = \max_k \{|p_{ik} - p_{jk}|\}. \tag{6.2}$$

Saatu etäisyysmatriisi sijoitetaan klusterointialgoritmiin, joka jakaa aineiston neljään ryhmään. Näistä ensimmäiseen valikoituu 10, toiseen 6, kolmanteen 4 ja neljänteen 7 tehtävänantoa (Taulukko 13). Myös muut menetelmät antavat suunnilleen samanlaisia tuloksia. Nyt tehtävänantojen ja arvosanojen yhteyttä kuvaava taulu on huomattavasti helpommin tulkittavissa, kuten Taulukosta 14 nähdään.

Saadun klusteroinnin mielekkyyttä voi loglineaaristen mallien tapauksessa tutkia vielä kontrastein. Tämän voi tehdä kahdella tavalla. Näistä ensimmäisessä luokat yh-

Taulukko 13: Neljä ryhmää ja niihin sisältyvät tehtävänannot

Ryhmä	Tehtävänannot
Huonot	6, 7, 11, 13, 21, 25, 26, 29, 30, 31
Melko huonot	4, 5, 8, 32, 33, 34
Melko hyvät	1, 9, 27, 28
Hyvät	2, 3, 10, 12, 22, 23, 24

Taulukko 14: Arvosanajakaumat neljällä ryhmällä

Ryhmä	I	A	B	C	M	E	L	Yhteensä
Huonot	228 1.48 59.38	1642 10.65 61.00	3245 21.04 62.57	5531 35.86 55.92	3238 20.99 47.70	1210 7.84 38.23	331 2.15 24.74	15425 100 (%) 52.39
Melko huonot	97 1.19 25.26	700 8.58 26.00	1283 15.73 24.74	2718 33.33 27.48	2005 24.59 29.54	942 11.55 29.76	410 5.03 30.64	8155 100 (%) 27.70
Melko hyvät	41 1.47 10.68	192 6.89 7.13	367 13.16 7.08	816 29.27 8.25	695 24.93 10.24	447 16.03 14.12	230 8.25 17.19	2788 100 (%) 9.47
Hyvät	18 0.59 4.69	158 5.14 5.87	291 9.46 5.61	826 26.85 8.35	850 27.63 12.52	566 18.40 17.88	367 11.93 27.43	3076 100 (%) 10.45
Yht.	384 1.30 100 (%)	2692 9.14 100 (%)	5186 17.61 100 (%)	9891 33.59 100 (%)	6788 23.05 100 (%)	3165 10.75 100 (%)	1338 4.54 100 (%)	29444 100 (%)

distetään summaamalla kuhunkin klusteroinnista saatuun ryhmään kuuluvat muuttujat yhteen. Toisessa kullekin alkuperäiselle luokalle, tässä tapauksessa tehtävänannoille, annetaan ryhmästä kertova numero ja tämä laitetaan malliin alkuperäisen muuttujan sijaan. Lopuksi kontrastein voidaan vielä katsoa, eroavatko ryhmät toisistaan. Nämä kaksi tapaa tuottavat mallin, jonka vapausasteet ja uskottavuudet sekä näin ollen myös informaatiokriteerien arvot ovat identtiset.

Kontrastien mukaan toinen ja kolmas ryhmä eivät sittenkään eroa toisistaan, mutta ensimmäinen ja neljäs eroavat näistä kummastakin. Lopulliseen malliin tulee siis vain kolme luokkaa: tehtävänannot, joista on kirjoitettu hyvin, keskinkertaisesti ja huonosti (Taulukko 15). Näiden mukaiset arvosanajakaumat löytyvät Taulukosta 16.

Taulukko 15: Kolme ryhmää ja niihin sisältyvät tehtävänannot

Ryhmä	Tehtävänannot
Huonot	6, 7, 11, 13, 21, 25, 26, 29, 30, 31
Keskinkertaiset	1, 4, 5, 8, 9, 27, 28, 32, 33, 34
Hyvät	2, 3, 10, 12, 22, 23, 24

Taulukko 16: Arvosanajakaumat kolmella ryhmällä

Ryhmä	I	A	B	C	M	E	L	Yhteensä
Huonot	228 1.48 59.38	1642 10.65 61.00	3245 21.04 62.57	5531 35.86 55.92	3238 20.99 47.70	1210 7.84 38.23	331 2.15 24.74	15425 100 (%) 52.39
Keskin- kertaiset	138 1.26 35.94	892 8.15 33.14	1650 15.08 31.82	3534 32.29 35.73	2700 24.67 39.78	1389 12.69 43.89	640 5.85 47.83	10943 100 (%) 37.17
Hyvät	18 0.59 4.69	158 5.14 5.87	291 9.46 5.61	826 26.85 8.35	850 27.63 12.52	566 18.40 17.88	367 11.93 27.43	3076 100 (%) 10.45
Yht.	384 1.30 100 (%)	2692 9.14 100 (%)	5186 17.61 100 (%)	9891 33.59 100 (%)	6788 23.05 100 (%)	3165 10.75 100 (%)	1338 4.54 100 (%)	29444 100 (%)

Tarkasteltaessa näihin ryhmiin kuuluvia tehtävänantoja huomataan, että heikompia arvosanoja tuottaneet tehtävänannot ovat olleet suositumpia kuin hyviä arvosanoja tuottaneet. Kun 27 tehtävänantoa laitetaan järjestykseen kirjoittajamäärän mukaan (Taulukko 12), ovat kaikki hyvän arvosanan tuottaneet viimeisen 13 tehtävänannon joukossa, kun taas huonoimpia arvosanoja tuottaneita niistä on vain kaksi. Luultavasti kyse ei kuitenkaan ole siitä, että suositummasta aiheesta olisi vaikeampi saada hyvä arvosana, vaan siitä, että hyvät kirjoittajat valitsevat haastavampia aiheita. Koska opiskelijan aiemmasta osaamistasosta ei ole tietoa, ei valintamekanismiin kuitenkaan päästä käsiksi.

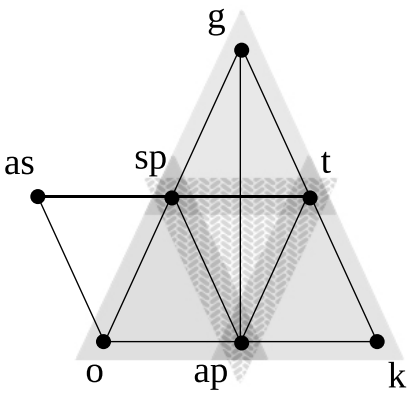
7 Tutkimustulokset

Edellisessä luvussa saatiin jo kiinnostava tulos, kun jaettiin tehtävänantoja ryhmiin klusterointimenetelmän avulla. Tarkastellaan seuraavaksi laadullisen luokittelun seurauksena saatavia tuloksia. Sovellusalan asiantuntijat ovat luokitelleet laadullisin perustein tehtävänantoja erityyppisiin ryhmiin, jotka voidaan jakaa tehtävänannon muotoiluun ja tehtävän viritykseen liittyviin muuttujiin. Lisäksi tarkastellaan luonteeltaan yksikäsitteisiä muuttujia, joita ovat kokeen arvosana, kieli sekä kokelaan sukupuoli. Nämä määräytyvät suoraan aineiston perusteella. Tarkasteluissa ovat mukana sekä kevään 2006 että kevään 2007 tulokset.

Loglineaarista mallia valittaessa oli aineistossa pieni virhe, sillä luokitteluissa ei ollut otettu huomioon niitä, jotka ovat kirjoittaneet äidinkielen kahdesti. Näitä henkilöitä oli aineistossa kuitenkin yhteensä vain reilut 200 ja heistäkin ongelma koski ainoastaan niitä, jotka olivat kirjoittaneet paremman arvosanan ensimmäisellä kerralla tai vaihtaneet kokeen kieltä. Täten virhe ei varmastikaan vaikuta parhaan mallin valikoitumiseen eikä suuritöisiä ajoja ollut tarpeen uusia. Loglineaarisen mallin pohjalta muodostetuissa ristiintaulukoissa virhe on korjattu.

7.1 Tehtävänannon muotoilu

Tarkastellaan aluksi tehtävänannon muotoiluun liittyvien muuttujien aihepiiri (ap), otsikko (o), teema (t) ja genre (g) suhteita kokeen arvosanan (as) ja kielen (k) sekä kokelaan sukupuolen (sp) kanssa. Parhaaksi loglineaariseksi malliksi valikoituu Schwarzin informaatiokriteerin perusteella `sp*k+o*as+sp*as+ap*g+ap*t*sp+ap*t*k+ap*o*sp+t*g*sp`. Sukupuolen ja kielen yhdysvaikutus ei kuitenkaan vaikuta kovin merkittävältä. Kun tarkastellaan kyseistä taulua (Taulukko 17), ei eroja ole juurikaan nähtävissä. Luultavasti yhdysvaikutuksen merkitsevyys johtuu suuresta otoskoosta ja siitä, että taulussa on vain yksi vapausaste, jolloin informaatiokriteeriin ei tule kyseisestä termistä juurikaan sakkoa, vaikka uskottavuus kuitenkin kasvaa parametrien määrän lisääntyessä. Koska termi `sp*k` vaikuttaa turhalta, jätetään se lopullisesta mallista pois, vaikka informaatiokriteerin arvo hiukan kasvaakin.



Kuva 7.1: Log-lineaarinen malli `sp*as+o*as+ap*g+sp*ap*t+k*ap*t+sp*ap*o+sp*t*g`. Viivat merkitsevät kahden välisiä yhdysvaikutuksia, harmaat kolmiot kolmen välisiä.

Lopullisissa tarkasteluissa riittää siis tarkastella mallia `o*as+sp*as+ap*g+ap*t*sp+ap*t*k+ap*o*sp+t*g*sp` (Kuva 7.1). Siinä on kahden muuttujan yhdysvaikutukset sukupulen ja arvosanan, otsikon ja arvosanan sekä aihepiirin ja genren välillä sekä kolmen muuttujan yhdysvaikutukset sukupuolen, aihepiirin ja teeman; kielen, aihepiirin ja teeman; sukupuolen, aihepiirin ja otsikon sekä sukupuolen, teeman ja genren välillä. Seitsemänulotteinen ristiintaulukko voidaan siis esittää 11 kaksiulotteisen tau-

Taulukko 17: Kielten jakauma sukupuolten suhteen.

	Suomi	Ruotsi	Yhteensä
Naiset	30052 92.80 57.69	2330 7.20 58.24	32382 100 (%) 57.73
Miehet	22043 92.95 42.31	1671 7.05 41.76	23714 100 (%) 42.27
Yhteensä	52095 92.87 100 (%)	4001 7.13 100 (%)	56096 100 (%)

lun avulla. Tarkastellaan aluksi arvosanojen jakaumia sukupuolten suhteen (Taulukko 18).

Taulukko 18: Arvosanajakaumat sukupuolen suhteen.

Ryhmä	I	A	B	C	M	E	L	Yhteensä
Naiset	356 1.10 35.96	1709 5.28 37.49	4435 13.69 46.68	10569 32.64 57.14	8909 27.51 66.04	4588 14.17 70.88	1819 5.62 70.20	32385 100 (%) 57.73
Miehet	634 2.67 64.04	2850 12.02 62.51	5065 21.36 53.32	7927 33.43 42.86	4582 19.32 33.96	1885 7.95 29.12	772 3.26 29.80	23715 100 (%) 42.27
Yht.	990 1.76 100 (%)	4559 8.13 100 (%)	9500 16.93 100 (%)	18496 32.97 100 (%)	13491 24.05 100 (%)	6473 11.54 100 (%)	2591 4.62 100 (%)	56100 100 (%)

Äidinkielen keväällä 2006 ja keväällä 2007 kirjoittaneista naisia on 32385 ja miehiä 23715, naisia on siis lähes kolme viidesosaa kokelaista. Kokonaisuudessaan ylioppilaskokeessa arvosanat annetaan suurin piirtein normaalijakauman mukaisesti, joten suurin osa, kolmannes kokelaista, on saanut keskimmäisen arvosanan cum laude (C). Magnan (M) on saanut noin neljäsosa, eximian (E) reilu kymmenesosa ja parhaan arvosanan laudaturin (L) vajaat viisi prosenttia kokelaista. Alempia arvosanoja on saatu vähän vähemmän: lubenterin (B) on saanut noin 15 prosenttia, approbaturin (A) vajaa kymmenesosa ja hylätyn arvosanan improbaturin (I) alle kaksi prosenttia kokelaista.

Taulukosta nähdään jälleen, että naisilla korkeampien arvosanojen osuus on keskimääräistä suurempi ja matalampien arvosanojen osuus keskimääräistä pienempi. Miehillä tilanne on toisin päin: hyviä arvosanoja on vähemmän, huonoja enemmän.

Katsotaan seuraavaksi, miltä näyttävät arvosanajakaumat sen mukaan, onko tehtävänannossa annettu aineelle valmis otsikko vai ei (Taulukko 19).

Taulukko 19: Arvosanajakaumat otsikon suhteen.

