

G Luku 1 Yksinkertainen korrespondenssianalyysi

Jussi Hirvonen

25.4.2018

Sisältö

1	Data	1
1.1	Luvun 1 tavoitteet	4
1.2	Perhe ja muuttuvat sukupuoliroolit - ISSP:n kyselytutkimuksen data 2012	4
1.3	Aineiston rajaaminen	4
1.4	Rajaukset	5
1.5	Puuttuvat tiedot (erävastauskato)	7
2	Yksinkertainen korrespondenssianalyysi - kahden luokittelumuuttujan taulukko	10
2.1	Äiti työssä	10
2.2	Korrespondenssianalyysin käsitteet	16
3	Tulkinnan perusteita	17
4	Yksinkertaisen korrespondenssianalyysin laajennuksia	17
4.1	Lisämuuttujat: ikäluokka ja sukupuoli	17
4.2	ABBA	17
4.3	Päällekkäiset matriisit (stacked matrices)	17

Kommentteja ja versionhallintaa:

- edit: oma kommentti, ei varsinaista tekstiä
- kirjastot/paketit ladataan jokaisessa Rmd-dokumentissa
- bib-formaatin viitetietokantaa tullaan kokeilemaan
- kuvasuhde (aspect ratio) edelleen epäselvä juttu!
- Datan käsittely ja hallinta +SPSS:n sallima kolme puuttuvan tiedon koodia saadaan mukaan read_spss-funktion (haven) parametrilla USER_NA = TRUE (mutta tarkistettava!) (25.5.18)
- faktoreita ei ainakaan toistaiseksi muuteta ordinaaliasteikolle, CA ei tästä välitä
- pidetään muuttujien ja tiedosojen nimeäminen selkeänä, tarkistetaan aika ajoin
- Taulukot: lisättiin riviprosentti- ja sarakeprosenttitaulut (25.5.18)

1 Data

Ladataan käytettävät paketit:

```
# pitääkö laittaa järjestykseen, vanhemmat ensin?  
library(rgl)  
library(ca)  
library(haven)  
library(dplyr)
```

```

##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##   filter, lag
## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union
library(knitr)
library(tidyverse)

## -- Attaching packages ----- tidyverse 1.2.0 --
## v ggplot2 2.2.1      v readr   1.1.1
## v tibble  1.4.2      v purrr   0.2.4
## v tidyr   0.7.2      v stringr 1.2.0
## v ggplot2 2.2.1      v forcats 0.2.0
## Warning: package 'tibble' was built under R version 3.4.3
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
library(lubridate)

##
## Attaching package: 'lubridate'
## The following object is masked from 'package:base':
##
##   date
library(rmarkdown)
library(ggplot2)
library(furniture)

## Warning: package 'furniture' was built under R version 3.4.4
## furniture 1.7.6: learn more at tysonbarrett.com
library(likert)

## Warning: package 'likert' was built under R version 3.4.4
## Loading required package: xtable
##
## Attaching package: 'likert'
## The following object is masked from 'package:dplyr':
##
##   recode
library(scales) # ggplot2 - kuznetsov

##
## Attaching package: 'scales'

```

```

## The following object is masked from 'package:purrr':
##
##   discard

## The following object is masked from 'package:readr':
##
##   col_factor

library(reshape2) # ggplot2 - kuvalle

##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##   smiths

# ehkä viimeiseksi sessionInfo() ja tämä koodi piiloon?
sessionInfo()

## R version 3.4.2 (2017-09-28)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 16299)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United Kingdom.1252
## [2] LC_CTYPE=English_United Kingdom.1252
## [3] LC_MONETARY=English_United Kingdom.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United Kingdom.1252
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets  methods    base
##
## other attached packages:
## [1] reshape2_1.4.2 scales_0.5.0   likert_1.3.5   xtable_1.8-2
## [5] furniture_1.7.6 rmarkdown_1.7   lubridate_1.7.1 forcats_0.2.0
## [9] stringr_1.2.0 purrr_0.2.4     readr_1.1.1   tidyr_0.7.2
## [13] tibble_1.4.2 ggplot2_2.2.1  tidyverse_1.2.0 knitr_1.17
## [17] dplyr_0.7.4 haven_1.1.0     ca_0.70       rgl_0.98.1
##
## loaded via a namespace (and not attached):
## [1] Rcpp_0.12.13 lattice_0.20-35 assertthat_0.2.0 rprojroot_1.2
## [5] digest_0.6.12 psych_1.7.8 mime_0.5 R6_2.2.2
## [9] cellranger_1.1.0 plyr_1.8.4 backports_1.1.1 evaluate_0.10.1
## [13] httr_1.3.1 pillar_1.2.1 rlang_0.2.0 lazyeval_0.2.1
## [17] readxl_1.0.0 rstudioapi_0.7 foreign_0.8-69 htmlwidgets_0.9
## [21] munsell_0.4.3 shiny_1.0.5 broom_0.4.2 compiler_3.4.2
## [25] httpuv_1.3.5 modelr_0.1.1 pkgconfig_2.0.1 mnormt_1.5-5
## [29] htmltools_0.3.6 gridExtra_2.3 crayon_1.3.4 grid_3.4.2
## [33] nlme_3.1-131 jsonlite_1.5 gtable_0.2.0 magrittr_1.5
## [37] cli_1.0.0 stringi_1.1.5 bindrcpp_0.2 xml2_1.1.1
## [41] tools_3.4.2 glue_1.2.0 hms_0.3 parallel_3.4.2
## [45] yaml_2.1.14 colorspace_1.3-2 rvest_0.3.2 bindr_0.1.1

