Ristitaulukot ja visualisointi

Eetu Tammi

LuK-tutkielma
Tammikuu 2024

MATEMATIIKAN JA TILASTOTIETEEN LAITOS

TURUN YLIOPISTO
Matematiikan ja tilastotieteen laitos

EETU TAMMI: Ristitaulukot ja visualisointi
LuK-tutkielma, 21 s.
Tilastotiede
Tammikuu 2024

Tässä kandidaatintutkielmassa käsitellään ristitaulukoita eli kontingenssitaulukoita. Tutkielma käsittelee ensin ristitaulukoiden perusteita ja millaisiin tilanteisiin ristitaulukoita voidaan soveltaa. Tämän jälkeen tutustutaan erityisesti 2×2 -taulukkoon, jonka avulla käydään läpi erilaisia ristitaulukoiden tunnuslukuja sekä riippumattomuuden testaus eli χ^2-testi.

Kandidaatintutkielman jälkimmäisessä osassa keskitytään tutkimaan erilaisia visualisointimenetelmiä ristitaulukoille sekä perehdytään korrespondenssianalyysiin.

Avainsanat: ristitaulukko, riippumattomuus, visualisointi, korrespondenssianalyysi

Sisällys

1 Johdanto

Ristitaulukot ovat tehokas työkalu tilastollisten tietojen analysoinnissa ja vertailussa eri ryhmien välillä. Ne tarjoavat mahdollisuuden havainnollistaa eri luokkien välisiä riippuvuuksia ja eroja. Tässä kandidaatintutkielmassa tarkastelemme ristitaulukoiden käyttöä, niiden visualisointia ja merkitystä aineistojen analysoinnissa. Lopuksi käydään läpi lyhyesti korrespondenssianalyysia. Tämän työn tärkeimpiä lähteitä ovat Tilastollisen päättelyn peruskurssin luentomoniste [2, luku 6] ristitaulukoiden teorian osalta sekä teos *Handbook of Data Visualization* [5, luku III.12] ristitaulukoiden visualisoinnissa.

Ristitaulukot tunnetaan myös nimellä kontingenssitaulukot. Ristitaulukot kuvaavat monimuuttujien frekvenssijakaumaa, mutta tässä kandidaatintutkielmassa keskitytään kaksiulotteisiin tapauksiin. Ristitaulukoita visualisoitaessa tarkoituksena on löytää erilaisia yhteyksiä muuttujien välillä visuaalisin keinoin. χ^2-testi on tärkeässä osassa, kun tutkitaan muuttujien riippumattomuutta toisistaan. Korrespondenssianalyysissä pyritään ymmärtämään, kuinka suuri osa aineiston vaihtelusta selittyy tietyssä ulottuvuudessa eli dimensiossa.

Ristitaulukoita käytetään monilla eri aloilla, mutta tunnetuin käyttökohde ristitaulukoille on erilaiset lääketieteen sovellukset. Esimerkiksi koronapikatestin positiiviset ja negatiiviset tulokset sekä oikea tieto koronavirusinfektiosta voidaan mallintaa 2×2 -ristitaulukkona. Tässä tutkielmassa pyritään tarkastelemaan ristitaulukoiden käyttöä eri tilanteissa, niiden merkitystä tilastollisessa analyysissa sekä erilaisia menetelmiä ja työkaluja, joita voidaan hyödyntää ristitaulukoiden visualisoinnissa. Selvitetään myös, kuinka visualisointien avulla voidaan tehdä päätöksiä ja löytää yhteyksiä aineiston muuttujien välillä.

2 Ristitaulukot

Tässä luvussa käydään läpi, mikä on ristitaulukko. Luvussa on seurattu lähdettä [1] sekä Tilastollisen päättelyn peruskurssin luentomonistetta [2, luku 6].

Tässä kandidaatintutkielmassa keskitytään kaksiulotteisiin ristitaulukoihin, jotka kuvaavat kaksimuuttujaisen aineiston frekvenssijakaumaa. Ristitaulukoiden avulla voidaan myös kuvata useampimuuttujaisia aineistoja. Tämä edellyttää useampia kaksiulotteisia ristitaulukoita. Ristitaulukot kuvaavat hyvin kahden tai useamman muuttujan välisiä suhteita. Muuttujat ristitaulukoissa voivat olla sekä numeerisia että luokallisia muuttujia, mutta niiden tulee olla diskreettejä tai luokallisia eli muuttujien luokan tulee olla toisensa poissulkevia eli havainnon tulee kuulua yksiselitteisesti tiettyyn luokkaan. Havaintoja kerättäessä ristitaulukkoon käytetään yleisesti yksinkertaista satunnaisotantaa.

Ristitaulukot ovat erittäin yleinen työkalu tilastotieteessä. Ristitaulukoita käytetään yleisesti osana eri alojen tutkimuksia. Ristitaulukoita käytettäessä erityisesti kiinnostuksen kohteena ovat muuttujien väliset suhteet kuten riippuvuus ja assosiaatio.

2.1 Kaksiulotteisen ristitaulukon yleinen muoto

				Y			
		y_1	$\cdots$	y_j	$\cdots$	y_s	
	x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1s}	$n_{1\cdot}$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
X	x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{is}	$n_{i\cdot}$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_r	n_{r1}	$\cdots$	n_{rj}	$\cdots$	n_{rs}	$n_{r\cdot}$
		$n_{\cdot 1}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot s}$	n

Taulukko 1: Ristitaulukko

Ristitaulukossa n_{ij} on niiden havaintojen määrä, jossa muuttuja X saa arvon x_i ja muuttuja Y saa arvon y_j. Luvut $n_{i\cdot}$ ja $n_{\cdot j}$ kuvaavat muuttujien X ja Y reunafrekvenssijakaumia eli ovat taulukon rivi- ja sarakesummia. Luku $n_{i\cdot}$ kertoo montako kertaa x_i esiintyy aineistossa, ja $n_{\cdot j}$ kertoo montako kertaa y_j esiintyy aineistossa. Kaikkien solujen n_{ij} summaa kuvaa n, joka on myös havaintojen kokonaismäärä.