Ryhmä	I	A	B	C	M	E	L	Yhteensä
Valmis otsikko	216 2.11 21.82	783 7.65 17.17	1556 15.20 16.38	3229 31.55 17.46	2467 24.10 18.29	1356 13.25 20.95	628 6.14 24.24	10235 100 (%) 18.24
Ei otsikkoa	774 1.69 78.18	3776 8.23 82.83	7944 17.32 83.62	15267 33.29 82.54	11024 24.04 81.71	5117 11.16 79.05	1963 4.28 75.76	45866 100 (%) 81.76
Yht.	990 1.76 100 (%)	4559 8.13 100 (%)	9500 16.93 100 (%)	18496 32.97 100 (%)	13491 24.05 100 (%)	6473 11.54 100 (%)	2591 4.62 100 (%)	56100 100 (%)

Suurin osa kokelaista, suunnilleen neljä viidestä, on kirjoittanut sellaisesta tehtävänannosta, jolle ei ole annettu valmista otsikkoa. Valmiista otsikosta kirjoittaneilla on hieman enemmän kahta korkeinta arvosanaa ja A:n, B:n tai C:n kirjoittaneita on taas vähän vähemmän kuin muilla, mutta jostain syystä hylätyn arvosanan kohdalla heidän osuutensa onkin suurempi. Vaikka erot ovatkin merkitseviä, eivät ne kaiken kaikkiaan ole kovin suuria.

Viimeinen mallin kahden muuttujan yhdysvaikutuksiin liittyvä ristiintaulukko kuvaa genren jakaumaa aihepiireittäin (Taulukko 20).

Taulukko 20: Genren jakauma aihepiireittäin.

	Määritelty	Ei määr.	Viittaa tyyppiin	Yhteensä
Aatteellinen	— — —	1558 51.10 4.38	1491 48.90 8.99	3049 100 (%) 5.43
Media & viihde	118 1.13 3.02	2272 21.69 6.38	8087 77.19 48.74	10477 100 (%) 18.67
Urheilu	— — —	1672 100.00 4.70	— — —	1672 100 (%) 2.98
Talous	— — —	1406 100.00 3.95	— — —	1406 100 (%) 2.51

Jatkuu seuraavalla sivulla

Taulukko 20 – Jatkoa edelliseltä sivulta

	Määritelty	Ei määr.	Viittaa tyyppiin	Yhteensä
Yhteiskunta	153 0.86 3.91	14176 79.68 39.82	3463 19.46 20.88	17792 100 (%) 31.71
Kirjallisuus	445 10.71 11.38	2231 53.68 6.27	1480 35.61 8.92	4157 100 (%) 7.41
Kulttuuri	3091 28.31 79.03	5758 52.74 16.17	2069 18.95 12.47	10918 100 (%) 19.46
Luonto	104 4.11 2.66	2424 95.89 6.81	— — —	2529 100 (%) 4.51
Avoin	— — —	4102 100.00 11.53	— — —	4102 100 (%) 7.31
Yhteensä	3911 6.97 100 (%)	35600 63.46 100 (%)	16589 29.57 100 (%)	56101 100 (%)

Suurin osa kokelaista, lähes kolmasosa, on kirjoittanut yhteiskunnallisesta aihepiiristä. Seuraavaksi suosituimpia aihepiirejä ovat olleet kulttuuri sekä media ja viihde, joista molemmista on kirjoittanut vajaa viidesosa kokelaista. Muiden aihepiirien osuus onkin jo selvästi pienempi: kirjallisuusaiheen sekä sellaisen tehtävänannon, jossa aihepiiri on avoin, on valinnut noin seitsemän prosenttia vastaajista, aatteellisen ja luontoaiheen valinneita on viitisen prosenttia ja urheilu- sekä talousaiheen valinneita alle kolme prosenttia vastaajista.

Genreä tarkasteltaessa huomataan, että suunnilleen kaksi kolmesta on kirjoittanut tehtävänannosta, jossa kirjoitelman tekstilajia ei ole määritelty. Vajaat 30 prosenttia on kirjoittanut tehtävänannosta, jonka ohjeistus viittaa tekstityyppiin, ja noin seitsemän prosenttia sellaisesta tehtävänannosta, jossa on määritelty kirjoituksen tekstilaji.

Aatteellisesta tehtävänannosta kirjoittaneista noin puolet on valinnut tehtävänannon, jonka tekstilajia ei ole määritelty, ja toinen puolikas sellaisen, jossa on viittaus tekstityyppiin. Media- ja viihdeaiheen valinneista enemmän kuin kolme neljästä on kirjoittanut tehtävänannosta, jossa viitataan tekstityyppiin, viidesosa sellaisesta, jossa tekstilajia ei ole määritelty, ja vain reilu prosentti on valinnut valmiiksi määritellyn tekstilajin. Urheilu- ja talousaiheita on suomen- ja ruotsinkielisissä kokeissa vain yksi kumpaakin, ja kummassakaan näistä tehtävänannoista ei ole määritelty tekstilajia.

Yhteiskunnallisen aiheen valinneista neljä viidestä on kirjoittanut aineen, jonka

tekstilajia ei ole valmiiksi määritelty. Tehtävänannoista, joissa on viittauksia tekstityyppiin, on kirjoittanut suunnilleen viidennes ja vain alle prosentti on valinnut valmiiksi määritellyn tekstilajin. Kirjallisuusaiheesta kirjoittaneista reilu puolikas on valinnut määrittelemättömän tekstilajin ja kymmenesosa valmiiksi määritellyn. Kolmasosa on kirjoittanut tehtävänannosta, jonka ohjeistus viittaa tekstityyppiin.

Kulttuuriaiheesta kirjoittaneista noin 30 prosentilla tekstilaji on valmiiksi määritelty ja suunnilleen puolella sitä ei ole määritelty etukäteen. Viittauksia tekstityyppiin on vajaalla viidenneksellä. Luontoaiheen valinneista alle viisi prosenttia on kirjoittanut tehtävänannosta, jossa tekstilaji on valmiiksi määritelty, ja loput sellaisesta, jossa tekstilajia ei ole valmiiksi määritelty. Kaikki avoimen aihepiirin tehtävänannot ovat sellaisia, joissa tekstilajia ei ole määritelty.

Seuraavaksi tarkastellaan neljää kolmen muuttujan yhdysvaikutusta, joista kukin esitetään kahtena kaksisuuntaisena ristiintaulukkona. Aloitetaan teeman jakaumasta aihepiireittäin naisilla (Taulukko 21).

Taulukko 21: Teeman jakauma aihepiireittäin naisilla.

	Avoin	Väljästi määritelty	Määritelty	Yhteensä
Aatteellinen	392 19.52 2.81	712 35.46 10.59	904 45.02 7.72	2008 100 (%) 6.20
Media & viihde	1941 42.29 13.92	2389 52.05 35.53	260 5.66 2.22	4590 100 (%) 14.17
Urheilu	— — —	— — —	393 100.00 3.35	393 100 (%) 1.21
Talous	— — —	— — —	240 100.00 2.05	240 100 (%) 0.74
Yhteiskunta	2777 24.37 19.91	2609 22.89 38.81	6010 52.74 51.30	11396 100 (%) 35.19
Kirjallisuus	1269 38.41 9.10	498 15.07 7.41	1537 46.52 13.12	3304 100 (%) 10.20

Jatkuu seuraavalla sivulla

Taulukko 21 – Jatkoa edelliseltä sivulta

	Avoin	Väljästi määritelty	Määritelty	Yhteensä
Kulttuuri	3769 58.97 27.02	332 5.19 4.94	2290 35.83 19.55	6391 100 (%) 19.73
Luonto	978 78.74 7.01	183 14.73 2.72	81 6.52 0.69	1242 100 (%) 3.84
Avoin	2821 100.00 20.23	— — —	— — —	2821 100 (%) 8.71
Yhteensä	13947 43.07 100 (%)	6723 20.76 100 (%)	11715 36.17 100 (%)	32385 100 (%)

Naisista kolmasosa on kirjoittanut yhteiskunnallisesta aiheesta, viidennes kulttuuriaiheesta ja noin 15 prosenttia media- tai viihdeaiheesta. Kirjallisuusaihe on kiinnostanut kymmenesosaa naiskokelaista ja vajaat yhdeksän prosenttia on kirjoittanut tehtävänannosta, jonka aihepiiri on avoin. Aatteellisesta aiheesta on kirjoittanut noin kuusi prosenttia ja luontoaiheesta neljä. Urheilu- ja talousaihe on kiinnostanut alle prosenttia kokelaista. Teemaltaan avoimesta tehtävänannosta on kirjoittunut reilut 40 prosenttia naisista, noin kolmasosalla tehtävänannoissa teema on määritelty valmiiksi ja viidesosalla se on määritelty väljästi.

Niistä naisista, jotka ovat kirjoittaneet aatteellisesta aihepiiristä, on 45 prosenttia valinnut tehtävänannon, jossa teema on määritelty, 35 prosenttia tehtävänannon, jossa teema on väljästi määritelty, ja 20 prosenttia teemaltaan avoimen tehtävänannon. Media- ja viihdeaiheen valinneista vähän yli puolet on kirjoittanut tehtävänannosta, jossa teema on väljästi määritelty, ja reilut 40 prosenttia tehtävänannosta, jossa teema on avoin. Valmiiksi määritellyn teeman sisältävistä tehtävänannoista on kirjoittanut vain viisi prosenttia naisista.

Urheilu- ja talousaiheet ovat kummatkin sellaisia, joiden tehtävänannossa teema on valmiiksi määritelty. Yhteiskunnallisista aiheista kirjoittaneista noin puolet kirjoitti tehtävänannosta, jossa teema on määritelty, neljäsosa tehtävänannosta, jossa teema on avoin, ja lähes neljäsosa tehtävänannoista, joissa teema on väljästi määritelty. Kirjallisuusaiheen valinneista naisista määritelty teema on lähes puolella, avoin teema vajaalla 40 prosentilla ja väljästi määritelty teema 15 prosentilla.

Kulttuuriaiheen valinneista suurin osa, lähes kolme viidestä, on kirjoittanut tehtävänannosta, jossa on avoin teema. Määritelty teema on 35 prosentilla ja väljästi määritelty teema vain viidellä prosentilla. Luontoaiheesta kirjoittaneista neljä viidestä on valinnut teemaltaan avoimen tehtävänannon, noin 15 prosenttia väljästi määritellyn ja vajaat seitsemän prosenttia määritellyn teeman. Kaikki aihepiiriltään avoimet

tehtävänannot ovat myös teemaltaan avoimia.

Katsotaan vielä aihepiirien jakaumia teemoittain. Teemaltaan avoimista tehtävänannoista kirjoittaneista kulttuuriaiheesta on kirjoittanut vajaat 30 prosenttia ja yhteiskunnallisesta sekä aihepiiriltään avoimesta tehtävänannosta noin viidennes. Kirjallisuus- ja luontoaiheen valinneita on vajaat kymmenen prosenttia ja aatteellisesta aiheesta kirjoittaneita alle kolme prosenttia.

Teemaltaan väljästi määritellystä tehtävänannosta vajaat 40 prosenttia on kirjoittanut yhteiskunnallisesta aiheesta ja media- tai viihdeaiheesta lähes yhtä moni. Aatteellisen aiheen valinneita on kymmenesosa, kirjallisuusaiheen valinneita vähän vähemmän. Kulttuuri- ja luontoaiheen valitsi vain pieni osa.

Tehtävänannoista, joissa teema on valmiiksi määritelty, on puolet kirjoitettu yhteiskunnallisesta aiheesta. Kulttuuriaiheesta kirjoittaneita on viidesosa ja kirjallisuusaiheen valinneita 13 prosenttia. Aatteellisen aiheen on valinnut vajaat kahdeksan prosenttia ja muista aihepiireistä kirjoittaneita on enää kahdesta kolmeen prosenttia.

Seuraavaksi tarkastellaan miesten taulua teeman jakaumista aihepiireittäin (Taulukko 22).

Taulukko 22: Teeman jakauma aihepiireittäin miehillä.

	Avoin	Väljästi määritelty	Määritelty	Yhteensä
Aatteellinen	179 17.20 1.69	208 19.98 3.44	654 62.82 9.22	1041 100 (%) 4.39
Media & viihde	1841 31.28 17.42	3702 62.90 61.21	343 5.83 4.83	5886 100 (%) 24.82
Urheilu	— — —	— — —	1279 100.00 18.02	1279 100 (%) 5.39
Talous	— — —	— — —	1166 100.00 16.43	1166 100 (%) 4.92
Yhteiskunta	2953 46.17 27.93	1442 22.55 23.84	2001 31.29 28.20	6396 100 (%) 26.97
Kirjallisuus	223 26.17 2.11	394 46.24 6.51	235 27.58 3.31	852 100 (%) 3.59

Jatkuu seuraavalla sivulla

Taulukko 22 – Jatkoa edelliseltä sivulta

	Avoin	Väljästi määritelty	Määritelty	Yhteensä
Kulttuuri	3005 66.38 28.43	127 2.81 2.10	1395 30.82 19.66	4527 100 (%) 19.09
Luonto	1088 84.60 10.29	175 13.61 2.89	23 1.79 0.32	1286 100 (%) 5.42
Avoin	1282 100.00 12.13	— — —	— — —	1282 100 (%) 5.41
Yhteensä	10571 44.58 100 (%)	6048 25.50 100 (%)	7096 29.92 100 (%)	23715 100 (%)

Miehistä yhteiskunnallisista ja media- tai viihdeaiheista kirjoittaneita on kumpiakin lähes neljäsosa kokelaista ja kulttuuriaiheesta kirjoittaneita suunnilleen viidennes. Eroa naisiin on kahdessa ensimmäisessä reilusti, sillä naisista yhteiskunnallisesta aiheesta kiinnostuneita on kymmenen prosenttiyksikköä enemmän kun taas media- tai viihdeaiheen kiinnostus on kymmenisen prosenttiyksikköä pienempää. Muut aihepiirit ovat kiinnostaneet vain neljästä viiteen prosenttia kokelaista. Erityisen huomionherättävää on kirjallisuusaiheen vähäinen kiinnostus naisiin verrattuna: miehistä sen on valinnut alle neljä prosenttia, kun naisista aihe kiinnosti yli kymmenesosaa kokelaista. Vastaavasti urheilu- ja talousaiheet ovat olleet miesten mielestä kiinnostavampia; naisista näitä on valinnut vain prosentti kumpaakin.