```

Yksinkertainen korrespondenssianalyysi on kahden luokitteluasteikon muuttujan riippuvuuksien geometrista analyysiä. Lähtökohta on kahden muuttujan ristiintaulukointi, alkuperäinen data voi olla muillakin asteikoilla mitattua. Menetelmän ydin on tarkastella molempien muuttujien – taulukon rivien ja sarakkeiden – riippuvuuksia kaksiulotteisena kuvana. Kuvaa kutsutaan myös kartaksi, ja tulkinnan ensimmäinen askel on kartan “koordinaatiston” tulkinta. Kaikki etäisyydet kuvassa ovat suhteellisia, vain rivi- ja sarakepisteiden etäisyydet kuvan origosta voidaan tulkita tarkasti. Koordinaatiston tulkinta aloitetaan “katsomalla mitä on oikealla ja vasemmalla, ja mitä on ylhäällä ja alhaalla” (viite LeRoux et.al, Bezecri-sitaatti). Vaikka pisteiden etäisyyksiä edes rivi- ja sarakepisteiden välillä ei voi tarkkaan tulkita (approksimaatioita), projektiossa kaukana toisistaan olevat pisteet ovat kaukana toisistaan myös alkuperäisessä “pistepilvessä”.

1.1 Luvun 1 tavoitteet

1. Data - tässä tiiviimmin, aineiston kuvailu tarkemmin liitteeseen. Perustella rajaukset ja kertoa miten ne tehdään.
2. Ensimmäinen taulukko: profiilit, massat, keskiarvoprofiilit, khii2 - riippumattomuustesti ja etäisyysmitta
3. Hyvin tiivis esitys CA:n perusideasta, mutta ilman aivan simppleitä kolmiulotteisia kuvia (niitä on jo)
4. Ensimmäinen symmetrinen kartta, perustulkinta (mitä kuvasta voidaan sanoa, mitä ei)
5. Lyhyt viittaus graafisen esityksen tulkintapulmiin, jotka eivät ole kovin pahoja. Niihin palataan kaksoiskuva-jaksossa.
6. Tulkinnan syventäminen - “kaikki selitetetään”

1.2 Perhe ja muuttuvat sukupuoliroolit - ISSP:n kyselytutkimuksen data 2012

Hieman historiaa datasta, sosiaalisesti määräytyneen sukupuoliroolit (gender) tutkimusaiheena neljässä kansainvälisessä kyselytutkimuksessa.

ISSP Research Group (2016): International Social Survey Programme: Family and Changing Gender Roles IV - ISSP 2012. GESIS Data Archive, Cologne. ZA5900 Data file Version 4.0.0, doi:10.4232/1.12661

[Aineistot] (https://search.gesis.org/research_data/ZA5900) 2012

[Muuttujakuvaukset ja muut tiedot] (<http://zacat.gesis.org/webview/index.jsp?object=http://zacat.gesis.org/obj/fStudy/ZA5900>)

[Suomenkielinen lomake (ZA5900_q-fi-fi.pdf)] (<https://dbk.gesis.org/dbksearch/sdesc2.asp?no=5900&db=e&doi=10.4232/1.12661>)

[Käyttöehdot:] (<https://www.gesis.org/en/services/data-analysis/more-data-to-analyze/data-archive-service/>)

[Tiedonkeruumenetelmä ja otoskoko:] (<https://dbk.gesis.org/dbksearch/sdesc2.asp?no=5900&db=e&doi=10.4232/1.12661>) Viimeisin Portugali 29.06.2014 - 31.01.2015, ensimmäinen Bulgaria 16.08.2011 - 20.09.2011. Suurin osa muista 2012-13, kuten Suomi (21.09.2012 - 07.12.2012).

edit: aineiston kuvailua voi ja kannattaakin jatkossa tarkentaa, ja laittaa se liitteeksi. Dokumentointi on hyvin tarkka, tiedot löytyvät haastattelumenetelmistä (parerilomake, tietokoneavusteinen haastattelu, jne), maakohtaisten taustamuuttujien harmonisoinnista maittain, otantamenetelmistä jne. Esittelen vain aineiston tärkeimmät rajaukset.

1.3 Aineiston rajaaminen

Aineistossa (jatkossa ISSP2012) on kyselytutkimukseen tulokset 41 maasta. Lisäksi aineistossa on runsaasti demografisia ja muita taustatietoja. R-koodista selviää käytetty versio (SPSS-tiedoston nimi) ja rajauksessa

käytetyt muuttujat.