2.2 Yksinkertainen esimerkki 2×2 -ristitaulukosta

Käytetään esimerkkinä aineistoa, joka kuvaa luottokorttimaksun laiminlyömistä. Aineisto on löydetty kaggle.com-sivulta [3]. Aineistossa on 25 eri muuttujaa. Tässä esimerkissä rajataan aineistosta kiinnostuksen kohteeksi sukupuoli sekä seuraavan kuukauden luottokorttimaksun maksamatta jättäminen. Aineisto on kerätty Taiwanissa vuonna 2005.

	Y (Maksun laiminlyöminen)		
X(Sukupuoli)	Kyllä	Ei	Yhteensä
Nainen	137	506	643
Mies	86	271	357
Yhteensä	223	777	1000

Taulukko 2: Maksun laiminlyöminen sukupuolen mukaan

Aineistossa on 30000 havaintoa, mutta otetaan tutkittavaksi 1000 havainnon kokoinen otos. Taulukkoa (taulukko 2) visuaalisesti tutkittaessa ei voida todeta sarakkeiden frekvenssien vaihtelevan paljoa rivien suhteen eli muuttujien välillä ei visuaalisen tutkimisen perusteella voida todeta olevan riippuvuutta. Muuttujien välistä riippuvuutta pystytään tutkimaan erilaisilla tunnusluvuilla, joista lisää luvussa 3, sekä χ^2-testillä, josta lisää luvussa 5.

3 Tunnuslukuja

3.1 Vetosuhde

Vetosuhde (odds ratio) on yksinkertaisin tunnusluku ristitaulukoille. Se kertoo, kuinka todennäköisesti kaksi tapahtumaa A ja B esiintyvät yhdessä (taulukko 3). Vetosuhde toimii vain 2×2 -taulukoille. Jos kyseessä olisi suurempi taulukko, voitaisiin se jakaa 2×2 -osataulukoiksi ja laskea vetosuhteet niille. Vetosuhde saa positiivisia arvoja. Vetosuhteen ollessa 1 (tai lähellä arvoa 1) ovat tapahtumat A ja B riippumattomia. Tapahtumat ovat positiivisesti korreloituneita, jos vetosuhde on suurempaa kuin 1. Tapahtumat ovat negatiivisesti korreloituneita, jos vetosuhde on alle 1.

	A	
B	A tapahtuu	A ei tapahdu
B tapahtuu	p_{11}	p_{01}
B ei tapahdu	p_{10}	p_{00}

Taulukko 3: Ristitaulukko tapahtumille A ja B

jossa p_{ij}:t ovat suhteellisia osuuksia, jotka summautuvat ykköseksi:

$$p_{ij} = \frac{n_{ij}}{n} \tag{1}$$

Vetosuhteen laskeminen perustuu vetokertoimiin (odds), jonka kaava tietylle tapahtumalle on seuraava:

$$Odds = \frac{p}{1-p} \tag{2}$$

jossa p on kyseisen tapahtuman todennäköisyys.

Vetosuhde tapahtumille A ja B voidaan muodostaa seuraavasti:

$$\mathbf{OR} = \frac{P(A \cap B) \cdot P(\neg A \cap \neg B)}{P(\neg A \cap B) \cdot P(A \cap \neg B)} = \frac{p_{11}p_{00}}{p_{01}p_{10}} \tag{3}$$

jossa merkintä A kuvaa tapahtuman A tapahtumista ja merkintä $\neg A$ kuvaa

tapahtuman A komplementtitapahtumaa.

Lasketaan seuraavaksi luvun 2.2 esimerkin vetosuhde. Muunnetaan ensin ristitaulukko niin, että ristitaulukosta nähdään tapahtumien suhteelliset osuudet p_{ij}. Lasketaan tämän jälkeen esimerkin vetosuhde (4)

	Y (Maksun laiminlyöminen)	
X(Sukupuoli)	Kyllä	Ei
Nainen	0.14	0.51
Mies	0.09	0.27

Taulukko 4: Todennäköisyys jättää luottokorttimaksu maksamatta

$$\mathbf{OR} = \frac{0.14 \cdot 0.27}{0.51 \cdot 0.09} \approx 0.82 \tag{4}$$

Vetosuhteeksi saatu arvo on melko lähellä arvoa 1, joten muuttujien välillä ei ole voimakasta yhteyttä. Vetosuhde on kuitenkin alle 1 eli muuttujien välinen yhteys on negatiivinen eli miehen on hieman todennäköisempää jättää luottokorttimaksu maksamatta kuin naisen. [4]

3.2 ϕ-kerroin

Tässä alaluvussa on seurattu lähdettä [1]. Yksinkertainen tunnusluku 2×2-ristitaulukoille kuvaamaan satunnaismuuttujien välistä assosiaatiota on ϕ-kerroin (ϕ-coefficient):

$$\phi = \pm\sqrt{\frac{\chi^2}{n}} \tag{5}$$

Etumerkki ϕ-kertoimelle saadaan laskemalla 2×2 -taulukon päädiagonaalin alkioiden tulon ja ei-diagonaalisten alkioiden tulon erotus. Jos erotus on positiivinen, ϕ-kerroin on positiivinen. Jos erotus on negatiivinen, ϕ-kerroin on negatiivinen. Sen arvot vaihtelevat 0:sta (muuttujien välillä ei ole assosiaatiota) 1:een tai -1:een (täydellinen assosiaatio tai täydellinen käänteinen assosiaatio).