Aatteellisesta aiheesta kirjoittaneista miehistä kolme viidestä on valinnut teemaltaan määritellyn tehtävänannon, viidesosa teemaltaan väljästi määritellyn tehtävänannon ja vajaa viidennes teemaltaan avoimen tehtävänannon. Vastaavat naiset ovat valinneet valmiiksi määriteltyjä tehtävänantoja lähes 20 prosenttiyksikköä vähemmän ja vastaavasti väljästi määriteltyjä noin 15 prosenttiyksikköä enemmän. Media- tai viihdeaiheen valinneista miehistä kolme viidestä on kirjoittanut tehtävänannosta, jossa teema on väljästi määritelty, kolmasosa tehtävänannosta, jossa teema on avoin, ja reilut viisi prosenttia tehtävänannosta, jossa on määritelty teema. Naisista teemaltaan avoimen tehtävänannon on valinnut noin kymmenen prosenttiyksikköä useampi ja väljästi määriteltyyn saman verran harvempi.

Yhteiskunnallisen aiheen valinneista miehistä vajaa puolet on kirjoittanut teemaltaan avoimesta tehtävänannosta, reilu viidennes teemaltaan väljästi määritellystä tehtävänannosta ja vajaa kolmannes teemaltaan määritellystä tehtävänannosta. Naisista avoimet tehtävänannot ovat kiinnostaneet vain neljäsosaa, kun taas valmiiksi määritellyt teemat jopa puolta. Kirjallisuusaiheen valinneista miehistä vajaa puolet on kirjoittanut tehtävänannosta, jossa teema on väljästi määritelty, ja reilut neljäs-

osat teemaltaan avoimesta ja määritellystä tehtävänannosta. Naisilla lukemat ovat aivan erilaiset: avoin teema on kiinnostanut 40 prosenttia, väljästi määritelty vain 15 prosenttia ja määritelty lähes puolta kirjoittajista.

Kulttuuriaiheesta kirjoittaneista kaksi kolmesta on kirjoittanut teemaltaan avoimesta tehtävänannosta ja vajaa kolmannes teemaltaan määritellystä tehtävänannosta. Teemaltaan väljästi määritellystä tehtävänannosta on kirjoittanut alle kolme prosenttia. Naisten osuudet ovat samansuuntaiset. Luontoaiheen valinneista miehistä jopa 85 prosenttia on kirjoittanut tehtävänannosta, jossa teema on avoin. Väljästi määritelty tehtävänanto on 14 prosentilla ja määritelty alle kahdella prosentilla. Naisten osuudet ovat jälleen samansuuntaiset, tosin valmiiksi määritelty teema on kiinnostanut useampaa ja avoin taas hieman harvempaa.

Tarkastellaan vielä aihepiirien jakaumat teemoittain. Sellaisista tehtävänannoista, joissa teema on avoin, on kirjoittanut lähes 30 prosenttia yhteiskunnallisesta ja kulttuuriaiheesta. Media- tai viihdeaiheesta on kirjoittanut vajaa viidennes ja luontoaiheesta sekä avoimesta aihepiiristä noin kymmenesosa. Aatteellisesta ja kirjallisuusaiheesta kirjoittaneita on vähän. Suurimmat erot naisiin löytyvät yhteiskunnallisten (naisilla osuus 19.91 %), kirjallisuus- (9.10 %) ja avoimen aiheen (20.23 %) kohdalta.

Teemaltaan väljästi määritellyistä tehtävänannoista kirjoittaneista kolme viidestä on valinnut media- tai viihdeaiheen ja vajaa neljäsosa yhteiskunnallisen aiheen. Kirjallisuusaiheen valinneita on vajaat seitsemän prosenttia, muiden aihepiirien osuudet ovat pieniä. Naiset eroavat miehistä eniten aatteellisen (naisten osuus 10.59 %), media- tai viihdeaiheen (35.53 %) ja yhteiskunnallisen aihepiirin (38.81 %) osalta.

Tehtävänannoista, joissa teema on määritelty, valitsi 30 prosenttia yhteiskunnallisen aiheen ja 20 prosenttia kulttuuriaiheen. Urheiluaiheesta on kirjoittanut 18 prosenttia ja talousaiheesta vain vähän harvempi. Aatteellisista tehtävänannoista kirjoittaneita on vajaat kymmenen prosenttia, muiden aihepiirien osuus on pieni. Eroa naisiin on eniten urheiluaiheessa (naisilla 3.35 %), talousaiheessa (2.05 %), yhteiskunnallisessa (51.30 %) ja kirjallisuusaiheessa (13.12 %).

Katsotaan seuraavaksi teeman jakaumaa aihepiireittäin suomenkielisessä kokeessa (Taulukko 23).

Taulukko 23: Teeman jakauma aihepiireittäin suomenkielisessä kokeessa.

	Avoin	Väljästi määritelty	Määritelty	Yhteensä
Aatteellinen	538 18.48 2.33	816 28.02 6.66	1558 53.50 9.32	2912 100 (%) 5.59
Media & viihde	3392 35.92 14.66	5932 62.83 48.43	118 1.25 0.71	9442 100 (%) 18.12

Jatkuu seuraavalla sivulla

Taulukko 23 – Jatkoa edelliseltä sivulta

	Avoin	Väljästi määritelty	Määritelty	Yhteensä
Urheilu	—	—	1490	1490
	—	—	100.00	100 (%)
	—	—	8.91	2.86
Talous	—	—	1278	1278
	—	—	100.00	100 (%)
	—	—	7.64	2.45
Yhteiskunta	5192	3869	7260	16321
	31.81	23.71	44.48	100 (%)
	22.45	31.59	43.43	31.33
Kirjallisuus	1492	871	1633	3996
	37.34	21.80	40.87	100 (%)
	6.45	7.11	9.77	7.67
Kulttuuri	6660	432	3380	10472
	63.60	4.13	32.28	100 (%)
	28.79	3.53	20.22	20.10
Luonto	1978	328	—	2307
	85.78	14.22	—	100 (%)
	8.55	2.68	—	4.43
Avoin	3878	—	—	3878
	100.00	—	—	100 (%)
	16.77	—	—	7.44
Yhteensä	23130	12248	16717	52095
	44.40	23.51	32.09	100 (%)
	100 (%)	100 (%)	100 (%)	

Yhteensä suomenkielisiä kirjoittajia on 52095 kokelasta. Heistä reilut 30 prosenttia on kirjoittanut yhteiskunnallisesta aiheesta ja noin 20 prosenttia kulttuuriaiheesta sekä media- tai viihdeaiheesta. Kirjallisuusaiheesta sekä aihepiiriltään avoimesta tehtävänannosta on kirjoittanut reilut seitsemän prosenttia kokelaista. Luontoaihe on kiinnostanut reilua neljää prosenttia ja urheilu sekä talous kahdesta kolmeen prosenttia kokelaista.

Teemaltaan avoimesta tehtävänannosta on kirjoittanut useampi kuin kaksi viidestä, teemaltaan määritellystä tehtävänannosta kolmasosa ja teemaltaan väljästi määritellystä tehtävänannosta vajaa viidennes kokelaista.

Niistä suomenkielisistä, jotka ovat valinneet aatteellisen tehtävänannon, on yli puolet kirjoittanut tehtävänannosta, jossa teema on määritelty, reilu neljäsosa tehtä-

vänannosta, jossa teema on määritelty väljästi, ja vajaa viidennes teemaltaan avoimesta tehtävänannossa. Media- ja viihdeaiheesta kirjoittaneista kolme viidesosaa on valinnut väljästi määritellyn teeman ja reilu kolmannes teemaltaan avoimen tehtävänannon. Määritellyn teeman on valinnut alle prosentti kokelaista.

Yhteiskunnallisen aiheen valinneista 44 prosenttia on kirjoittanut tehtävänannosta, jossa teema on määritelty, vajaa kolmannesta tehtävänannosta, jossa teema on avoin, ja alle viidesosa tehtävänannosta, jossa teema on määritelty väljästi. Kirjallisuusaiheesta kirjoittaneista lähes yhtä moni on kirjoittanut tehtävänannoista, joissa teema on avoin, ja tehtävänannoista, joissa teema on määritelty. Tehtävänannoista, joissa teema on vain väljästi määritelty, on kirjoittanut noin viidesosa suomenkielisistä.

Niistä kokelaista, jotka ovat kirjoittaneet kulttuuriaiheesta, on lähes kaksi kolmasosaa valinnut teemaltaan avoimen tehtävänannon ja vajaa kolmannes valmiiksi määritellyn teeman. Teemaltaan väljästi määritellyistä tehtävänannoista kirjoitti vain noin neljä prosenttia kokelaista. Luontoaiheesta kirjoittaneista 85 prosenttia on valinnut tehtävänannon, jossa teema on avoin, ja vajaat 15 prosenttia tehtävänannon, jossa teema on määritelty väljästi.

Teemaltaan avoimesta tehtävänannosta kirjoittaneista lähes 30 prosenttia on valinnut kulttuuriaiheen ja reilu viidesosa yhteiskunnallisen aiheen. Avoimesta aihepiiristä on kirjoittanut kuudesosa ja media- tai viihdeaiheesta vain vähän harvempi. Luontoaihe on kiinnostanut alle kymmenesosaa, kirjallisuus vielä harvempaa ja aatteellisesta aiheesta on kirjoittanut enää pari prosenttia.

Väljästi määritellyn teeman sisältävän tehtävänannon valinneista lähes puolet on kirjoittanut media- tai viihdeaiheesta ja vajaa kolmasosa yhteiskunnallisesta aiheesta. Kirjallisuus- ja aatteellisesta aiheesta kirjoittaneita on noin seitsemän prosenttia, kulttuuri- ja luontoaiheen valinneiden osuus on pieni.

Valmiiksi määritellyn teeman sisältävän tehtävänannon valinneista yli 40 prosenttia on kirjoittanut yhteiskunnallisesta aiheesta ja viidesosa kulttuuriaiheesta. Vajaa kymmenesosa on kirjoittanut kirjallisuus- ja aatteellisesta aiheesta ja vähän harvempi urheilu- ja talousaiheesta.

Seuraavaksi on vuorossa teeman jakauma aihepiireittäin ruotsinkielisessä kokeessa (Taulukko 24).

Taulukko 24: Teeman jakauma aihepiireittäin ruotsinkielisessä kokeessa.

	Avoin	Väljästi määritelty	Määritelty	Yhteensä
Aatteellinen	33 24.09 2.38	104 75.91 19.89	— — —	137 100 (%) 3.42
				Jatkuu seuraavalla sivulla

Taulukko 24 – Jatkoa edelliseltä sivulta

	Avoin	Väljästi määritelty	Määritelty	Yhteensä
Media & viihde	390 37.72 28.14	159 15.38 30.40	485 46.91 23.18	1034 100 (%) 25.84
Urheilu	— — —	— — —	182 100.00 8.70	182 100 (%) 4.55
Talous	— — —	— — —	128 100.00 6.12	128 100 (%) 3.20
Yhteis-kunta	538 36.62 38.82	182 12.39 34.80	749 50.99 35.80	1469 100 (%) 36.72
Kirjalli-suus	— — —	21 13.13 4.02	139 86.88 6.64	160 100 (%) 4.00
Kulttuuri	113 25.39 8.15	27 6.07 5.16	305 68.54 14.58	445 100 (%) 11.12
Luonto	88 39.64 6.35	30 13.51 5.74	104 46.85 4.97	222 100 (%) 5.55
Avoin	224 100.00 16.16	— — —	— — —	224 100 (%) 5.60
Yhteensä	1386 34.64 100 (%)	523 13.07 100 (%)	2092 52.29 100 (%)	4001 100 (%)

Ruotsinkielisen kokeen tehneitä on yhteensä 4001. Heistä kolmannes on kirjoittanut yhteiskunnallisesta aiheesta ja neljäsosa media- tai viihdeaiheesta. Kulttuuriaiheesta on kirjoittanut kymmenesosa ja luontoaiheesta sekä avoimesta aihepiiristä vajaat kuusi prosenttia. Aatteellinen, urheilu-, talous- ja kirjallisuusaihe on kiinnostanut kolmesta viiteen prosenttia kirjoittajista. Eniten eroa suomenkielisen kokeen tehneisiin löytyy media- tai viihdeaiheessa (suomalaisilla osuus 18.12 %), kirjallisuusaiheessa (7.67 %) ja kulttuuriaiheessa (20.10 %).

Sellaisista tehtävänannoista, joissa teema on määritelty, on kirjoittanut yli puolet

kokelaista. Teemaltaan avoimista tehtävänannoista on kirjoittanut kolmasosa ja väljästi määritellyistä teemoista 13 prosenttia ruotsinkielisistä. Suomenkielisten jakauma on täysin erinäköinen: avoin teema on kiinnostanut 44 prosenttia, väljästi määritelty teema 24 prosenttia ja määritelty teema 32 prosenttia kokelaista. Myös teemojen aihepiireittäiset jakaumat eroavat lähes täysin ruotsinkielisen kokeen tehneistä.