1.4 Rajaukset

1. Eurooppa ja samankaltaiset maat (28)

Bulgaria, Czech Republic, Denmark, Finland, France, Germany, Great Britain, Ireland, Latvia, Lithuania, Norway, Poland, Sweden, Slovakia Slovenia, Spain, Switzerland, Australia, Austria, Canada, Croatia, Iceland, Russia, United States, Belgium, Hungary, Netherlands, Portugal

Pois jätettiin 13: Argentiina, Turkki, Venezuela, Etelä-Afrikka, Korea, Intia, Kiina, Taiwan, Filippiinit, Meksiko, Israel, Japani, Chile.

2. Maat joissa varsinaisissa tutkimuskysymyksissä on käytetty poikkeavia luokituksia tms. Esimerkiksi Espanjan datassa on jätetty pois neutraali "en samaa enkä eri mieltä" - vaihtoehto, Unkarin datassa on omia versioita kysymyksistä jne. Espanja jätetään ainakin aluksi pois vertailukelpoisuuden vuoksi, Unkari ehkä myös.

3. kaikki havainnot, joissa on puuttuvia tietoja. Tämä raja on kyselytutkimuksessa ankara, tai oikeastaan keltainen. Oikea menettely olisi imputoida jollain menetelmällä puuttuvat tiedot, mutta raja otantatutkimuksen menetelmät tutkielman ulkopuolelle (aiheesta löytyy artikkeleita...). Yksittäisten vastausten puuttuminen eli erävastauskato ohitetaan aluksi, mutta siihen palataan. Korrespondenssianalyysiin on helppo ottaa mukaan myös puuttuvat tiedot, sillä data on luokitteluasteikon dataa. Yksikkövastauskato eli otokseen poimitut joita ei ole tavoitettu ollenkaan on kansallisen tason ongelma, joka on ratkaistu vaihtelevin tavoin. Tiedot löytyvät aineiston dokumentaatiosta. Aineistossa on myös mukana painomuuttujat, mutta ne soveltuvat vain jokaisen maan omaan aineistoon.

edit: Tähän täsmennetään miten puuttuvia tietoja käsitellään.

4. Datan hallinta

Aineistoa käsitellään ja muokataan niin, että jokaisen analyysin voi mahdollisimman yksinkertaisesti toistaa suoraan alkuperäisestä datasta.

Aineiston muokkauksen (muuttujien ja havaintojen valikointi, muunnokset ja uusien muuttujien luonti jne.) dokumentoidaan r-koodiin.

```
# kolme maa-muuttujaa datassa. V3 erottaa joidenkin maiden alueita, V4 on koko maan
#two country code variables based on the ISO Code 3166. One identifies
#countries as a whole, the other one possible subsamples, such as East and West Germany. The cross
#tabulations shown in this Variable Report are based on a third, alphanumerical country code variable,
#which also identifies subsamples."
#V3 - Country/ Sample ISO 3166 Code (see V4 for codes for whole nation states)
# V3 erot valituissa maissa
# 5601 BE-FLA-Belgium/ Flanders
# 5602 BE-WAL-Belgium/ Wallonia
# 5603 BE-BRU-Belgium/ Brussels
# 27601 DE-W-Germany-West
# 27602 DE-E-Germany-East
# 62001 PT-Portugal 2012: first fieldwork round (main sample)
# 62002 PT-Portugal 2012: second fieldwork round (complementary sample)
# Myös tämä on erikoinen, näyttää olevan vakio kun V4 = 826:
# 82601 GB-GBN-Great Britain
# Portugalissa aineistoa täydennettiin, koska siinä oli puutteita. Jako ei siis ole oleellinen,
# mutta muut ovat. Tähdellä merkityt maat valitaan johdattelevaan esimerkkiin.
# Maat:
# 36 AU-Australia
# 40 AT-Austria
```

```

# 56 BE-Belgium*
# 100 BG-Bulgaria*
# 124 CA-Canada
# 191 HR-Croatia
# 203 CZ-Czech Republic
# 208 DK-Denmark*
# 246 FI-Finland*
# 250 FR-France
# 276 DE-Germany*
# 348 HU-Hungary*
# 352 IS-Iceland
# 372 IE-Ireland
# 428 LV-Latvia
# 440 LT-Lithuania
# 528 NL-Netherlands
# 578 NO-Norway
# 616 PL-Poland
# 620 PT-Portugal
# 643 RU-Russia
# 703 SK-Slovakia
# 705 SI-Slovenia
# 724 ES-Spain
# 752 SE-Sweden
# 756 CH-Switzerland
# 826 GB-Great Britain and/or United Kingdom
# 840 US-United States
#
# Belgian ja Saksan alueet:
# V3
# 5601 BE-FLA-Belgium/ Flanders
# 5602 BE-WAL-Belgium/ Wallonia
# 5603 BE-BRU-Belgium/ Brussels
# 27601 DE-W-Germany-West
# 27602 DE-E-Germany-East

#valittavien maiden kolminumeroinen ISO 3166 - koodi vektoriin
incl_countries <- c(36, 40, 56,100, 124, 191, 203, 208, 246, 250, 276, 348, 352, 372, 428, 440,
                    528, 578, 616, 620, 643, 703, 705, 724, 752, 756, 826, 840)
ISSP2012.data <- read_spss("data/ZA5900_v4-0-0.sav", user_na = TRUE)
#
# lisäys 25.4.2018 user_na
# "If TRUE variables with user defined missing will be read into labelled_spss objects.
# If FALSE, the default, user-defined missings will be # converted to NA"
# https://www.rdocumentation.org/packages/haven/versions/1.1.0/topics/read_spss
#
#
#
#str(ISSP2012.data) #61754 obs. of 420 variables
ISSP2012jh1.data <- filter(ISSP2012.data, V4 %in% incl_countries)
#length((ISSP2012jh1.data))
#names(ISSP2012jh1.data)
#str(ISSP2012jh1.data) #37816 obs. of 420 variables
#