Lasketaan seuraavaksi ϕ-kerroin esimerkkiaineistolle:

$$\phi = -\sqrt{\frac{1.03}{1000}} \approx -0.03$$

χ^2-testisuure lasketaan luvussa 5. Saatu ϕ-kerroin voidaan tulkita siten, että luottokorttimaksun maksamatta jättämisellä ja sukupuolella on heikko negatiivinen assosiaatio. Eli tässä tilanteessa miehet jättävät maksun maksamatta hieman todennäköisemmin.

3.3 Cramerin V

Tässä alaluvussa on seurattu lähdettä [1]. Cramerin V (kaava (6)) on assosiaation mittari, joka saa arvoja 0:sta (muuttujien välillä ei ole assosiaatiota) 1:een (muuttujien välillä on täydellinen assosiaatio). Se määritellään kaavalla

$$V = \sqrt{\frac{\chi^2}{n(k-1)}} \tag{6}$$

jossa k on rivien tai sarakkeiden määrä, riippuen siitä, kumpi on pienempi. Cramerin V sopii erityisesti suuremmille kuin 2×2 -ristitaulukoille, sillä Cramerin V ja ϕ-kerroin ovat etumerkkiä lukuun ottamatta samat 2×2 -ristitaulukoille.

4 Visualisointi

Tässä luvussa tutustutaan ristitaulukoiden visualisointiin erilaisten kuvaajien avulla. Tässä luvussa on seurattu *Handbook of Data Visualization* -teosta [5, luku III.12]. Aineiston visualisoinnissa on kyse numeerisen aineiston esittämisestä graafisesti kuvaajina, kaavioina tai muina graafisina elementteinä, jotta aineistoa olisi helpompi ymmärtää ja tulkita. Tavoitteena aineistosta on tehdä helpommin hahmotettava sekä havainnollinen, jotta päätöksenteko ja tiedon analysointi olisi tehokkaampaa.

4.1 Aineiston esittely

Esimerkkiaineistona visualisoinnissa on käytetty samaa aineistoa ja samaa otosta kuin alaluvussa 2.2, mutta nyt kiinnostuksen kohteeksi valitaan henkilön ikäryhmä sekä annetun luoton määrä. Luokitellaan muuttujat ikä ja luoton määrä neljään eri luokkaan, jotta ristitaulukon muodostaminen on mahdollista.

	Y (Ikäryhmä)			
X(Luoton määrä (NT$))	21–30	31–40	41–50	yli 50
alle 100 000	198	79	83	24
100 000–199 999	88	98	44	15
200 000–299 999	45	98	37	10
yli 300 000	41	78	46	16

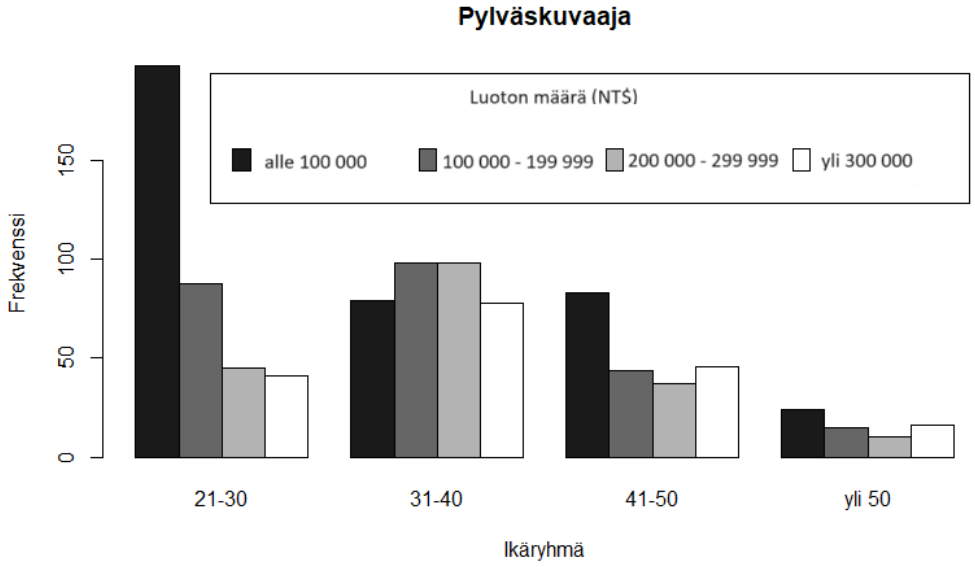
Taulukko 5: Luoton määrä ikäryhmän mukaan

Otoksessa on 1000 havaintoa. Aineisto on kerätty Taiwanissa, joten luoton määrä on ilmoitettu Taiwanin dollareissa (NT$). 100 tuhatta Taiwanin dollaria vastaa noin 3000 euroa.

4.2 Pylväskuvaaja

Pylväskuvaaja on yksi yleisimmistä ja helpoimmin tulkittavista kuvaajista aineiston visualisoinnissa. Pylväskuvaaja on hyvä työkalu eri luokkia vertailtaessa, sillä pylväskuvaajat ovat varsin helppolukuisia ja yksinkertaisia. Luodaan esimerkkiaineistolle pylväskuvaaja seuraavalla R-koodilla:

```
x <- table(Luottoryhma,Ikaryhma)
barplot(x, beside = TRUE, col=gray(seq(0.1,1,length=4)),
main="Pylvaskuvaaja", xlab = "Ikaryhma",
ylab="Frekvenssi")
```