Aatteellisesta aihepiiristä kirjoittaneista kolme neljäsosaa on valinnut tehtävänannon, jossa teema on määritelty väljästi, ja neljäsosa tehtävänannon, jossa teema on jätetty avoimeksi. Tehtävänannosta, jossa teema on määritelty, ei kukaan ole ollut kiinnostunut. Koska aineistossa ovat mukana kaikki kirjoittajat eikä kyseessä ole otoksesta johtuva nolla, on tämäkin merkitty rakenteelliseksi nollaksi, vaikka sen lähtökohta onkin erilainen kuin niillä rakenteellisilla nollilla, jotka ovat syntyneet kokeen rakenteen myötä.

Yhteiskunnallisista aiheista kirjoittaneista puolet on valinnut tehtävänannon, jossa teema on määritelty valmiiksi, reilu kolmasosa tehtävänannon, jossa teema on jätetty avoimeksi ja reilut kymmenen prosenttia tehtävänannon, jossa teema on määritelty vain väljästi. Kirjallisuusaiheesta löytyy jälleen sellainen rakenteellinen nolla, joka on syntynyt, kun kukaan ei ole ollut kiinnostunut kyseisestä tehtävänannosta. Tehtävänannosta, jossa teema on määritelty valmiiksi, on kirjoittanut 87 prosenttia ja väljästi määritellystä teemasta 13 prosenttia.

Kulttuuriaiheesta kirjoittaneista kaksi kolmasosaa valitsi tehtävänannon, jossa teema on määritelty, ja neljäsosa tehtävänannon, jossa teema on avoin. Väljästi määritelty teema on kuudella prosentilla kokelaista. Luontoaiheesta kirjoittaneista vajaa puolet kirjoitti määritellyn teeman sisältävästä tehtävänannosta, 40 prosenttia teemaltaan avoimesta tehtävänannosta ja vajaat 14 prosenttia väljästi määritellyn teeman sisältävästä tehtävänannosta.

Teemaltaan avoimesta tehtävänannosta kirjoittaneista lähes 40 prosenttia on valinnut yhteiskunnallisen aiheen ja vajaat 30 prosenttia media- tai viihdeaiheen. Avoin aihepiiri on noin kuudesosalla, kulttuuriaihe vajaalla kymmenesosalla ja luontoaihe kuudella prosentilla. Aatteellinen aihepiiri on kiinnostanut ainoastaan paria prosenttia ruotsinkielisistä. Suurimmat erot suomenkielisiin löytyvät media- tai viihdeaiheen (suomenkielisillä 14.66 %), yhteiskunnallisen aiheen (22.45 %) ja kulttuuriaiheen (28.79 %) valinneista.

Sellaisesta tehtävänannosta, jossa teema on määritelty vain väljästi, on kolmasosa kirjoittanut yhteiskunnallisesta aiheesta ja vain vähän harvempi media- tai viihdeaiheesta. Aatteellisesta aiheesta on kirjoittanut viidesosa, muiden aihepiirien osuudet ovat pieniä. Suomenkieliset eroavat eniten aatteellisen (6.66 %) ja media- tai viihdeaiheen (48.43 %) osalta.

Tehtävänannoista, joissa teema on määritelty, on reilu kolmannes kokelaista kirjoittanut yhteiskunnallisista aiheista ja vajaa neljäsosa media- tai viihdeaiheesta. Kulttuuriaihe on kiinnostanut 15 prosenttia ja urheiluaihe yhdeksää prosenttia ruotsinkielisistä. Talous- ja kirjallisuusaiheen on valinnut noin kuusi prosenttia ja luontoaiheen viisi prosenttia kokelaista. Eroa suomenkielisiin on eniten aatteellisen (9.32 %), media- tai viihdeaiheen (0.71 %) sekä yhteiskunnallisen (43.43 %) aiheen kohdalla.

Katsotaan seuraavaksi otsikon vaikutusta eri tehtävänannoilla naisten valitsemissa tehtävänannoissa (Taulukko 25).

Taulukko 25: Otsikointi aihepiireittäin naisilla

	Valmis	Ei otsikkoa	Yhteensä
Aatteellinen	— — —	2008 100.00 7.61	2008 100 (%) 6.20
Media & viihde	802 17.47 13.40	3788 82.53 14.35	4591 100 (%) 14.17
Urheilu	— — —	393 100.00 1.49	393 100 (%) 1.21
Talous	240 100.00 4.01	— — —	240 100 (%) 0.74
Yhteiskunta	580 5.09 9.69	10815 94.91 40.97	11396 100 (%) 35.19
Kirjallisuus	1301 39.38 21.74	2003 60.62 7.59	3304 100 (%) 10.20
Kulttuuri	241 3.77 4.03	6150 96.23 23.30	6391 100 (%) 19.73
Luonto	— — —	1242 100.00 4.70	1242 100 (%) 3.84
Avoin	2820 100.00 47.13	— — —	2821 100 (%) 8.71
Yhteensä	5985 18.48 100 (%)	26400 81.52 100 (%)	32385 100 (%)

Naisista vajaa viidesosa on kirjoittanut tehtävänannosta, jossa on annettu valmis otsikko, loput ovat valinneet tehtävän, jossa otsikko on laadittava itse. Kaikki aatteelliset, urheilu- ja luontoaiheet ovat sellaisia, joiden tehtävänannoissa ei ole annettu valmista otsikkoa. Talous- ja avoimessa aiheessa kaikki tehtävänannot sen sijaan sisältävät valmiin otsikon.

Media- tai viihdeaiheen valinneista useampi kuin neljä viidestä on kirjoittanut tehtävänannosta, jolla ei ole valmista otsikkoa, ja vajaa viidennes valmiista otsikosta. Yhteiskunnallisista aiheista kirjoittaneista vain viisi prosenttia on kirjoittanut valmiista otsikosta, loput ovat valinneet tehtävänannon, joissa otsikko on vapaavalintainen.

Kirjallisuusaiheen valinneista naisista kolme viidestä on saanut keksiä otsikon itse, kaksi viidestä on kirjoittanut valmiista otsikosta. Kulttuuriaiheesta kirjoittaneet ovat lähes kaikki valinneet tehtävänannon, jossa otsikko on vapaavalintainen. Valmiista otsikosta on kirjoittanut alle neljä prosenttia.

Katsotaan vielä aihepiirien jakaumia sen mukaan, onko tehtävänannossa annettu valmis otsikko vai ei. Valmiista otsikosta kirjoittaneista suurin osa, lähes puolet, on kirjoittanut avoimesta aihepiiristä. Viidesosalla aiheena on kirjallisuus, 13 prosentilla media tai viihde ja vain vähän vähemmällä yhteiskunta. Talous- ja kulttuuriaiheen kiinnostus on ollut vain nelisen prosenttia.

Niistä, jotka ovat kirjoittaneet tehtävänannosta, jossa otsikko on vapaavalintainen, on kaksi viidestä valinnut yhteiskunnallisen aiheen, lähes neljäsosa kulttuuriaiheen ja vajaat 15 prosenttia media- tai viihdeaiheen. Aatteellisen ja kirjallisuusaiheen on valinnut alle kahdeksan prosenttia, urheilu- ja luontoaiheen osuudet ovat vielä pienempiä.

Katsotaan seuraavaksi otsikoinnin jakaumia aihepiireittäin miesten kohdalla (Taulukko 26).

Taulukko 26: Otsikointi aihepiireittäin miehillä.

	Valmis	Ei otsikkoa	Yhteensä
Aatteellinen	— — —	1041 100.00 5.35	1041 100 (%) 4.39
Media & viihde	331 5.62 7.79	5555 94.38 28.54	5886 100 (%) 24.82
Urheilu	— — —	1279 100.00 6.57	1279 100 (%) 5.39

Jatkuu seuraavalla sivulla

Taulukko 26 – Jatkoa edelliseltä sivulta

	Valmis otsikko	Ei otsikkoa	Yhteensä
Talous	1166 100.00 27.44	— — —	1166 100 (%) 4.92
Yhteiskunta	241 3.77 5.67	6155 96.23 31.62	6396 100 (%) 26.97
Kirjallisuus	447 52.46 10.52	405 47.54 2.08	852 100 (%) 3.59
Kulttuuri	783 17.30 18.42	3744 82.70 19.23	4527 100 (%) 19.09
Luonto	— — —	1286 100.00 6.61	1286 100 (%) 5.42
Avoin	1282 100.00 30.16	— — —	1282 100 (%) 5.41
Yhteensä	4250 17.92 100 (%)	19465 82.08 100 (%)	23715 100 (%)

Valmiista otsikosta kirjoittaneiden miesten osuus on yhtä suuri kuin naisten, vajaa viidesosa. Media- tai viihdeaiheen valinneista miehistä noin kuusi prosenttia on kirjoittanut valmiista otsikosta, loput ovat keksineet sen itse. Naisten joukossa valmiista otsikosta kirjoittaneita on paljon enemmän, lähes viidesosa. Yhteiskunnallisen aiheen valinneissa valmiista otsikosta kirjoittaneiden miesten osuus on alle neljä prosenttia eikä naistenkaan kovin paljon suurempi.

Kirjallisuusaiheesta kirjoittaneista mieskokelaista lähes yhtä moni on kirjoittanut valmiista otsikosta ja keksinyt otsikon itse. Naisista valmiista otsikosta on kirjoittanut vähän harvempi, noin 40 prosenttia. Kulttuuriaiheesta kiinnostuneista miehistä vajaa viidennes on valinnut valmiin otsikon, reilut 80 prosenttia on keksinyt otsikon itse. Naisilla valmiin otsikon valinneita on alle neljä prosenttia.

Valmiista otsikosta kirjoittaneet miehet ovat valinneet useimmiten talous- tai avoimen aiheen, näiden kummankin osuus on suunnilleen 30 prosenttia. Vajaa viidennes on kirjoittanut kulttuuriaiheesta ja suunnilleen kymmenesosa kirjallisuusaiheesta. Media- tai viihdeaiheen kiinnostus on vajaat kahdeksan prosenttia ja yhteiskunnalli-

sen aiheen alle kuusi. Eniten eroa naisiin on talousaiheen (naisilla 4.01 %), kirjallisuusaiheen (21.74 %), kulttuuriaiheen (4.03 %) ja avoimen aiheen (47.13 %) kohdalla.

Otsikon itse keksineet ovat valinneet mieluiten yhteiskunnallisia tai media- ja viihdeaiheita, kummankin kohdalla kirjoittajia on suunnilleen 30 prosenttia. Kulttuuriaiheen on valinnut vajaa viidesosa. Media- tai viihdeaiheen, urheiluaiheen ja luontoaiheen kirjoittajien osuus vaihtelee viiden ja seitsemän välillä, kirjallisuusaihe on kiinnostanut vain kahta prosenttia näistä miehistä. Suurimmat erot naisiin löytyvät media- tai viihdeaiheesta (14.35 %) ja yhteiskunnallisesta aiheesta (40.97 %) kirjoittaneista.

Nyt on jäljellä enää teeman jakaumat genren suhteen. Katsotaan jälleen ensin naisten jakaumia (Taulukko 27).

Taulukko 27: Teeman jakaumat genren suhteen naisilla.

	Avoin	Väljästi määritelty	Määritelty	Yhteensä
Määritelty	1955 87.90 14.02	— — —	269 12.10 2.30	2224 100 (%) 6.87
Ei määritelty	6714 32.30 48.14	4086 19.66 60.78	9984 48.04 85.22	20784 100 (%) 64.18
Viittaa tyyppiin	5278 56.29 37.84	2637 28.12 39.22	1462 15.59 12.48	9377 100 (%) 28.95
Yhteensä	13947 43.07 100 (%)	6723 20.76 100 (%)	11715 36.17 100 (%)	32385 100 (%)

Niistä naisista, jotka ovat kirjoittaneet valmiiksi määritellystä genrestä, on lähes 90 prosenttia kirjoittanut tehtävänannosta, jossa teema on avoin, ja reilu kymmenesosa on valinnut tekstilajinsa itse. Niistä tehtävänannoista kirjoittaneita, joiden tekstilajia ei ole määritelty, on lähes puolet valinnut teemaltaan määritellyn tehtävänannon, kolmasosa teemaltaan avoimen tehtävänannon ja viidennes väljästi määritellyn tehtävänannon.

Kirjoittajista, jotka ovat valinneet tehtävänannon, jossa ohjeistus viittaa tekstityyppiin, on yli puolet valinnut teemaltaan avoimen tehtävänannon, vajaat 30 prosenttia väljästi määritellyn teeman ja reilut 15 prosenttia määritellyn teeman sisältävän tehtävänannon.

Teemaltaan avoimista tehtävänannoista kirjoittaneista lähes puolet on valinnut tehtävän, jossa genreä ei ole määritelty, lähes 40 prosenttia tehtävän, jossa ohjeistus viittaa tekstityyppiin, ja 14 prosenttia tehtävän, jossa tekstilaji on määritelty.