```

```
#EDIT: tiivistä, nämä ovat vain kokeiluja ja datan kaivelua (15.4.2018)
#
# V5 - V67 kysymyksiä, joillain mailla omat vastaukset joihinkin omina muuttujina, esim. # ES_V5 muut.
#$ V5      :Class 'labelled'  atomic [1:37816] 5 1 2 2 1 NA 2 4 2 2 ...
# .. - attr(*, "label")= chr "Q1a Working mom: warm relationship with children as a not working mom"
# .. - attr(*, "format.spss")= chr "F1.0"
# .. - attr(*, "labels")= Named num [1:8] 0 1 2 3 4 5 8 9
# .. - attr(*, "names")= chr [1:8] "NAP: ES" "Strongly agree" "Agree" "Neither agree nor disagree"
# $ ES_V5   :Class 'labelled'  atomic [1:37816] NA NA NA NA NA NA NA NA NA ...
# .. - attr(*, "label")= chr "Q1a Working mom: warm relationship with children as a not working mom"
# .. - attr(*, "format.spss")= chr "F1.0"
# .. - attr(*, "display_width")= int 4
# .. - attr(*, "labels")= Named num [1:7] 0 1 2 3 4 8 9
# .. - attr(*, "names")= chr [1:7] "NAP: other countries" "Strongly agree" "Agree" "Disagree" ...
#HU_V18
#V18$label
#attr(ISSP2012jh1.data$V6,'labels')
#attr(ISSP2012jh1.data$ES_V6,'labels')
```

Yllä esimerkiksi muuttujan V6 metatiedot. Perusvaihtoehdot ovat 1 - 5, ja joillain mailla on vaihtoehtona ollut myös “Can’t choose”, muilla taas on vain puuttuva tieto (No answer, 9).

Espanjan aineiston metatiedot muuttujalla ES_V6 taas ovat

```
attr(ISSP2012jh1.data$ES_V5,'labels')
```

```
## NAP: other countries      Strongly agree      Agree
##                0                1                2
##                Disagree    Strongly disagree    Can't choose
##                3                4                8
##                No answer
##                9
```

```
temp1 <- ISSP2012jh1.data %>% filter(V4 == 724) %>% select(ES_V6, C_ALPHAN)
#str(temp1)
temp1$ES_V6 <- factor(temp1$ES_V6 )
summary(temp1)
```

```
## ES_V6      C_ALPHAN
## 1: 195   Length:2595
## 2:1117   Class :character
## 3: 898   Mode  :character
## 4: 278
## 8:  91
## 9:  16
```

```
#typeof(ISSP2012jh1.data)
#class(ISSP2012jh1.data)
#storage.mode(ISSP2012jh1.data)
#attributes(ISSP2012jh1.data)
```

1.5 Puuttuvat tiedot (erävastauskato)

EDIT tämä muuttuu kun puuttuvissa on tarkentava tieto mukana. Tuskin käytetään, joten tekstiä voi reippaasti supistaa. (25.4.18)

Datassa ei ole eroteltu vastausvaihtoehtoa “Can’t choose” (8) ja “No answer” (9), ne on (luultavasti) yhdistetty ja koodattu puuttuviksi havainnoiksi. Dokumentaatiosta selviää (s.13), että vaihtoehtoon 8 on valinnut 30 ja loput neljä “puuttuvaa tietoa” ovat erävastaukatoa (tai kieltäytymistä tms.). Jokaisen kysymyksen vastauksista löytyy aineiston dokumentaatiosta taulukko, joissa puuttuva tieto on eritelty tarkemmin.