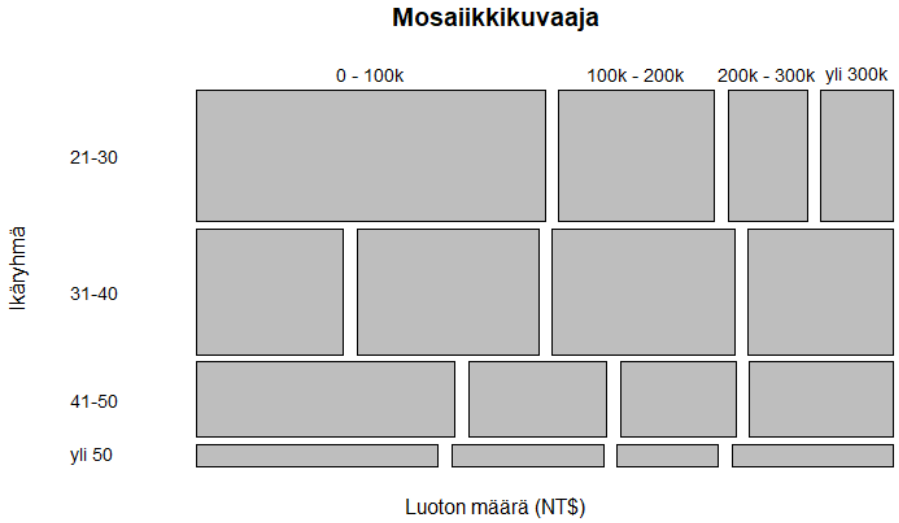
Kuva 1: Pylväskuvaaja

Esimerkkiaineistolle luodussa pylväskuvaajassa (kuva 1) huomataan, että alle 100 000 NT$:n luotot ovat yleisempiä 21–30-vuotiaiden ikäryhmässä. Muiden ikäryhmien jakauma luoton määrän suhteen on huomattavasti tasaisempi.

4.3 Mosaiikkikuvaaja

Mosaiikkikuvaaja on visuaalinen työkalu, joka muodostuu pinta-alasuhteellisista laatoista, jotka on järjestetty suorakaiteen muotoiseen mosaiikkiin. Laatat saadaan jakamalla suorakulmio rekursiivisesti. Mosaiikkikuvaajaa muodostettaessa ensimmäinen jako on hallitseva, joten ensimmäiseksi jakomuuttujaksi tulisi valita selittävä muuttuja. Jako tapahtuu aineiston frekvenssien mukaan. Tämän jälkeen jako suoritetaan toisen muuttujan suhteen. Tällöin jo muodostetut laatat jaetaan toisen muuttujan frekvenssien mukaan. Mosaiikkikuvaaja on hyödyllinen esimerkiksi silloin, kun halutaan nähdä, kuinka tietyt ominaisuudet tai ilmiöt jakautuvat eri luokkien kesken. Tutkitaan alaluvun 4.1 aineistoa seuraavaksi mosaiikkikuvaajan avulla. Luodaan esimerkkiaineistolle mosaiikkikuvaaja:

```
mosaicplot(x, sort = 2:1, main = "Mosaiikkikuvaaja",
ylab = "Ikaryhma", xlab = "Luoton_maara(NT)", las = 1)
```



Kuva 2: Mosaiikkikuvaaja

Mosaiikkikuvaajassa (kuva 2) ensimmäiseksi jakomuuttujaksi valitaan ikäryhmä, joka toimii selittävänä muuttujana. Tämän jälkeen jako suoritetaan luoton määrän mukaan. Jos muuttujilla ei olisi yhteyttä toisiinsa, olisi mosaiikkikuvaaja säännöllinen. Kuvaajasta selvästi nähdään, että 21–30-vuotiailla annetun luoton määrä on pienempi kuin muilla ikäryhmillä. Mosaiikkikuvaajasta huomataan myös, että 31–40-vuotiailla on suurempia luoton määriä enemmän kuin muilla ikäryhmillä. Eli muuttujien välillä on yhteyttä.

4.4 Assosiaatiokuvaajat

Assosiaatiokuvaajat ovat visualisointityökaluja, joita käytetään havainnollistamaan ja tutkimaan tilastollisia yhteyksiä ja assosiaatioita kahden tai useamman muuttujan välillä. Assosiaatiokuvaajat auttavat havainnollistamaan, kuinka eri muuttujien arvot liittyvät toisiinsa. Assosiaatiokuvaajia

tulkitaan muodostettujen palkkien avulla. Positiivinen palkki tarkoittaa, että ryhmien välillä on positiivista assosiaatiota. Negatiivinen palkki kertoo taas negatiivisesta assosiaatiosta. Assosiaatiokuvaajat perustuvat yleisesti Pearsonin residuaaleihin:

$$r_{ij} = \frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}} \tag{7}$$

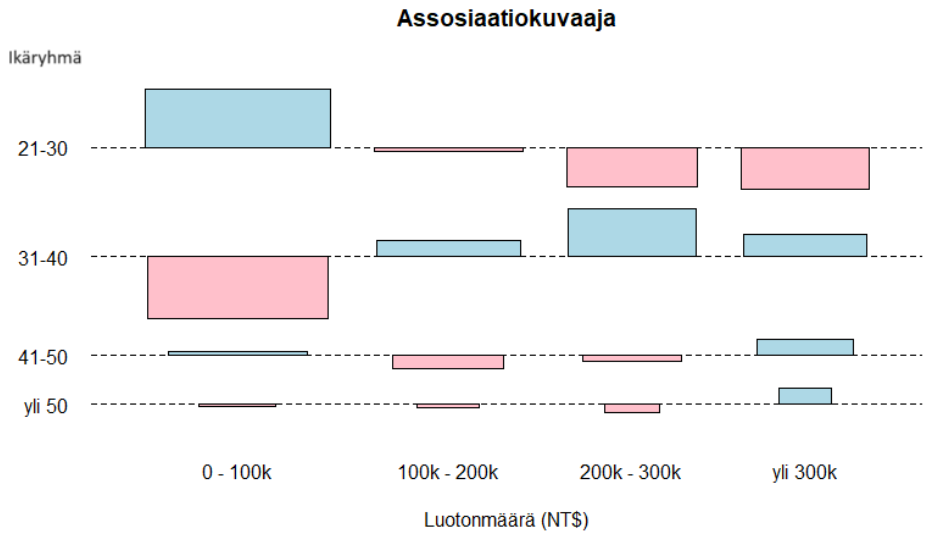
jossa n_{ij} on havaittu frekvenssi ja e_{ij} on odotetun frekvenssin estimaatti:

$$e_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n} \tag{8}$$

Odotetun frekvenssin estimaatin lauseke muodostuu rivi- ja sarakesummien n_i ja n_j tulon ja kaikkien havaintojen osamäärästä.