Väljästi määritellyn teeman sisältävästä tehtävänannosta kirjoittaneista kolme viidestä on valinnut tehtävän, jossa tekstilajia ei ole määritelty, ja loput tehtävän, jossa ohjeistus viittaa tekstityyppiin. Määritellyn teeman sisältävistä tehtävänannoista kirjoittaneista 85 prosenttia on valinnut tehtävän, jossa tekstilajia ei ole määritelty, 12 prosenttia tehtävän, jossa viitataan tekstityyppiin, ja vain kaksi prosenttia on valinnut genrensä itse.

Viimeiseksi katsotaan vielä miesten taulu, josta löytyvät teeman jakaumat genren suhteen (Taulukko 28).

Taulukko 28: Teeman jakaumat genren suhteen miehillä.

	Avoin	Väljästi määritelty	Määritelty	Yhteensä
Määritelty	1521 90.16 14.39	— — —	166 9.84 2.34	1687 100 (%) 7.11
Ei määritelty	6486 43.78 61.36	2374 16.02 39.25	5956 40.20 83.93	14816 100 (%) 62.48
Viittaa tyyppiin	2564 35.55 24.26	3674 50.94 60.75	974 13.51 13.73	7212 100 (%) 30.41
Yhteensä	10571 44.58 100 (%)	6048 25.50 100 (%)	7096 29.92 100 (%)	23715 100 (%)

Määritellystä genrestä kirjoittaneista miehistä 90 prosenttia on valinnut teemaltaan avoimen tehtävänannon, loput tehtävänannon, jossa teema on määritelty. Naisten osuudet ovat melko samanlaisia. Vapaasti genrensä valinneista reilut 40 prosenttia on kirjoittanut teemaltaan avoimesta tehtävänannosta, samoin valmiiksi määritellystä teemasta. Väljästi määritellyn teeman on valinnut 16 prosenttia. Naisiin löytyy eroa lähinnä avoimen (naisten vastaava osuus 32.30 %) ja määritellyn (48.04 %) teeman kohdalla

Niistä miehistä, joiden tehtävänannosta löytyy viittauksia tekstityyppiin, on puolet kirjoittanut väljästi määritellystä teemasta, reilulla kolmanneksella teema on avoin ja 13 prosentilla valmiiksi määritelty. Naisista sen sijaan avoimesta teemasta on kiinnostunut yli puolet, väljästi määritellystä teemasta vain alle 30 prosenttia.

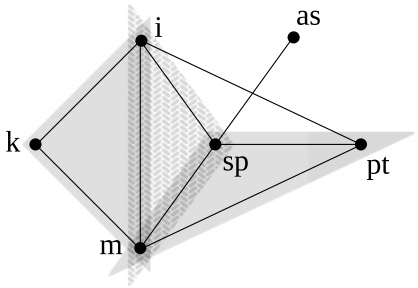
Teemaltaan avoimesta tehtävänannosta kirjoittaneista kolme viidestä on päässyt valitsemaan genrensä itse, neljäsosalla tehtävänannossa on viittauksia tekstityyppiin ja alle 15 prosentilla genre on valmiiksi määritelty. Naisiin verrattuna eroa on niiden tehtävänantojen suhteen, joissa genreä ei ole määritelty (naisilla 48.14 %) tai tekstityyppiin on viittauksia ohjeistuksessa (37.84 %).

Väljästi määritellystä teemasta kirjoittaneista kolme viidestä on valinnut tehtävänannon, jossa on viittauksia tekstityyppiin, lopuilla genren on saanut valita vapaasti. Naisten osuudet ovat vastakkaiset. Valmiiksi määritellystä teemasta kirjoittaneista 84 prosenttia on päässyt valitsemaan genrensä itse, 14 prosentilla on viittauksia tekstityyppiin ja vain reilut kaksi prosenttia on kirjoittanut valmiiksi määritellystä genrestä. Naisten osuudet ovat samansuuntaisia.

7.2 Tehtävän viritys

Tarkastellaan seuraavaksi tehtävän viritykseen liittyvien muuttujien pohjatekstin (pt), vaaditun intertekstuaalisen varannon (i) sekä tehtävänannon moodin (m) yhteyttä arvosanaan, kokeen kieleen sekä kokelaan sukupuoleen. Parhaaksi loglineaariseksi malliksi valikoituu Schwarzin informaatiokriteerin perusteella `sp*as+pt*i+sp*pt*m+sp*i*m+k*i*m`. Siinä on kahden muuttujan yhdysvaikutus sukupuolen ja arvosanan sekä pohjatekstin ja vaaditun intertekstuaalisen varannon välillä sekä kolmen muuttujan yhdysvaikutukset sukupuolen, pohjatekstin ja moodin; sukupuolen, intertekstuaalisen varannon ja moodin sekä kielen, intertekstuaalisen varannon ja moodin välillä.

Koska sukupuolen ja arvosanan välistä taulua on tarkasteltu jo edellisessä luvussa, aloitetaan tarkastelemalla vaaditun intertekstuaalisen varannon jakaumaa pohjatekstien mukaan (Taulukko 29).



Kuva 7.2: Log-lineaarinen malli `sp*as+pt*i+sp*pt*m +sp*i*m+k*i*m`. Viivat merkitsevät kahden välisiä yhdysvaikutuksia, harmaat kolmiot kolmen välisiä.

Taulukko 29: Vaaditun intertekstuaalisen varannon jakauma pohjatekstien mukaan.

	Oma	Kulttuuri	Yhteiskunta	Historia	Kult. ja yht.kunta	Yhteensä
Ei pohjatekstiä	9396 47.93 44.07	3152 16.08 22.57	5383 27.46 42.61	— — —	1672 8.53 26.00	19603 100 (%) 34.94
Kuva	1133 60.62 5.31	736 39.38 5.27	— — —	— — —	— — —	1869 100 (%) 3.33
Asiateksti	413 3.57 1.94	4652 40.17 33.31	4775 41.23 37.79	— — —	1740 15.03 27.06	11580 100 (%) 20.64
Essee, kolumni, pakina	4566 64.56 21.41	1983 28.04 14.20	339 4.79 2.68	— — —	185 2.62 2.88	7073 100 (%) 12.62
Fiktio: proosa	— — —	1008 49.05 7.22	— — —	— — —	1047 50.95 16.28	2055 100 (%) 3.66

Jatkuu seuraavalla sivulla

Taulukko 29 – Jatkoa edelliseltä sivulta

	Oma	Kult- tuuri	Yhteis- kunta	Historia	Kult. ja yht.kunta	Yhteensä
Fiktio: runo	— — —	911 96.20 6.52	— — —	— — —	36 3.80 0.56	947 100 (%) 1.69
Usk. & aatteell.	920 37.13 4.31	506 20.42 3.62	— — —	— — —	1052 42.45 16.36	2478 100 (%) 4.42
Sitaatti	4895 53.41 22.96	— — —	1956 21.34 15.48	1745 19.04 100.00	569 6.21 8.85	9165 100 (%) 16.34
Useita eri	— — —	1019 76.62 7.30	181 13.61 1.43	— — —	130 9.77 2.02	1330 100 (%) 2.37
Yhteensä	21323 38.01 100 (%)	13967 24.90 100 (%)	12634 22.52 100 (%)	1745 3.11 100 (%)	6431 11.46 100 (%)	56100 100 (%)

Kolmasosa kokelaista on kirjoittanut sellaisesta tehtävänannosta, jossa ei ole lainkaan pohjatekstiä. Joka viides on kirjoittanut asiatekstin pohjalta, kuudesosalla tehtävänannossa on sitaatti ja 13 prosenttia on kiinnostanut esseestä, kolumnista tai pakinasta. Muiden aineistojen osuudet ovat pieniä: kuvista on kirjoittanut kolme prosenttia, proosasta vajaat neljä, runosta alle kaksi ja uskonnollisesta tai aatteellisesta pohjatekstistä reilut neljä prosenttia kokelaista. Noin kaksi prosenttia on kirjoittanut tehtävänannosta, jossa on useita erityyppisiä aineistoja.

Sellaisesta tehtävänannosta, joka vaatii vain omaa kokemusta, on kirjoittanut suunnilleen kaksi viidestä kokelaasta. Kulttuuritietoa vaativan tehtävänannon on valinnut neljäsosa ja lähes yhtä moni on tarvinnut aineessaan yhteiskunnallista tietoa. Sekä kulttuurista että yhteiskunnallista tietoa on odotettu joka kymmenenneltä, historiatietoa vain kolmelta prosentilta kokelaista.

Niistä, joilla ei ole ollut aineistoa kirjoituksensa pohjana, on lähes puolet kirjoittanut tehtävänannosta, jossa ei vaadita muuta kuin omaa kokemusta. Reilulta neljäsosalta on odotettu yhteiskunnallista tietoa ja 16 prosentilta kulttuuritietoa. Näitä molempia pitäisi löytyä vajaalta kymmenesosalta.

Kuvaa kirjoituksensa pohjana käyttäneistä kolme viidestä on kirjoittanut vain oman kokemuksensa pohjalta, lopuilta on odotettu kulttuurin tuntemusta. Asiateksteistä kirjoittaneista 40 prosentilta on vaadittu kulttuuritietoa ja yhtä isolta osalta yhteiskunnan tuntemusta. Kummankin hallitsemista on odotettu 15 prosentilta, ja vain kaksi prosenttia on valinnut aiheen, jossa ei tarvita muuta kuin omaa kokemusta. Esseestä, kolumnista tai pakinasta kirjoittaneista 65 prosenttia on selvinnyt ai-

heestaan pelkän oman kokemuksen perusteella. Vajaat 30 prosenttia on saanut esitellä kulttuuritietojaa ja alle viisi prosenttia yhteiskunnan tuntemustaan. Sekä kulttuurista että yhteiskunnallista tietoa on vaadittu vajaalta kolmelta prosentilta kokelaista.

Proosasta kirjoittaneista puolelta on odotettu kulttuurin tuntemista, toiselta puolikkaalta lisäksi yhteiskunnallista tietoa. Runon pohjatekstikseen valinneista 96 prosentilta on edellytetty kulttuuritietoa ja neljältä prosentilta sekä kulttuuri- että yhteiskunnallista tietoa. Uskonnollisesta tai aatteellisesta aineistosta kirjoittaneista reilut kaksi viidesosaa on saanut esitellä kulttuuritietonsa lisäksi myös yhteiskunnan tuntemustaan. Pelkän oman kokemuksen perusteella on voinut kirjoittaa vajaat 40 prosenttia ja viidesosalta on odotettu kulttuuritietoa.

Sitaatin pohjalta kirjoittaneista yli puolet ei ole tarvinnut muuta kuin omaa kokemusta kirjoituksensa pohjaksi. Viidesosalta on odotettu yhteiskunnallista tietoa ja vain vähän harvempi on päässyt esittelemään historian tuntemustaan. Kulttuuri- ja yhteiskunnallista tietoa on odotettu kuudelta prosentilta. Niistä kokelaista, jotka ovat kirjoittaneet usean erityyppisen aineiston pohjalta, on kolmelta neljästä vaadittu kulttuuritietoa ja 14 prosentilta yhteiskunnallista tietoa. Kymmenesosalta odotettiin kummankin tuntemusta.

Tarkastellaan seuraavaksi moodien jakaumia pohjatekstien mukaan naisten valitsemissa tehtävänannoissa (Taulukko 30).

Taulukko 30: Moodien jakauma pohjatekstien mukaan naisilla.

	Korkea-kult.	Yleis-kielinen	Arkinen	Interper-soonainen	Yhteensä
Ei pohja-tekstiä	646 6.11 34.79	7385 69.85 33.15	423 4.00 100.00	2119 20.04 27.07	10573 100 (%) 32.65
Kuva	456 35.76 24.56	802 62.90 3.60	— — —	17 1.33 0.22	1275 100 (%) 3.94
Asia-teksti	— — —	4341 64.25 19.49	— — —	2415 35.75 30.84	6756 100 (%) 20.86
Essee, kolumni, pakina	— — —	1773 57.60 7.96	— — —	1305 42.40 16.67	3078 100 (%) 9.50
Fiktio: proosa	— — —	1673 100.00 7.51	— — —	— — —	1673 100 (%) 5.17

Jatkuu seuraavalla sivulla

Taulukko 30 – Jatkoa edelliseltä sivulta

	Korkea-kult.	Yleis-kielinen	Arkinen	Interper-soonainen	Yhteensä
Fiktio: runo	753 95.80 40.55	33 4.20 0.15	— — —	— — —	787 100 (%) 2.43
Usk. & aatteell.	— — —	904 55.94 4.06	— — —	712 44.06 9.09	1616 100 (%) 4.99
Sitaatti	— — —	4522 81.42 20.30	— — —	1032 18.58 13.18	5554 100 (%) 17.15
Useita eri	2 0.19 0.11	843 78.49 3.78	— — —	229 21.32 2.93	1074 100 (%) 3.32
Yhteensä	1857 5.73 100 (%)	22276 68.78 100 (%)	423 1.31 100 (%)	7829 24.17 100 (%)	32385 100 (%)

Suurin osa naiskokelaista, yli kaksi kolmasosaa, on kirjoittanut moodiltaan yleiskielisestä tehtävänannosta. Interpersoonainen moodi on neljäsosalla, korkeakulttuurinen alle kuudella prosentilla ja arkinen enää reilulla prosentilla kokelaiden valitsemista tehtävänannoista.