Muiden kuin Espanjan vastaukset kysymykseen V6 jakautuvat näin:

```
temp2 <- ISSP2012jh1.data %>% filter(!(V4 == 724)) %>% select(V6, C_ALPHAN)
```

```
#str(temp1)
```

```
temp2$V5 <- factor(temp2$V6 )
```

```
temp2$maa <- factor(temp2$C_ALPHAN)
```

```
summary(temp2)
```

```
##          V6          C_ALPHAN          V5          maa
## Min.      :1.000  Length:35221      1:2881  FR      : 2409
## 1st Qu.:2.000   Class :character    2:9019  BE      : 2202
## Median :3.000   Mode  :character    3:6829  CZ      : 1804
## Mean    :3.181                                4:9576  DE      : 1766
## 3rd Qu.:4.000                                5:5675  AU      : 1612
## Max.    :5.000                                8: 875  RU      : 1525
## NA's    :1241                                9: 366  (Other):23903
```

```
temp2 %>% tableX(V6,maa,type = "count")
```

```
##          maa
## V6      AT  AU  BE  BG  CA  CH  CZ  DE  DK  FI  FR  GB-GBN
## 1       218  82  193 118  51  89  174 165  70  47  256  37
## 2       447 405  454 395 215 431  392 376 238 188  551 247
## 3       171 285  440 205 181 222  403 199 152 149  424 208
## 4       205 568  554 190 317 365  415 538 232 423  469 331
## 5        98 215  381  13 194 112  355 441 696 303  624 105
## Missing 43  57  180  82  14  18   65  47  15  61  85  22
## Total   1182 1612 2202 1003 972 1237 1804 1766 1403 1171 2409 950
##          maa
## V6      HR  HU  IE  IS  LT  LV  NL  NO  PL  PT  RU  SE  SI
## 1       75  219  56  13  50  188  59  23  110  73  244  29  39
## 2       265 288  250 138 438  395  296 186  395 495  542 124 272
## 3       190 225  197 186 396  156  242 226  155 157  360 219 200
## 4       327 190  478 552 220  209  445 579  365 215  254 276 365
## 5       133  75  197 271  22  38  196  365  64  52  42  354 131
## Missing 10  15  37  12  61  14  77  65  26  9  83  58  27
## Total   1000 1012 1215 1172 1187 1000 1315 1444 1115 1001 1525 1060 1034
##          maa
## V6      SK  US  Total
## 1      117  86  2881
## 2      246 350  9019
## 3      229 652  6829
## 4      298 196  9576
## 5      198  0  5675
## Missing 40  18  1241
## Total   1128 1302 35221
```

Esimerkiksi Ruotsin puuttuviksi tiedoiksi koodatuista 29 havainnosta 19 valitsi “can’t choose”(8) ja 10 kieltäytyi vastaamasta (9) tms. Dokumentti, s.12.

Tarkastellaan aineiston puuttuvia havaintoja hieman tarkemmin. Puuttuvat tiedot on koodattu aineistoon

näin: 0: Not applicable (NAP), Not available (NAV) 7: (97,997, 9997,...): Refused 8: (98, 998, 9998,...): Don't know 9: (99, 999, 9999,...): No answer

NAP ja NAV määritellään

“GESIS adds ‘Not applicable’(NAP) codes for questions that have filters. NAP indicates that only a subsample and not all of respondents were asked. Also in the case of country specific variables, all the other countries are coded NAP.

GESIS adds ‘Not available’ for variables, which in single countries may not have been conducted for whatever reason.”

EDIT: Puuttuneisuuden lyhyttä kuvailua, ja rajausten vaikutus havaintojen lukumäärään muutamaan taulukkoon. Voi siirtää liitteisiin (25.4.2018)

#Tähän lisätään erävastauskadon kuvailua - 10.4.2018

#Nyt edetään esimerkkien, jossa mukana kuusi maata ja kysymys V6 - palataan kun aineistoa laajennetaan.

1.5.1 Poikkeavat kysymykset

Aineistossa on ns. substanssimuuttujia 63 (V5 - V67). Suurin osa on kerätty jollain haastattelumenetelmällä, ja yleisin vastausvaihtoehto on viiden arvon Likert-skaala (1 = täysin samaa mieltä, samaa mieltä, en samaa enkä eri mieltä, eri mieltä, täysin eri mieltä =5). Eri maiden lomakkeissa on vaihtelua puuttuviksi tiedoiksi koodattujen muiden vastausten välillä. Esimerkikseen Suomen lomakkeessa on kuudes vaihtoehto “en osaa sanoa”, ja lisäksi on koodattu vastaamisesta kieltäytyminen tai muuten puuttuva tieto. SPSS-aineistossa nämä kaikki on koodattu puuttuviksi havainnoiksi.

Espanjan lisäksi Unkarin osatutkimuksessa kysymyksen V18 V19 V20 vastausvaihtoehdot ovat poikkeavat siten, että keskimäinen neutraali vaihtoehto on jätetty pois (em.dok, s. 48).

edit: nämä merkinnät ovat muistiinpanoja, kun tarkemmin luin muuttujadokumenttia. Kysymyksissä on vaihtelua, ja tavallaan niin pitääkin olla kansainvälisessä kyselytutkimuksessa. Vastajien on ymmärrettävä kysymyksen suurinpiirtein samalla tavalla. Kaikki on tarkasti dokumentoitu.

Islannissa kysymykseen V28 (Consider a couple who both work full-time and now have a new born child. One of them stops working for some time to care for their child. Do you think there should be paid leave available and, if so, for how long?) on tarjolla oma vastausvaihtoehto ((97) “Yes, but don't know how many months”). Kysymyseen “V29 - Q9 Paid leave: Who should pay ja V30(Paid leave: How to divide between parents) Bulgarian kysely on poikkeava (0 NAP (code 0,98 in V28), s. 91).

Hollannin vastausvaihtoehdoissa kysymykseen V35 (Elderly people: Provider of domestic help) on oma variantti “5 Employers”, jonka kuitenkin on valinnut vain 6 vastajaa (0,5 %).

V39, V40, V41, V42, V43, V44, V45, V46, V47, V48, V50, V51, V52, V53, V54: paljon poikkeamia, aika vaikeaselkoisia kysymyksiä. Näitä ehkä pitää tutkailla... V55 (Life in general: How happy on the whole) ok.