Assosiaatiokuvaajan palkkien korkeus kuvaa Pearsonin residuaalien arvoa ja leveys havaintojen osuutta aineistosta. Luodaan assosiaatiokuvaaja esimerkkiaineistolle:

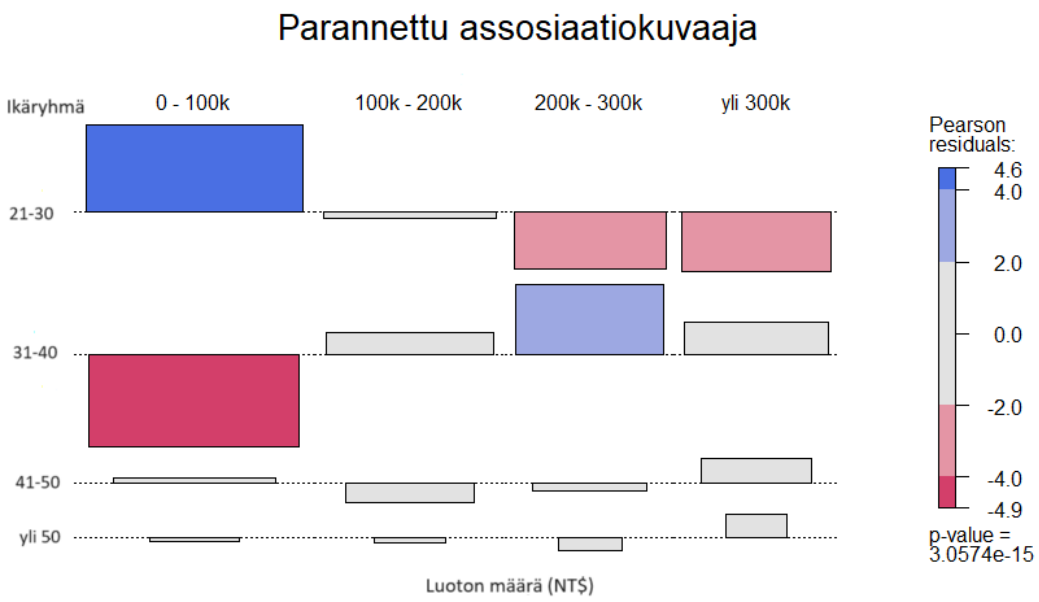
```
assocplot(x, col = c("lightblue", "pink"),
main = "Assosiaatiokuvaaja", xlab = "Luoton_maara(NT)",
ylab = "Ikaryhma", las = 1)
```



Kuva 3: Assosiaatiokuvaaja

Esimerkkiaineistosta tehdystä assosiaatiokuvaajasta (kuva 3) nähdään, että 21–30-vuotiailla on selvästi enemmän alle 100 000 NT$:n luottoja ja vähemmän suuria luottoja. Erityisesti 31–40-vuotiailla on vähemmän alle 100 000 NT$:n luottoja. Luodaan parannettu assosiaatiokuvaaja tilanteesta:

```
z <- table(Ikaryhma, Luottoryhma)
assoc(z, shade = TRUE, legend = TRUE, main = "Parannettu
assosiaatiokuvaaja", xlab = "Luotonmaara(NT)",
ylab = "Ikaryhma")
```



Kuva 4: Parannettu assosiaatiokuvaaja

Parannetussa assosiaatiokuvaajassa (kuva 4) tilastollisesti merkitsevät erot on värikoodattu. 21–30-vuotiailla on erityisen paljon alle 100 000 NT$:n ja vähemmän yli 200 000 NT$:n luottoja. 31–40-vuotiailla on selvästi muita vähemmän alle 100 000 NT$:n luottoja ja enemmän 200 000–300 000 NT$:n luottoja. Parannettu assosiaatiokuvaaja laskee myös aineiston χ^2-testin p-arvon. Tässä tapauksessa p-arvo on erittäin pieni ($p < 0.001$), joten muuttujien välillä on tilastollisesti merkitsevä yhteys. χ^2-testiä käydään tarkemmin läpi luvussa 5.

5 χ^2-testi

Tässä luvussa on seurattu lähdettä [6] sekä Tilastollisen päättelyn luentomonistetta [2, luku 6]. χ^2-testi (kaava (11)) on epäparametrinen tilastollinen testi, joka testaa nollahypoteesia, jossa luokalliset muuttujat X ja Y (taulukko 1) oletetaan riippumattomiksi. χ^2-testi vaatii suurehkon otoskoon, mikä tarkoittaa useampaa kymmentä havaintoa luokkien määrästä riippuen. Suurehko otoskoko sekä havaintojen riippumattomuus tarvitaan, jotta testisuure noudattaisi likimain χ^2-jakaumaa. χ^2-testiä käytetään määrittämään, onko odotettujen ja havaittujen frekvenssien välillä tilastollisesti merkitsevää eroa yhdessä tai useammassa ristitaulukon luokassa.