Niistä kokelaista, joiden valitsemassa tehtävänannossa ei ole annettu lainkaan aineistoa, on 70 prosenttia kirjoittanut moodiltaan yleiskielisestä tehtävänannosta. Interpersoonaisen tehtävänannon on valinnut viidesosa, korkeakulttuurinen ja arkinen moodi on kiinnostanut vain harvaa. Kuvan pohjalta kirjoittaneista naisista reilut 60 prosenttia on kirjoittanut yleiskielisestä tehtävänannosta ja 35 prosenttia korkeakulttuurisesta. Interpersoonainen tehtävänanto on kiinnostanut vain prosenttia näistä kokelaista.

Asiatekstistä kirjoittaneista naisista lähes kaksi kolmasosaa on kirjoittanut yleiskielisestä tehtävänannosta, loput interpersoonaisesta. Esseen, kolumnin tai pakinan pohjatekstikseen valinneista vajaat 60 prosenttia on kirjoittanut yleiskielisestä tehtävänannosta, loput interpersoonaisesta. Kaikki proosasta kirjoitetut aineet ovat yleiskielisiä ja runoistakin valtaosa. Yleiskielisiä tehtävänantoja näistä on vain neljä prosenttia.

Uskonnollisen tai aatteellisen pohjatekstin valinneista reilu puolikas on kirjoittanut yleiskielisestä tehtävänannosta, loput interpersoonaisesta. Sitaatin pohjalta kirjoittaneista edellisiä on noin 80 prosenttia, jälkimmäisiä 20 prosenttia. Useista erityyppisistä pohjateksteistä kirjoittaneista lähes 80 prosenttia on kirjoittanut yleiskielisestä ja reilut 20 prosenttia interpersoonaisesta tehtävänannosta. Korkeakulttuurisesta

tehtävänannosta kiinnostuneita on heistä vain häviävän pieni osa: kaksi henkilöä.

Korkeakulttuurisen moodin sisältävästä tehtävänannosta kirjoittaneista 40 prosenttia on kirjoittanut runosta ja 25 prosenttia kuvasta. Kahta useasta pohjatekstistä kirjoittanutta vaille loput 35 prosenttia ovat kirjoittaneet tehtävänannosta, johon ei liity pohjatekstiä. Kumpikin arkisen moodin sisältävistä tehtävänannoista on sellaisia, joihin ei liity pohjatekstiä.

Yleiskielisen tehtävänannon valinneista kolmasosa on kirjoittanut tehtävänannosta, jossa ei ole pohjatekstiä, ja sekä asiatekstistä että sitaatin pohjalta on kirjoittanut viidesosa. Esseestä, kolumnista tai pakinasta ja proosasta kirjoittaneita on vajaat kahdeksan prosenttia. Muista aineistoista eli kuvasta, runosta, uskonnollisesta tai aatteellisesta ja useita eri tyyppisiä pohjatekstejä sisältävistä tehtävänannoista on kirjoittanut korkeintaan nelisen prosenttia kustakin.

Interpersoonaisen moodin sisältävästä tehtävänannosta kirjoittaneista naiskokelaista noin 30 prosenttia on kirjoittanut asiatekstistä ja vain vähän harvempi sellaisesta tehtävästä, johon ei liity aineistoa. Essee, kolumni tai pakina on kiinnostanut kuudesosaa ja sitaatti vähän harvempaa. Uskonnollisen tai aatteellisen tekstin on valinnut vajaa kymmenesosa, kuvan tai useita erilaisia aineistoja sisältäviä tehtävänantoja on valinnut vain harva.

Tarkastellaan seuraavaksi moodien jakaumia pohjatekstien suhteen miehillä (Taulukko 31).

Taulukko 31: Moodien jakauma pohjatekstien mukaan miehillä.

	Korkea-kult.	Yleis-kielinen	Arkinen	Interper-soonainen	Yhteensä
Ei pohja-tekstiä	444 4.92 52.05	7159 79.28 48.33	191 2.12 100.00	1236 13.69 15.73	9030 100 (%) 38.08
Kuva	251 42.26 29.43	331 55.72 2.23	— — —	12 2.02 0.15	594 100 (%) 2.50
Asia-teksti	— — —	2470 51.20 16.67	— — —	2354 48.80 29.96	4824 100 (%) 20.34
Essee, kolumni, pakina	— — —	595 14.89 4.02	— — —	3400 85.11 43.27	3995 100 (%) 16.85

Jatkuu seuraavalla sivulla

Taulukko 31 – Jatkoa edelliseltä sivulta

	Korkea-kult.	Yleis-kielinen	Arkinen	Interper-soonainen	Yhteensä
Fiktio: proosa	—	382	—	—	382
	—	100.00	—	—	100 (%)
	—	2.58	—	—	1.61
Fiktio: runo	158	3	—	—	161
	98.14	1.86	—	—	100 (%)
	18.52	0.02	—	—	0.68
Usk. & aatteell.	—	654	—	208	862
	—	75.87	—	24.13	100 (%)
	—	4.42	—	2.65	3.63
Sitaatti	—	3045	—	566	3611
	—	84.33	—	15.67	100 (%)
	—	20.56	—	7.20	15.23
Useita eri	—	174	—	82	256
	—	67.97	—	32.03	100 (%)
	—	1.17	—	1.04	1.08
Yhteensä	853	14813	191	7858	23715
	3.60	62.46	0.81	33.14	100 (%)
	100 (%)	100 (%)	100 (%)	100 (%)	

Miehistä yli 60 prosenttia on kirjoittanut yleiskielisestä ja kolmasosa interpersoonaisesta tehtävänannosta. Kiinnostus muihin moodeihin on ollut pientä: korkeakulttuurisesta tehtävänannosta on kirjoittanut vajaat neljä prosenttia ja arkisesta alle prosentti miehistä. Naisilla osuudet ovat samansuuntaisia, tosin interpersoonainen teksti on kiinnostanut vain 24 prosenttia.

Sellaisesta tehtävänannosta, jossa ei ole annettu aineistoa, on lähes 80 prosenttia kirjoittanut yleiskielisestä ja noin 14 prosenttia interpersoonaisesta tehtävänannosta. Korkeakulttuurisesta ja arkisesta tehtävänannosta on kirjoittanut vain harva. Naisiin verrattuna eroa on lähinnä yleiskielisessä moodissa (naisten osuus 69.85 %). Kuvasta kirjoittaneista miehistä yli puolet on kirjoittanut yleiskielisestä tehtävänannosta ja reilut 40 prosenttia korkeakulttuurisesta. Interpersoonaisesta tehtävänannosta kirjoittaneita on vain kaksi prosenttia. Naisilla erot ovat suurempia, mutta samansuuntaisia.

Asiatekstin pohjalta kirjoittaneista miehistä puolet on kirjoittanut yleiskielisestä ja puolet interpersoonaisesta tehtävänannosta. Naisista yleiskielisestä tehtävänannosta kirjoittaneita on 64 prosenttia. Esseestä, kolumnista tai pakinasta kirjoittaneista miehistä yleiskielisen tehtävänannon on valinnut vain 15 prosenttia, loput ovat kirjoittaneet interpersoonaisesta tehtävänannosta. Naisilla sen sijaan yleiskielisestä tehtävänannosta kirjoittaneita on lähes 60 prosenttia. Runosta kirjoittaneista lähes kaikki ovat kirjoittaneet korkeakulttuurisesta tehtävänannosta, yleiskielisestä kirjoittaneita

on alle kaksi prosenttia.

Uskonnollisen tai aatteellisen tekstin pohjalta kirjoittaneista kolme neljäsosaa on valinnut yleiskielisen ja loput interpersoonaisen tehtävänannon. Naisten osuudet ovat samansuuntaisia, mutta huomattavasti lähempänä toisiaan. Sitaatin pohjalta kirjoittaneista lähes 85 prosenttia on kirjoittanut yleiskielisestä ja loput interpersoonaisesta tehtävänannosta. Naisten osuudet ovat lähellä miesten jakaumaa. Useita erityyppisiä aineistoja valinneista lähes 70 prosenttia on kirjoittanut yleiskielisestä tehtävänannosta, loput interpersoonaisesta. Korkeakulttuurinenkin tehtävänanto olisi ollut valittavana ruotsinkielisessä kokeessa, mutta siitä kiinnostuneita miehiä ei ole ollut. Naisten osuudet ovat samansuuntaisia, mutta yleiskielisestä tehtävänannosta kirjoittaneita on vielä kymmenen prosenttiyksikköä enemmän.

Niistä miehistä, jotka ovat kirjoittaneet korkeakulttuurisesta moodista, on yli puolet valinnut tehtävänannon, johon ei liity aineistoa. Noin 30 prosentilla aineistona on kuva ja vajaalla viidesosalla runo. Naisista runosta kirjoittaneita on huomattavasti enemmän, jopa kaksi viidestä, ja pohjatekstittömien tehtävänantojen osuus on lähes 20 prosenttiyksikköä pienempi.

Yleiskielisestä tehtävänannosta kirjoittaneista miehistä lähes puolet on jälleen kirjoittanut tehtävänannosta, johon ei liity aineistoa. Sitaatillinen tehtävänanto on kiinnostanut viidesosaa ja asiateksti kuudesosaa. Muiden aineistojen osuus on alle 4,5 prosenttia. Suurin ero naisiin löytyy pohjatekstittömistä tehtävänannoista (naisten osuus 33.15 %) ja jonkin verran eroa on myös proosasta ja esseestä, kolumnista tai pakinasta kirjoittaneiden välillä, joissa naisten osuudet ovat hieman suurempia.

Interpersoonaisesta tehtävänannosta kirjoittaneista miehistä yli 40 prosenttia on valinnut pohjatekstiksi esseen, kolumnin tai pakinan ja 30 prosenttia asiatekstin. Niistä tehtävänannoista, joissa ei ole annettu aineistoa, on kirjoittanut vajaa kuudesosa. Sitaatista kirjoittaneita on seitsemän prosenttia ja kuvasta, uskonnollisesta tai aatteellisesta ja useita erityyppisiä aineistoja sisältävistä tehtävänannoista on kirjoittanut vain harva. Erot naisiin ovat suurimmillaan pohjatekstittömistä tehtävänannoista (naisilla 27.07 %) ja erityisesti esseestä, kolumnista tai pakinasta kirjoittaneilla (16.67 %).

Katsotaan seuraavaksi moodien jakaumia vaaditun intertekstuaalisen varannon suhteen naisilla (Taulukko 32).

Taulukko 32: Moodien jakauma vaaditun intertekstuaalisen varannon mukaan naisilla.

	Korkea-kult.	Yleis-kielinen	Arkinen	Interper-soonainen	Yhteensä
Oma	—	10474	423	1910	12806
	—	81.79	3.30	14.91	100 (%)
	—	47.02	100.00	24.38	39.54

Jatkuu seuraavalla sivulla

Taulukko 32 – Jatkoa edelliseltä sivulta

	Korkea-kult.	Yleis-kielinen	Arkinen	Interper-soonainen	Yhteensä
Kult-tuuri	1211 13.86 65.21	4014 45.93 18.02	— — —	3514 40.21 44.88	8739 100 (%) 26.98
Yhteis-kunta	646 9.22 34.79	4084 58.31 18.33	— — —	2274 32.47 29.05	7004 100 (%) 21.63
Historia	— — —	233 100.00 1.05	— — —	— — —	233 100 (%) 0.72
Kult. & yht.kunta	— — —	3471 96.34 15.58	— — —	132 3.66 1.69	3603 100 (%) 11.12
Yhteensä	1857 5.73 100 (%)	22276 68.78 100 (%)	423 1.31 100 (%)	7829 24.17 100 (%)	32385 100 (%)

Niitä tehtävänantoja, joissa ei tarvita muuta kuin omaa kokemusta, valinneista yli 80 prosenttia on kirjoittanut yleiskielisestä, 15 prosenttia interpersoonaisesta ja vain reilut kolme prosenttia moodiltaan arkisesta tehtävänannosta. Kulttuurista tietoa vaativista tehtävänannoista kirjoittaneista 45 prosenttia on kirjoittanut yleiskielisestä ja vain vähän harvempi interpersoonaisesta tehtävänannosta. Korkeakulttuurinen tehtävänanto on 14 prosentilla.

Yhteiskunnallista tietoa vaativia tehtävänantoja valinneista lähes 60 prosenttia on kirjoittanut yleiskielisestä, kolmasosa interpersoonaisesta ja vajaa kymmenesosa korkeakulttuurisesta tehtävänannosta. Kaikki historiatietoa vaativat tehtävänannot ovat yleiskielisiä. Sekä kulttuurista että yhteiskunnallista tietoa vaativista tehtävistä kirjoittaneista vain neljä prosenttia on valinnut interpersoonaisen ja loput yleiskielisen tehtävänannon.

Korkeakulttuurisista tehtävänannoista kirjoittaneista 65 prosentilta on odotettu kulttuuritietoa ja lopuilta yhteiskunnan tuntemusta. Yleiskielisen tehtävänannon valinneista lähes puolet on saanut kirjoittaa pelkän oman kokemuksen pohjalta, vajaalta viidesosalta on odotettu kulttuurin tuntemusta ja toiselta viidesosalta yhteiskunnallista tietoa. Näitä molempia on joutunut esittelemään 15 prosenttia kokelaista ja vain prosentilta on toivottu historian tuntemusta.

Kaikki moodiltaan arkiset tehtävänannot ovat sellaisia, joissa riittää oma kokemus. Interpersoonaisen tehtävänannon valinneista 45 prosentilta on odotettu kulttuurin tuntemista, 30 prosentilta yhteiskunnallista tietoa ja neljäsosa on selvinnyt pelkällä omalla kokemuksella. Sekä kulttuurista että yhteiskunnallista tietoa on toivottu vajaalta prosentilta.