V56-57 poikkeamia, V58 (Health status) ok V59 “ketjutettu kysymys”, samoin V60-V64. s. 174 - puolison koulutus...

edit: täsmennettävä, periaatteessa vastaukset on harmonisoitu. Joistain maista joku tieto puuttuu, jos sitä ei ole kysytty. Joissain tapauksissa kysymysten vaihtoehdot poikkeavat standardista.

2 Yksinkertainen korrespondenssianalyysi - kahden luokittelu-muuttujan taulukko

Tässä esitellään yksinkertainen esimerkki, yksi kysymys (esim. V6) ja muutamia maita ristiintaulukoituna. Johdatteluna aiheeseen esitellään ca-käsitteet profiili, massa ja reunajakauma. Havainnollistetaan rivi- ja sarakeprofiilien vertailua vastaaviin keskiarvoprofiileihin.

Toiseksi riippumattomuushypoteesi ja χ^2 - riippumattomuustesti (pieni huomautus - on monta tapaa testata taulukon riippuvuuksia). Riippumattomuushypoteesi ehdollisena todennäköisyytenä reunajakauman suhteen.

χ^2 - etäisyys, yhteys hajontaa eli inertiaan ca-terminologiassa.

Dimensioiden vähentämisen idea.

Ensimmäinen symmetrinen kartta, tulkinnat ja yksinkertaisimmat perussäännöt ("mitä on oikealla ja vasemmalla"). Jos pisteet ovat alkuperäisessä "pilvessä" kaukana toisistaan, ne ovat sitä myös projektiossa. Kartta, mutta etäisyyksillä ei suoraa tulkintaa paitsi etäisyyksinllä origoon. Rivipisteiden suhteelliset etäisyydet, samoin sarakepisteiden, mutta ei muut.

2.1 Äiti työssä

Aineisto muuttujat V5-V9 ovat vastauksia (1-5 Likert, täysin samaa mieltä - täysin eri mieltä) seuraaviin kysymyksiin (suomenkielinen lomake, kysymys 23):

- (a) Työssäkäyvä äiti pystyy luomaan lapsiinsa aivan yhtä lämpimän ja turvallisen suhteen kuin äiti, joka ei käy työssä
- (b) Alle kouluikäinen lapsi todennäköisesti kärsii, jos hänen äitinsä käy työssä
- (c) Kaiken kaikkiaan perhe-elämä kärsii, kun naisella on kokopäivätyö
- (d) On hyvä käydä töissä mutta tosiasiaassa useimmat naiset haluavat ensisijaisesti kodin ja lapsia
- (e) Kotirouvana oleminen on aivan yhtä antoisaa kuin ansiotyön tekeminen

```
#vähän hankalaa jos Rmd-tiedoston 'scope' vaatii aina kaiken ajamisen joka tiedostossa!
incl_esim1 <- c(56, 100, 208, 246, 276, 348) #BE,BG,DK,FI,DE,HU)
ISSP2012.data <- read_spss("data/ZA5900_v4-0-0.sav", user_na = TRUE) # Alkuperäinen data
#
# lisäys 25.4.2018 user_na
# "If TRUE variables with user defined missing will be read into labelled_spss objects.
# If FALSE, the default, user-defined missings will be # converted to NA"
# https://www.rdocumentation.org/packages/haven/versions/1.1.0/topics/read_spss
#
# str(ISSP2012.data)
#61754 obs. of 420 variables ja 61754 obs. of 420 variables 25.4.18
#
# kuusi maata
ISSP2012esim1.dat <- filter(ISSP2012.data, V4 %in% incl_esim1)
#str(ISSP2012esim1.dat) #8557 obs. of 420 variables
#
# mukaan muuttujat, V3 jos halutaan jakaa Saksa ja Belgia
# SEX 1=male, 2=female AGE haastateltava ikä haastatteluhetkellä
#
ISSP2012esim1.dat <- select(ISSP2012esim1.dat, C_ALPHAN, V3,V4, V6, SEX, AGE)

#str(ISSP2012esim1.dat) #8557 obs. of 6 variables
```

```
#
#poistetaan havainnot, joissa puuttuvia tietoja
ISSP2012esim1.dat <- filter(ISSP2012esim1.dat, (!is.na(V6) & !is.na(SEX) & !is.na(AGE)))
#str(ISSP2012esim1.dat) #8143 obs. of 6 variables
ISSP2012esim1.dat %>% table1(C_ALPHAN, splitby = V6)
```

```
##
## -----
##                               V6
##           1           2           3           4           5
##           n = 810     n = 1935    n = 1367    n = 2125    n = 1906
## C_ALPHAN
## BE    191 (23.6%) 451 (23.3%) 438 (32%)   552 (26%)   381 (20%)
## BG    118 (14.6%) 395 (20.4%) 205 (15%)   190 (8.9%)   13 (0.7%)
## DE    165 (20.4%) 375 (19.4%) 198 (14.5%) 538 (25.3%) 438 (23%)
## DK     70 (8.6%)  238 (12.3%) 152 (11.1%) 232 (10.9%) 696 (36.5%)
## FI     47 (5.8%)  188 (9.7%)  149 (10.9%) 423 (19.9%) 303 (15.9%)
## HU    219 (27%)   288 (14.9%) 225 (16.5%) 190 (8.9%)  75 (3.9%)
## -----
```

Havaintojen lukumäärät voi tarkistaa [täältä] (<http://zacat.gesis.org/webview/index.jsp?object=http://zacat.gesis.org/obj/fStudy/ZA5900>) .