Olkoon satunnaismuuttujien X ja Y pistetodennäköisyysfunktiot seuraavat:

$$\begin{aligned} P(X = x_i) &= p_i \\ P(Y = y_j) &= q_j \end{aligned} \tag{9}$$

Nämä satunnaismuuttujat ovat riippumattomia, jos ja vain jos

$$P(X = x_i, Y = y_j) = p_{i,j} = P(X = x_i) \cdot P(Y = y_j) = p_i \cdot q_j \tag{10}$$

Nollahypoteesin ja havaintojen eroa voidaan mitata χ^2-testillä, jonka testisuure on

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \tag{11}$$

jossa n_{ij} on havaittu frekvenssi ja e_{ij} on nollahypoteesin mukaisen odotusarvon estimaatti (kaava (8)). Odotusarvon estimaatti on luku, joka estimoi kuinka monta havaintoa solussa odotusarvoisesti olisi, jos muuttujat olisivat riippumattomia toisistaan.

Testisuureen χ^2 suuret arvot todistavat nollahypoteesia vastaan, kun nollahypoteesissa on oletettu muuttujien olevan riippumattomat. χ^2-testisuure on jakautunut likimain χ^2-jakauman mukaan nollahypoteesin pätiessä. Ja-

kauman vapausaste määritetään seuraavasti:

$$df = (r-1)(s-1) \tag{12}$$

jossa r ja s ovat ristitaulukon rivien ja sarakkeiden lukumäärät.

Tämän jälkeen voidaan määrittää p-arvo. Testin likimääräinen p-arvo määritetään saadun testisuureen arvon perusteella

$$p = 1 - F(\chi^2) \tag{13}$$

jossa F on χ^2_{df} -jakauman kertymäfunktio.

P-arvo voidaan määrittää joko tilastollisella ohjelmalla tai taulukosta lukien. Lopuksi p-arvon avulla tehdään päätös nollahypoteesin hyväksymisestä tai hylkäämisestä tietyllä merkitsevyystasolla. Pienet p-arvot tukevat nollahypoteesin hylkäämistä, kun taas suuret p-arvot tukevat nollahypoteesia.

5.1 χ^2-testi esimerkkiaineistolle

Tässä alaluvussa käydään läpi χ^2-testin vaiheet käytännössä. Käytetään esimerkkiaineistoa luvusta 2.2, joka kuvaa luottokorttimaksun maksamatta jättämistä sukupuolen mukaan. Aluksi muotoillaan nolla- ja vastahypoteesit:

H_0 : Sukupuolella ja luottokorttimaksun maksamatta jättämisellä ei ole yhteyttä.
H_v : Sukupuolella ja luottokorttimaksun maksamatta jättämisellä on jonkinlainen yhteys.

Määritetään seuraavaan taulukkoon odotetut frekvenssit e_{ij} (taulukko 6):

	Y (Maksun laiminlyöminen)	
X(Sukupuoli)	Kyllä	Ei
Nainen	$\frac{643 \cdot 223}{1000} \approx 143.39$	$\frac{643 \cdot 777}{1000} \approx 499.61$
Mies	$\frac{357 \cdot 223}{1000} \approx 79.61$	$\frac{357 \cdot 777}{1000} \approx 277.39$

Taulukko 6: Odotetut frekvenssit

Lasketaan seuraavaksi χ^2-testisuure:

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} \\ &= \frac{(137-143.39)^2}{143.39} + \frac{(506-499.61)^2}{499.61} + \frac{(86-79.61)^2}{79.61} + \frac{(271-277.39)^2}{277.39} \\ &\approx 1.03\end{aligned}$$

Määritetään χ^2-jakauman vapausaste (kaava (12)): $df = (2-1)(2-1) = 1$. Tällöin χ^2_1-jakauman kriittinen alue löytyy oikeasta hännästä, kriittinen arvo on 3.841, kun merkitsevyystasoksi on valittu 0.05. Tarkan p-arvon määrittämiseksi käytetään R-ohjelmistoa:

```
1-pchisq(1.03, 1)
> 0.3101587
```

P-arvoksi saadaan $p \approx 0.31$, joka ei todista nollahypoteesia vastaan eli luottokorttimaksun maksamatta jättämisellä miesten ja naisten välillä ei ole tilastollisesti merkitsevää eroa.

χ^2-testi voitaisiin tehdä myös suoraan R-ohjelmistolla seuraavalla koodilla:

```
chisq.test(Sukupuoli, Maksun laiminlyominen)
> X-squared = 1.0263, df = 1, p-value = 0.311
```

2×2 -taulukon testaus voitaisiin toteuttaa myös kahden suhteellisen osuuden testinä. Siinä testisuuren jakauma nollahypoteesin pätiessä on standardinormaalijakauma eli $N(0,1)$-jakauma. Tällöin testisuureen toinen potenssi on sama kuin χ^2-testisuure. Kahden suhteellisen osuuden testistä lisää Tilastol-

lisen päättelyn peruskurssin luentomonisteessa [2, s. 54-55].

6 Rivi- ja sarakeprofiilit sekä inertia ja korrespondenssianalyysi

Tämä luku perustuu monimuuttujamenetelmät-kurssin diasarjaan [7, osio 9]. Rivi- ja sarakeprofiilit sekä inertia luovat taustaa korrespondenssianalyysille, joka esitellään lyhyesti alaluvussa 6.3.

6.1 Rivi- ja sarakeprofiilit

Riviprofiilit ovat tilastollinen käsite, jonka tarkoituksena on vertailla sarakemuuttujien suhteellisia frekvenssijakaumia eri rivien kesken. Riviprofiileja analysoitaessa voidaan käyttää erilaisia graafisia menetelmiä kuten pylväskuvaajia, viivakuvaajia, hajontakuvia ja lämpökarttoja. Profiileja tulkittaessa yleensä keskitytään luokkien välisiin eroihin tai luokkien eroon keskimääräisestä profiilista.