Seuraavaksi on vuorossa vastaava moodin ja intertekstuaalisen varannon yhteyttä kuvaava taulu miehille (Taulukko 33).

Taulukko 33: Moodien jakauma vaaditun intertekstuaalisen varannon mukaan miehillä.

	Korkea-kult.	Yleis-kielinen	Arkinen	Interper-soonainen	Yhteensä
Oma	—	4892	191	3434	8517
	—	57.44	2.24	40.32	100 (%)
	—	33.03	100.00	43.70	35.91
Kult-tuuri	409	2363	—	2456	5228
	7.82	45.20	—	46.98	100 (%)
	47.95	15.95	—	31.25	22.05
Yhteis-kunta	444	3276	—	1910	5630
	7.87	58.19	—	33.93	100 (%)
	52.05	22.12	—	24.31	23.74
Historia	—	1512	—	—	1512
	—	100.00	—	—	100 (%)
	—	10.21	—	—	6.38
Kult. & yht.kunta	—	2770	—	58	2828
	—	97.95	—	2.05	100 (%)
	—	18.70	—	0.74	11.92
Yhteensä	853	14813	191	7858	23715
	3.60	62.46	0.81	33.14	100 (%)
	100 (%)	100 (%)	100 (%)	100 (%)	

Vain omaa kokemusta aineessaan tarvitsevat ovat valinneet useimmin tehtävänantoja, joissa on yleiskielinen moodi. Interpersoonaisesta tehtävänannosta on kirjoittanut 40 prosenttia ja arkisesta vain kaksi prosenttia. Naisilla vastaavia yleiskielisiä tehtävänantoja on valinnut jopa yli 80 prosenttia ja interpersoonaisia vain 15 prosenttia. Niistä, joilta odotetaan kulttuuritietoa, on korkeakulttuurisen tehtävänannon valinnut vajaat kahdeksan prosenttia. Loput ovat kirjoittaneet suunnilleen tasan yleiskielisestä ja interpersoonaisesta tehtävänannosta. Naisten osuudet ovat samantapaisia.

Yhteiskunnallista tietoa aineessa tarvitsevista lähes 60 prosenttia on kirjoittanut yleiskielisestä, kolmasosa interpersoonaisesta ja vajaat kahdeksan prosenttia moodiltaan korkeakulttuurisesta tehtävänannosta. Eroa naisiin ei juurikaan ole. Sekä kulttuuri- että yhteiskunnallista tietoa vaativista tehtävänannoista kirjoittaneista lähes kaikki ovat valinneet yleiskielisen tehtävänannon, interpersoonaisesta tehtävänannosta on kirjoittanut vain kaksi prosenttia. Edelleen tulokset vastaavat naisten

jakaumaa.

Moodiltaan korkeakulttuurisen tehtävänannon valinneista lähes yhtä moni kokelas on kirjoittanut kulttuuritietoa ja yhteiskunnan tuntemusta vaativista tehtävänannoista. Naisilla sen sijaan kulttuuritietoa vaativista tehtävänannoista on kirjoittanut 65 prosenttia. Interpersoonaisista tehtävänannoista kirjoittaneista miehistä yli 40 prosenttia on valinnut tehtävänannon, jossa ei vaadita muuta kuin omaa kokemusta. Vajaalta kolmasosalta on odotettu kulttuuritietoa ja noin neljäsosalta yhteiskunnan tuntemista. Sekä kulttuurista että yhteiskunnallista tietoa on vaadittu alle prosentilta näistä kokelaista. Naisilla vain omaa kokemusta vaatineista tehtävänannoista on kirjoittanut ainoastaan neljäsosa mutta kulttuuritietoa vaativista jopa 45 prosenttia.

Yleiskielisestä tehtävänannosta kirjoittaneista kolmasosa on saattanut suoriutua kokeestaan pelkällä omalla kokemuksella. Reilu viidesosa on tarvinnut aineessaan yhteiskunnan tuntemusta ja vajaa viidennes sekä kulttuurista että yhteiskunnallista tietoa. Pelkällä kulttuurin tuntemuksella on saattanut selvitä 15 prosenttia kyseisistä kokelaista ja historiatietoa on odotettu kymmenesosalta. Eniten eroa naisiin on vain omaan kokemusta (naisilla lähes puolet) ja historiatietoa vaativissa tehtävänannoissa (naisilla vain prosentti).

Tarkastellaan seuraavaksi moodien jakaumaa vaaditun intertekstuaalisen varannon mukaan suomenkielisen kokeen tehneillä (Taulukko 34).

Taulukko 34: Moodien jakauma vaaditun intertekstuaalisen varannon mukaan suomenkielisessä kokeessa.

	Korkea-kult.	Yleis-kielinen	Arkinen	Interper-soonainen	Yhteensä
Oma	—	14035	432	4684	19151
	—	73.29	2.26	24.46	100 (%)
	—	40.43	100.00	32.58	36.77
Kult-tuuri	1585	6072	—	5838	13495
	11.75	44.99	—	43.26	100 (%)
	61.60	17.49	—	40.61	25.90
Yhteis-kunta	988	7168	—	3854	12010
	8.23	59.68	—	32.09	100 (%)
	38.40	20.65	—	26.81	23.05

Jatkuu seuraavalla sivulla

Taulukko 34 – Jatkoa edelliseltä sivulta

	Korkea-kult.	Yleis-kielinen	Arkinen	Interper-soonainen	Yhteensä
Historia	—	1633	—	—	1633
	—	100.00	—	—	100 (%)
	—	4.70	—	—	3.13
Kult. & yht.kunta	—	5806	—	—	5806
	—	100.00	—	—	100 (%)
	—	16.73	—	—	11.15
Yhteensä	2573	34714	432	14376	52095
	4.94	66.64	0.83	27.60	100 (%)
	100 (%)	100 (%)	100 (%)	100 (%)	

Ne, jotka ovat kirjoittaneet tehtävänannoista, joissa ei tarvita muuta kuin omaa kokemusta, ovat valinneet selvästi useimmin yleiskielisen tehtävänannon. Interpersoonaisesta tehtävänannosta on kirjoittanut joka neljäs ja arkisesta kaksi prosenttia. Kulttuuritietoa vaativista tehtävänannoista kirjoittaneista 45 prosenttia on valinnut yleiskielisen, vähän harvempi interpersoonaisen ja noin joka kymmenes moodiltaan korkeakulttuurisen tehtävänannon.

Niistä, joiden aineilta vaaditaan yhteiskunnallista tietämystä, on kolme viidestä kirjoittanut yleiskielisestä, kolmasosa interpersoonaisesta ja alle kymmenesosa korkeakulttuurisesta tehtävänannosta. Historian tuntemusta ja sekä kulttuuri- että yhteiskunnallista tietoa vaativat tehtävänannot ovat kaikki yleiskielisiä suomenkielisessä kokeessa.

Moodiltaan korkeakulttuurisesta tehtävänannosta kirjoittaneista yli 60 prosenttia on valinnut kulttuuritietoa ja loput yhteiskunnan tuntemusta vaativan tehtävänannon. Interpersoonaisen moodin sisältävistä tehtävänannoista kirjoittaneista 40 prosenttia on valinnut kulttuurin tuntemusta ja reilu neljäsosa yhteiskunnallista tietoa vaativan aiheen. Loput ovat saattaneet pärjätä pelkän oman kokemuksen varassa.

Moodiltaan yleiskielisen tehtävänannon valinneista 40 prosenttia on kirjoittanut vain omaa kokemusta vaativasta tehtävänannosta. Viidesosalta on odotettu yhteiskunnan, vähän harvemmalta kulttuuritietoa ja kuudesosalta näitä molempia. Historiatietoa on tarvinnut alle viisi prosenttia.

Katsotaan vielä viimeiseksi moodien jakaumia vaaditun intertekstuaalisen varannon mukaan ruotsinkielisessä kokeessa (Taulukko 35).

Taulukko 35: Moodien jakauma vaaditun intertekstuaalisen varannon mukaan ruotsinkielisessä kokeessa.

	Korkeakult.	Yleiskielinen	Arkinen	Interpersoonainen	Yhteensä
Oma	—	1329	182	659	2170
	—	61.24	8.39	30.37	100 (%)
	—	56.05	100.00	50.27	54.24
Kulttuuri	35	304	—	132	471
	7.43	64.54	—	28.03	100 (%)
	25.55	12.82	—	10.07	11.77
Yhteiskunta	102	192	—	330	624
	16.35	30.77	—	52.88	100 (%)
	74.45	8.10	—	25.17	15.60
Historia	—	112	—	—	112
	—	100.00	—	—	100 (%)
	—	4.72	—	—	2.80
Kult. & yht.kunta	—	434	—	190	624
	—	69.55	—	30.45	100 (%)
	—	18.30	—	14.49	15.60
Yhteensä	137	2371	182	1311	4001
	3.42	59.26	4.55	32.77	100 (%)
	100 (%)	100 (%)	100 (%)	100 (%)	

Vain omaa kokemusta kirjoituksessaan tarvitsevista ruotsinkielisistä 60 prosenttia on valinnut yleiskielisen, 30 prosenttia interpersoonaisen ja vajaa kymmenesosa moodiltaan arkisen tehtävänannon. Suomenkielisillä sen sijaan yleiskielisistä tehtävänannoista on kirjoittanut yli 70 prosenttia ja muista kahdesta hieman pienempi osuus. Niistä ruotsinkielisistä, joiden teksteissä odotetaan näkyvän kulttuuritieto, on lähes kaksi kolmasosaa kirjoittanut yleiskielisestä, vajaat 30 prosenttia interpersoonaisesta ja seitsemisen prosenttia korkeakulttuurisesta tekstistä. Suomenkielisillä yleiskielisistä ja interpersoonaisista tehtävänannosta kirjoittaneita on lähes yhtä paljon.

Yhteiskunnallista tietoa aineissaan tarvitsevista puolet on valinnut interpersoonaisen, 30 prosenttia yleiskielisen ja kuudesosa korkeakulttuurisen tehtävänannon. Suomenkielisillä korkeakulttuurisesta tehtävästä kirjoittaneiden osuus on suunnilleen puolet pienempi ja yleiskielisestä tehtävänannosta kirjoittaneiden osuus lähes puolet suurempi. Interpersoonaisesta tehtävänannosta on kirjoittanut vain kolmasosa. Historiantuntemusta vaativat tehtävänannot ovat kaikki moodiltaan yleiskielisiä. Sekä kulttuuri- että yhteiskunnallista tietoa vaativista tehtävänannoista kirjoittaneista ruotsinkielisistä 70 prosenttia on valinnut yleiskielisen ja loput interpersoonaisen tehtävänannon. Suomenkielisessä kokeessa kaikki kulttuuri- ja yhteiskunnallista tietoa

vaativat tehtävänannot ovat yleiskielisiä.

Moodiltaan korkeakulttuurisesta tehtävänannosta kirjoittaneista ruotsinkielisistä neljäsosa on valinnut kulttuuritietoa ja loput yhteiskunnan tuntemusta vaativan tehtävänannon. Suomenkielisillä osuudet ovat päinvastaisia: korkeakulttuurisesta tehtävänannosta on kirjoittanut lähes kaksi kolmesta. Yleiskielisen tehtävänannon valinneista reilulta puolikkaalta on vaadittu vain omaa kokemusta ja vajaalta viidesosalta sekä kulttuurin että yhteiskunnan tuntemusta. Reilulta kymmenesosalta on odotettu kulttuuritietoa ja vajaalta kymmenykseltä yhteiskunnan tuntemusta. Viitisen prosenttia on päässyt esittelemään aineessaan historiatietämystään. Eniten eroa suomenkielisiin löytyy vain omaa kokemusta vaativista (osuus 40.43 %) ja yhteiskunnallista tietoa odottavista tehtävänannoista (20.65 %).

Kaikki moodiltaan arkiset tehtävänannot ovat sellaisia, joissa ei tarvita muuta kuin omaa kokemusta. Moodiltaan interpersoonaisesta tehtävänannosta kirjoittaneista puolet on kirjoittanut oman kokemuksen pohjalta ja neljäsosalta on toivottu yhteiskunnallista tietoa. Sekä kulttuurista että yhteiskunnallista tietämystä on odotettu noin 15 prosentilta, kymmenesosalle on riittänyt kulttuurin tuntemus. Eroa suomenkielisiin löytyy lähinnä omaa kokemusta vaativista tehtävänannoista kirjoittaneisiin, joita on vain kolmasosa. Toinen ero löytyy kulttuuri- ja yhteiskunnallista tietoa vaativista tehtävänannoista, joita ei suomenkielisessä kokeessa ole ollut tarjolla.

8 Diskussio

Lähtökohtana tälle työlle on toiminut Ylioppilastutkintolautakunnalta saatu laaja luokiteltu rekisteriaineisto. Aineiston riippuvuusrakenteita on selvitetty sovittamalla loglineaarisia tai kvasiloglineaarisia malleja ja vertaamalla näitä Schwarzin informaatiokriteerillä. Parhaaksi valikoituneen mallin perusteella päästiin alkuperäistä aineistoa pelkistämään.