Tehdään aineistoon muutama muutos, jotta sen käsittely on helpompaa.

```
# muutetaan muuttujia faktoreiksi
#
# Luokittelumuuttujien tasoille labelit
#
# sp (sukupuoli) m = 1, f = 2
sp_labels <- c("m", "f")
# 1 = täysin samaa mieltä, 2 = samaa mieltä, 3 = ei samaa eikä eri, 4 = eri mieltä, 5 = täysin eri miel
vastaus_labels <- c("ts", "s", "ese", "e", "te")

# Faktoreiksi
ISSP2012esim1.dat$maa <- factor(ISSP2012esim1.dat$C_ALPHAN)
ISSP2012esim1.dat$sp <- factor(ISSP2012esim1.dat$SEX, labels = sp_labels)
ISSP2012esim1.dat$V6 <- factor(ISSP2012esim1.dat$V6, labels = vastaus_labels)
#
#tsekkauksia
#ISSP2012esim1.dat %>% tableX(maa,V6,type = "count")
#summary(ISSP2012esim1.dat$sp)
#
#Apuvälineitä - lisätietoa muuttujista
# kun faktoroidaan V6, niin metadata katoaa?
#
# typeof(ISSP2012esim1.dat$V6) # what is it?
# class(ISSP2012esim1.dat$V6) # what is it? (sorry)
# storage.mode(ISSP2012esim1.dat$V6) # what is it? (very sorry)
# length(ISSP2012esim1.dat$V6) # how long is it? What about two dimensional objects?
# attributes(ISSP2012esim1.dat$V6) # does it have any metadata?
# str(ISSP2012esim1.dat) #8143 obs. of 8 variables

# Taulkoidaan data
```

```
ISSP2012esim1.dat %>% tableX(maa, V6, type = "count")
```

```
##          V6
## maa      ts  s    ese  e    te  Total
## BE      191 451  438  552  381  2013
## BG      118 395  205  190  13   921
## DE      165 375  198  538  438  1714
## DK       70 238  152  232  696  1388
## FI       47 188  149  423  303  1110
## HU      219 288  225  190  75   997
## Total   810 1935 1367 2125 1906 8143
```

```
ISSP2012esim1.dat %>% tableX(maa,V6,type = "cell_perc")
```

```
##          V6
## maa      ts  s    ese  e    te  Total
## BE      2.35 5.54  5.38  6.78  4.68  24.72
## BG      1.45 4.85  2.52  2.33  0.16  11.31
## DE      2.03 4.61  2.43  6.61  5.38  21.05
## DK      0.86 2.92  1.87  2.85  8.55  17.05
## FI      0.58 2.31  1.83  5.19  3.72  13.63
## HU      2.69 3.54  2.76  2.33  0.92  12.24
## Total   9.95 23.76 16.79 26.10 23.41 100.00
```

```
ISSP2012esim1.dat %>% tableX(maa,V6,type = "row_perc")
```

```
##          V6
## maa      ts  s    ese  e    te  Total
## BE      9.49 22.40 21.76 27.42 18.93 100.00
## BG     12.81 42.89 22.26 20.63  1.41  100.00
## DE      9.63 21.88 11.55 31.39 25.55 100.00
## DK      5.04 17.15 10.95 16.71 50.14 100.00
## FI      4.23 16.94 13.42 38.11 27.30 100.00
## HU     21.97 28.89 22.57 19.06  7.52  100.00
## All      9.95 23.76 16.79 26.10 23.41 100.00
```

```
ISSP2012esim1.dat %>% tableX(maa,V6,type = "col_perc")
```

```
##          V6
## maa      ts  s    ese  e    te  All
## BE     23.58 23.31 32.04 25.98 19.99 24.72
## BG     14.57 20.41 15.00  8.94  0.68 11.31
## DE     20.37 19.38 14.48 25.32 22.98 21.05
## DK      8.64 12.30 11.12 10.92 36.52 17.05
## FI      5.80  9.72 10.90 19.91 15.90 13.63
## HU     27.04 14.88 16.46  8.94  3.93 12.24
## Total  100.00 100.00 100.00 100.00 100.00 100.00
```