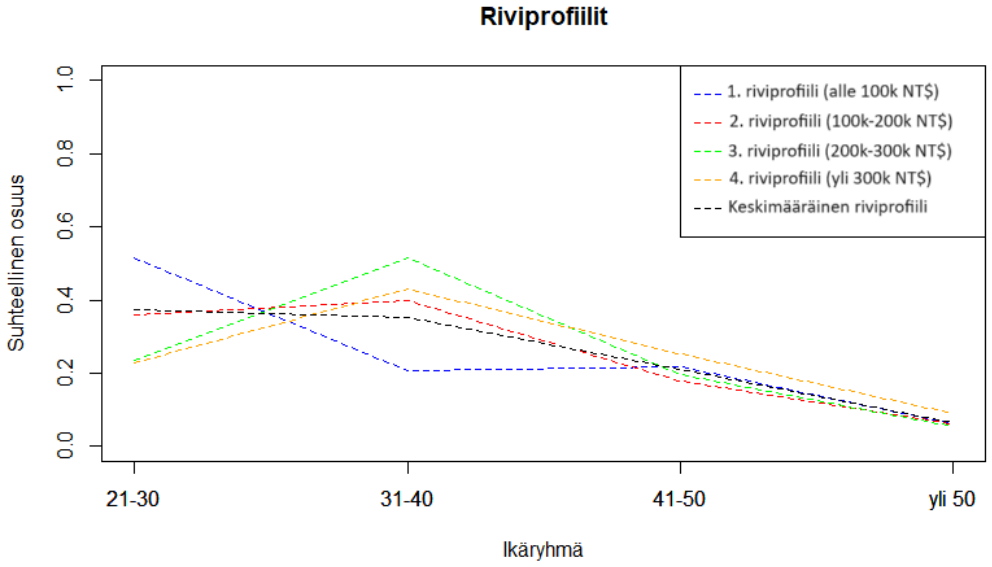
Riviprofiili i:lle riville määritetään seuraavasti:

$$\left(\frac{n_{i1}}{n_{i\cdot}}, \cdots, \frac{n_{is}}{n_{i\cdot}}\right) \tag{14}$$

Keskiarvoinen riviprofiili määritetään seuraavasti:

$$\left(\frac{n_{\cdot 1}}{n}, \cdots, \frac{n_{\cdot s}}{n}\right) \tag{15}$$

Käytetään esimerkkiaineistona luvun 4.1 aineistoa havainnollistamaan riviprofiileja. Muodostetaan riviprofiileista ja keskiarvoisesta riviprofiilista viivakuvaaja.



Kuva 5: Riviprofiilit

Kuvaajaan (kuva 5) on piirretty jokaisen luoton määrän riviprofiilit ikäryhmän suhteen. Musta katkoviiva kuvaa keskiarvoista riviprofiilia. Kuvaajasta voidaan tehdä hyvin samanlaisia huomioita kuin luvun 4.4 assosiaatiokuvaajista. 21–30-vuotiailla on suhteessa enemmän alle 100 000 NT$:n luottoja ja vähemmän yli 200 000 NT$:n luottoja. 31–40-vuotiailla on suhteessa vähemmän alle 100 000 NT$:n luottoja ja enemmän suuria luottoja. Muissa ikäryhmissä luottojen määrä on jakautunut tasaisemmin.

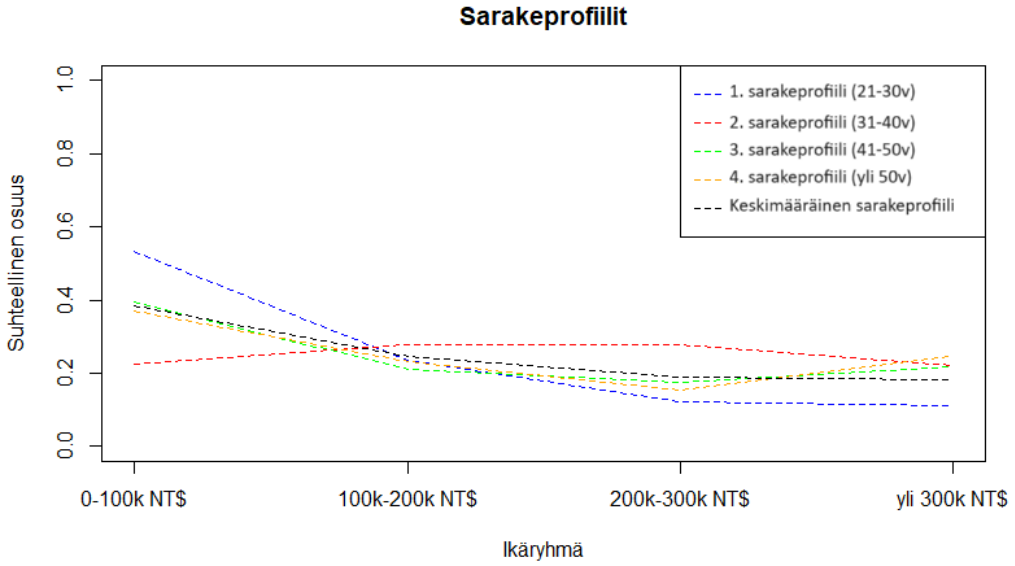
Sarakeprofiilit toimivat samalla tavalla kuin riviprofiilitkin, mutta tällöin mallinnetaan sarakkeita. Sarakeprofiili j:lle sarakkeelle määritetään seuraavasti:

$$\left(\frac{n_{1j}}{n_{\cdot j}}, \cdots, \frac{n_{rj}}{n_{\cdot j}}\right) \tag{16}$$