Laajojen moniulotteisten taulujen tarkastelu olisi alun perin ollut käytännössä mahdotonta tai ainakin erittäin suuritöistä. Pelkistäminen eli aineiston esittäminen alkuperäistä yksinkertaisempien informatiivisten frekvenssi- ja prosenttijakaumien avulla helpottaa tulkintaa huomattavasti, sillä laajojen taulujen sijaan voidaan tulkita oleellisia alempiulotteisia tauluja. Nämä ovat yleensä korkeintaan kolmiulotteisia eli huomattavasti helpommin tarkasteltavissa. Koska kaikki tarkasteluista pois jätettävä on epäinformatiivista, ei vääriä tulkintoja pääse syntymään.

Tehtävänantomuuttujalla on tässä aineistossa hyvin runsaasti luokkia. Tämä hankaloittaa tulkintaa, joten luokkia yhdisteltiin sekä tilastollisin menetelmin että sovellusalan asiantuntijoiden toimesta. Tällä lailla saatiin myös paljon kiinnostavaa tietoa erityyppisistä tehtävänannoista ja niiden yhteyksistä toisiinsa.

Tässä työssä pelkistäminen tehtiin tutkimustulosten kohdalla ainoastaan komponenteittain. Tästä voitaisiin kuitenkin vielä jatkaa ja tutkia, olisiko muuttujien luokkia mahdollista yhdistää. Pelkistämisen suuruus voitaisiin myös haluttaessa mitata vapausasteiden määrän pienentymisenä.

Menetelmän suurimpia hyötyjä on tulkinnan helppous. Tuloksia pystyy lukemaan myös ilman merkittävää tilastotieteen osaamista, kun tarkasteluissa palataan takaisin havaintoavaruuteen sen sijaan, että lukijalle esitettäisiin vaikeasti tulkittava loglineaarinen malli. Yksinkertaisuudestaan huolimatta menetelmällä on vahva teoreettinen pohja.

Tämän aineiston kohdalla menetelmä oli tehokas ja tuotti ainoastaan kahden ja kolmen muuttujan yhdysvaikutuksia, joista jälkimmäisetkin pystyi esittämään kahtena kaksiulotteisena tauluna. Jos aineisto on rakenteeltaan monimutkainen, se ei kuitenkaan pelkisty lainkaan tai juuri lainkaan, eikä tästä menetelmästä ole hyötyä. Tällaisessa tapauksessa on kuitenkin kyse aineiston luonteesta eikä menetelmän toimimattomuudesta.

Haasteena laajojen aineistojen käsittelyssä on itse toteutus. Moniulotteisten taulujen pelkistäminen vaatii vähintään kymmenien, helposti satojen mallien sovittamista, mikä vaatii paljon laskentatehoa ja aikaa – erityisesti, jos muuttujilla on paljon luokkia.

Toistaiseksi menetelmää on käytetty ainoastaan eksploratiivisesti, mutta jatkossa voisi olla tarpeen soveltaa sitä myös konfirmatoriseen tutkimukseen. Lisäksi käytännön työtä auttaisi helppokäyttöinen algoritmi.

Kiitosmaininnat

Haluaisin kiittää Kielitaidon kirjo -tutkimushankkeen jäseniä, erityisesti hankkeen johtajaa kielikasvatuksen professori Anneli Kauppista Jyväskylän yliopistosta sekä ruotsinkielisestä puolesta vastaavaa professori Hanna Lehti-Eklundia Helsingin yliopiston pohjoismaisten kielten ja pohjoismaisen kirjallisuuden laitoksesta. Olen kiitollinen mielenkiintoisesta tutkimusaiheesta ja aineiston hankinnasta, käymistämme avartavista keskusteluista sekä tietenkin tutkimusapulaisen pestistä, jolla on ollut hyvin kannustava vaikutus gradun etenemisessä.

Luonnollisesti suuri kiitos kuuluu myös ohjaajalleni professori Antti Penttiselle lukuisista kannustavista ja hyödyllisistä keskusteluistamme sekä avusta aiheen rajaamisessa ja käsittelyssä. Kiitokset lisäksi tutkija Kari Nissiselle avusta SAS-ohjelman ongelmien kanssa.

Lopuksi haluaisin tietenkin kiittää vielä lähipiiriäni kannustavista keskusteluista ja tuesta tutkielman teon aikana.

Viitteet

[1] Agresti A. (1990). *Categorical Data Analysis.* New York: Wiley.

[2] Akaike H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**: 716–723.

[3] Bishop Y. M. M. (1971). Effects of collapsing multidimensional contingency tables. *Biometrics* **27**: 545–562.

[4] Bishop Y. M. M., Fienberg S. E. ja Holland P. W. (1975). *Discrete Multivariate Analysis.* Cambridge, MA: MIT Press.

[5] Burnham K. P. ja Anderson D. R. (2004). Multimodel Inference. Understanding AIC and BIC in Model Selection. *Sociological Methods & Research* **33**: 261–304.

[6] Fienberg S. E. (1970) Quasi-independence and maximum likelihood estimation in incomplete contingency tables. *Journal of the American Statistical Association* **65**: 1610–1616.

[7] Kullback S. ja Leibler R. A. (1951) On Information and Sufficiency *The Annals of Mathematical Statistics* **22**: 76–86.

[8] Rissanen J. (1983) A Universal Prior for Integers and Estimation by Minumum Description Length. *The Annals of Statistics* **11**: 416–431.

[9] Savage, I. R. (1973) Incomplete contingency tables: condition for the existence of unique MLE. *Mathematics and Statistics. Essays in Honor of Harold Bergström.* 87–99 Göteborg: Chalmers Institute of Technology.

[10] Schwarz G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics* **6**: 461–464.

Liitteet

A Äidinkielen kokeiden tehtävänantojen koodaus

Aihepiiri

1	Aat	aatteellinen
2	Med	viihde, media
3	U	urheilu
4	Tal	talous
5	Yht	yhteiskunta
6	Kir	kirjallisuus
7	Kult	kulttuuri, taide
8	Lu	luonto
9	Av	avoin

Otsikko

1	V	valmis otsikko
2	E	ei valmista otsikkoa

Teema

1	A	avoin teema
2	V	väljästi määritelty teema
3	M	määritelty teema

Genre

1	M	tekstilaji määritelty
2	A	tekstilajia ei määritelty
3	Joh	ohjeistus viittaa tekstityyppiin (esim. erittele, pohdi, analysoi)

Pohjateksti

1	Ei	ei pohjatekstiä
2	K	kuva
3	As	asiateksti
4	Es	essee, kolumni, pakina
5	Fp	fiktio: proosa
6	Fr	fiktio: runo
7	Usk	uskonnollinen, aatteellinen
8	Sit	lyhyt sitaatti tai referointi tehtävänannon yhteydessä
9	Us	useita erityyppisiä aineistoja

Inter

1	Oma	edellyttää vain omaa kokemusta
2	Kult	edellyttää (vain) kulttuuritietoa
3	Yht	edellyttää (vain) yhteiskunnallista tietoa
4	His	edellyttää historiatietoa
5	YK	edellyttää sekä yhteiskunnallista että kulttuuritietoa

Moodi

1	Kork	korkeakulttuurinen
2	Kult	kulttuuripuhetta
3	Yl	yleiskielinen, tunnusmerkitön
4	Ark	arkinen, puheenomainen
5	Ip	interpersoonainen (puhuttelu, 1. tai 2. persoonan muoto, viittaus tai vetoaminen kirjoittajan kokemukseen)

B Tehtävänantojen luokitukset

Taulukko 36: Kevään 2006 suomenkielisen kokeen tehtävänantojen luokitukset.

	Tehtävänannon muotoilu				Tehtävän viritys		
Tehtävä	Aihe-piiri	Otsik-ko	Teema	Genre	Pohja-teksti	Inter	Moodi
1	Kult	E	A	A	K	Kult	Ip
2	Kir	E	M	A	Fp	Kult	Yl
3	Kir	E	V	Joh	Fr	Kult	Kork
4	Aat	E	M	A	Usk	Kult	Yl
5	Aat	E	M	A	Usk	YK	Yl
6	Yht	E	M	Joh	As	YK	Yl
7	Kult	E	A	M	As	Kult	Ip
8	Kult	E	M	A	As	Kult	Ip
9	Kult	E	M	A	Fp	Kult	Yl
10	Yht	E	M	A	Es	Yht	Ip
11	Med	E	A	Joh	As	Yht	Yl
12	Med	E	M	M	As	Yht	Yl
13	Kult	E	M	A	Us	Kult	Yl
Tehtävä	Aihep.	Ots.	Teema	Genre	Pohj.	Inter	Moodi
21	Med	E	V	Joh	Ei	Yht	Kork
22	Aat	E	A	Joh	Sit	YK	Yl
23	Yht	E	A	A	Sit	His	Yl
24	Yht	V	M	A	Ei	Yht	Yl
25	Kult	E	V	A	Ei	Oma	Ark
26	Yht	E	A	A	Sit	Oma	Yl
27	Kir	V	M	A	Ei	Kult	Ip
28	Lu	E	V	A	Sit	Yht	Yl
29	Kult	V	A	A	Ei	Kult	Yl
30	Lu	E	A	A	Ei	Yht	Ip
31	Med	E	A	A	Ei	Oma	Yl
32	Yht	E	A	A	Sit	His	Yl
33	Yht	E	A	Joh	Sit	Yht	Ip
34	Av	V	A	A	Ei	Oma	Yl

Taulukko 37: Kevään 2006 ruotsinkielisen kokeen tehtävänantojen luokitukset.

	Tehtävänannon muotoilu				Tehtävän viritys		
Tehtävä	Aihe-piiri	Otsik-ko	Teema	Genre	Pohja-teksti	Inter	Moodi
1	Kult	E	A	A	K	Kult	Ip
2	Kir	E	A	Joh	Us	Kult	Kork
3	Aat	E	M	Joh	Es	YK	Yl
4	Yht	E	A	A	Us	Yht	Ip
5	Kult	E	M	M	As	YK	Ip
6	Med	E	A	A	As	Oma	Yl
7	Kult	E	M	A	Us	Kult	Yl
8	Yht	E	M	M	As	Oma	Ip
9	Kult	E	M	A	Us	Kult	Yl
10	Kult	E	M	A	Fp	Kult	Yl
11	Kult	E	M	Joh	Us	Kult	Yl
12	Kult	E	M	Joh	Es	Oma	Ip
13	Kult	E	M	A	Fr	Kult	Kork
Tehtävä	Aihep.	Ots.	Teema	Genre	Pohj.	Inter	Moodi
21	Med	E	V	Joh	Ei	Yht	Kork
22	Yht	E	A	Joh	Sit	YK	Yl
23	Yht	E	A	A	Sit	His	Yl
24	Yht	V	M	A	Ei	Yht	Yl
25	Kult	E	V	A	Ei	Oma	Yl
26	Yht	E	A	A	Sit	Oma	Yl
27	Kir	V	M	A	Ei	Kult	Ip
28	Lu	E	V	A	Sit	Yht	Yl
29	Kult	V	A	A	Ei	Kult	Yl
30	Lu	E	A	A	Ei	Yht	Ip
31	Med	E	A	A	Ei	Oma	Yl
32	Yht	E	A	A	Sit	His	Yl
33	Yht	E	A	Joh	Sit	Yht	Ip
34	Av	V	A	A	Ei	Oma	Yl

Taulukko 38: Kevään 2007 suomenkielisen kokeen tehtävänantojen luokitukset.

	Tehtävänannon muotoilu				Tehtävän viritys		
Tehtävä	Aihe-piiri	Otsik-ko	Teema	Genre	Pohja-teksti	Inter	Moodi
1	Kir	V	V	A	Ei	Kult	Yl
2	Yht	E	V	A	Ei	Oma	Yl
3	Tal	V	M	A	Ei	Yht	Yl
4	Av	V	A	A	Ei	Oma	Yl
5	Yht	E	M	A	Sit	Oma	Yl
6	U	E	M	A	Ei	YK	Yl
7	Kult	E	M	A	Ei	Kult	Yl
8	Aat	E	V	Joh	Usk	Oma	Ip
9	Kir	E	A	Joh	Fp	YK	Yl
10	Kir	E	A	M	Fr	Kult	Kork
11	Yht	E	M	A	As	Yht	Yl
12	Med	E	V	Joh	Es	Oma	Ip
13	Kult	E	A	Joh	Es	Kult	Yl
14	Med	V	V	A	K	Oma	Yl

Taulukko 39: Kevään 2007 ruotsinkielisen kokeen tehtävänantojen luokitukset.

	Tehtävänannon muotoilu				Tehtävän viritys		
Tehtävä	Aihe-piiri	Otsik-ko	Teema	Genre	Pohja-teksti	Inter	Moodi
1	Kir	V	V	A	Ei	Kult	Yl
2	Yht	E	V	A	Ei	Oma	Ark
3	Tal	V	M	A	Ei	Yht	Yl
4	Av	V	A	A	Ei	Oma	Yl
5	Yht	E	M	A	Sit	Oma	Yl
6	U	E	M	A	Ei	YK	Yl
7	Kult	E	M	A	Ei	Kult	Yl
8	Aat	E	V	Joh	Usk	Oma	Ip
9	Yht	E	M	A	Us	YK	Ip
10	Kir	E	M	A	Fr	YK	Yl
11	Med	E	M	Joh	Es	Oma	Ip
12	Lu	E	M	M	Es	Oma	Yl
13	Yht	E	M	M	Es	Oma	Yl
14	Med	V	V	A	K	Oma	Yl