Keskiarvoinen sarakeprofiili määritetään seuraavasti:

$$\left(\frac{n_{1\cdot}}{n}, \cdots, \frac{n_{r\cdot}}{n}\right) \tag{17}$$

Muodostetaan myös sarakeprofiilit ja keskiarvoinen sarakeprofiili luvun 4.1 esimerkkiaineistolle:



Kuva 6: Sarakeprofiilit

Sarakeprofiilien (kuva 6) tulkinta on samankaltainen kuin riviprofiileissakin. Alle 100000 NT$:n luotot ovat yleisimpiä 21–30-vuotiailla, ja niitä on vähiten 31–40-vuotiailla. 100000–200000 NT$:n luotoissa ikäryhmät ovat jakautuneet varsin tasaisesti. Yli 200 000 NT$:n luottoja on eniten 31–40-vuotiailla ja vähiten 21–30-vuotiailla. Ikäryhmän 41–50 ja yli 50-vuotiaiden sarakeprofiilit seuraavat melko tarkasti keskimääräistä sarakeprofiilia.

6.2 Inertia

Inertia on tilastollinen käsite, joka liittyy korrespondenssianalyysiin vahvasti. Sen idea on samankaltainen kuin χ^2-testisuureella. Inertia mittaa aineistossa rivi- ja sarakemuuttujien välisten riippuvuuksien voimakkuutta kokonaisuutena. Vain kerroin $\frac{1}{n}$ erottaa inertian ja χ^2-testisuureen toisistaan. Inertia kertoo, kuinka suuri osa aineiston vaihtelusta selittyy tietyssä dimensiossa. Suurempi inertia tarkoittaa, että hajontaa selitetään enemmän kyseisessä dimensiossa. Ymmärtämällä inertian merkitys, voidaan paremmin kuvata ja

ymmärtää aineiston rakennetta ja sen riippuvuuksia. Inertia voidaan määritellä seuraavasti:

$$\phi^2 = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \tag{18}$$

jossa n_{ij} on havaittu frekvenssi ja e_{ij} on odotusarvon estimaatti (kaava (8)). Inertian neliöjuuri on jo aikaisemmin esitelty ϕ-kertoimena alaluvussa 3.2. Inertia voidaan jakaa rivikohtaisiin komponentteihin, jolloin erot eri dimensioiden välillä saadaan selville

$$\phi^2 = \sum_{i=1}^{r} \frac{n_{i\cdot}}{n} \left[\sum_{j=1}^{s} \frac{n}{n_{\cdot j}} \left(\frac{n_{ij}}{n_{i\cdot}} - \frac{n_{\cdot j}}{n} \right)^2 \right] = \frac{\chi^2}{n} \tag{19}$$

jossa hakasuluissa oleva summa on neliöity χ^2-etäisyys i:nnen riviprofiilin ja keskiarvoisen riviprofiilin välillä. Luvut $\left(\frac{n_{i\cdot}}{n}\right)$ ovat rivipainoja.

6.3 Korrespondenssianalyysi

Tässä alaluvussa esitellään korrespondenssianalyysi lyhyesti. Alaluku perustuu teokseen *Kyselytutkimuksen mittarit ja menetelmät* [8, luku 7]. Korrespondenssianalyysissa tarkastellaan ryhmien välisiä suhteita ja visualisoidaan ne kuvilla, jotka ovat keskeisemmässä roolissa kuin numeeriset tulokset verrattuna muihin monimuuttujamenetelmiin. Kuvat voivat olla haastavia, mutta antoisia.

Korrespondenssianalyysin tavoitteena on kuvata rivi- ja sarakemuuttujien välisiä riippuvuuksia. Tavoitteena on myös selvittää, missä ristitaulukon dimensioissa eroavaisuudet ilmenevät. Korrespondenssianalyysin lähtökohtana toimivat Pearsonin residuaalien (kaava (7)) muodostama matriisi, jota työstetään erilaisin matriisilaskennan keinoin. Korrespondenssianalyysin matemaattisesta käsittelystä lisää monimuuttujamenetelmät-kurssin diasarjassa [7, osio 9]. Tämän jälkeen tuloksia visualisoidaan erilaisten kuvaajien avulla.

Analyysin perustana on kahden muuttujan ristitaulukko, joka luokittelee muuttujat luokittelutasoisiksi. Korrespondenssianalyysi mahdollistaa yh-

teyksien tutkimisen monenlaisiin suhteisiin, ei vain lineaarisiin. Yleistyksissä tavallista taulukkoa laajennetaan eri tavoin, jotta siihen saadaan sisällytettyä useampia muuttujia. Korrespondenssianalyysia sovelletaan myös laadullisissa tutkimuksissa, missä muuttujat voivat olla kaikenlaisia mittaustasoiltaan.

Viitteet

[1] *Wikipedia: Contingency table*, `https://en.wikipedia.org/wiki/Contingency_table`, luettu 23.8.2023

[2] H.Pesonen: *Tilastollisen päättelyn peruskurssi.* Luentokalvot, luettu 23.8.2023.

[3] *Maksun laiminlyöminen -data*, `https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset`, ladattu 28.9.2023

[4] M.Szumilas: *Explaining Odds Ratios* (2010). Journal of the Canadian Academy of Child and Adolescent Psychiatry = Journal de l'Academie canadienne de psychiatrie de l'enfant et de l'adolescent, 19(3), 227–229.

[5] C.Chen et al: *Handbook of Data Visualization* (2008) (Springer Handbooks of Computational Statistics). Springer-Verlag TELOS, Santa Clara, CA, USA

[6] *Wikipedia: Chi-squared test*, `https://en.wikipedia.org/wiki/Chi-squared_test`, luettu 31.8.2023

[7] P.Nieminen: *Monimuuttujamenetelmät-kurssi.* Diasarja, luettu 12.10.2023

[8] K.Vehkalahti: *Kyselytutkimuksen mittarit ja menetelmät* (2008), Tammi.