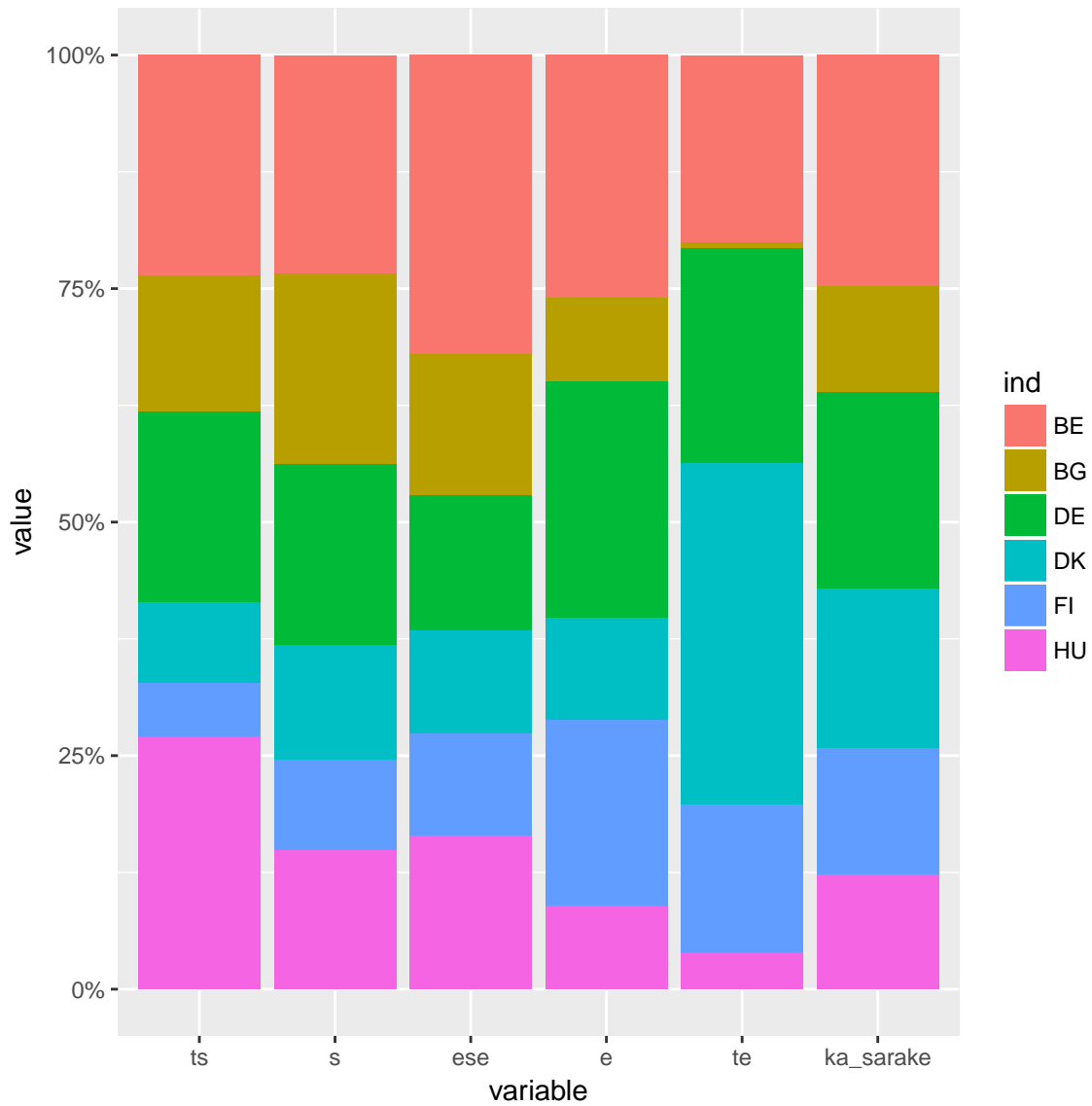
Taulukoissa on kuuden maan vastausten jakauma kysymykseen “Alle kouluikäinen lapsi todennäköisesti kärsii, jos hänen äitinsä käy työssä”. Taulukko on pieni, mutta havaintoja on melko paljon (N=8143). Alemman suhteellisten frekvenssien taulukon rivejä voi verrata toisiinsa ja alimpaan (“Total”) keskimääräiseen riviin, sarakemuuttujien eli vastausvaihtoehtojen reunajakaumaan. Vastavasti sarakkeita voi verrata rivimuuttujien reunajakaumasarakkeeseen (“Total2”). Eniten vastaajia on Belgiasta (25 %) ja Saksasta (21 %), vähiten Unkarista (12 %).

EDIT: Pienenkin taulukon pyörittely johdattelee hyvin, mihin korrespondenssianalyysiä tarvitaan. Näistä riippuvuuden rakenteet näkee ilmeisesti, jos on tarpeeksi nokkelia. Muiden pitää käyttää CA:ta.

edit: pitäisikö myös riviprofiileja havainnollistaa kuvalla?

```
#tauluG121 <- ISSP2012esim1.dat %>% tableX(maa, V6, type = "count")
#str(tauluG121)
#apu1 <- (tauluG121[-7, -6])
#str(apu1)
#apu1
#(rowSums(apu1))
#mutkikas kuvan piirto - sarakeprofiilit vertailussa
#ggplot vaatii df-rakenteen ja 'long data' - muotoon
#
# käytetään ca - tuloksia
apu1 <- (simpleCA1$N)
colnames(apu1) <- c("ts", "s", "ese", "e", "te")
rownames(apu1) <- c("BE", "BG", "DE", "DK", "FI", "HU")
apu1_df <- as.data.frame(apu1)
#lasketan rivien reunajakauma
apu1_df$ka_sarake <- rowSums(apu1_df)
#muokataan 'long data' - muotoon
apu1b_df <- melt(cbind(apu1_df, ind = rownames(apu1_df)), id.vars = c('ind'))

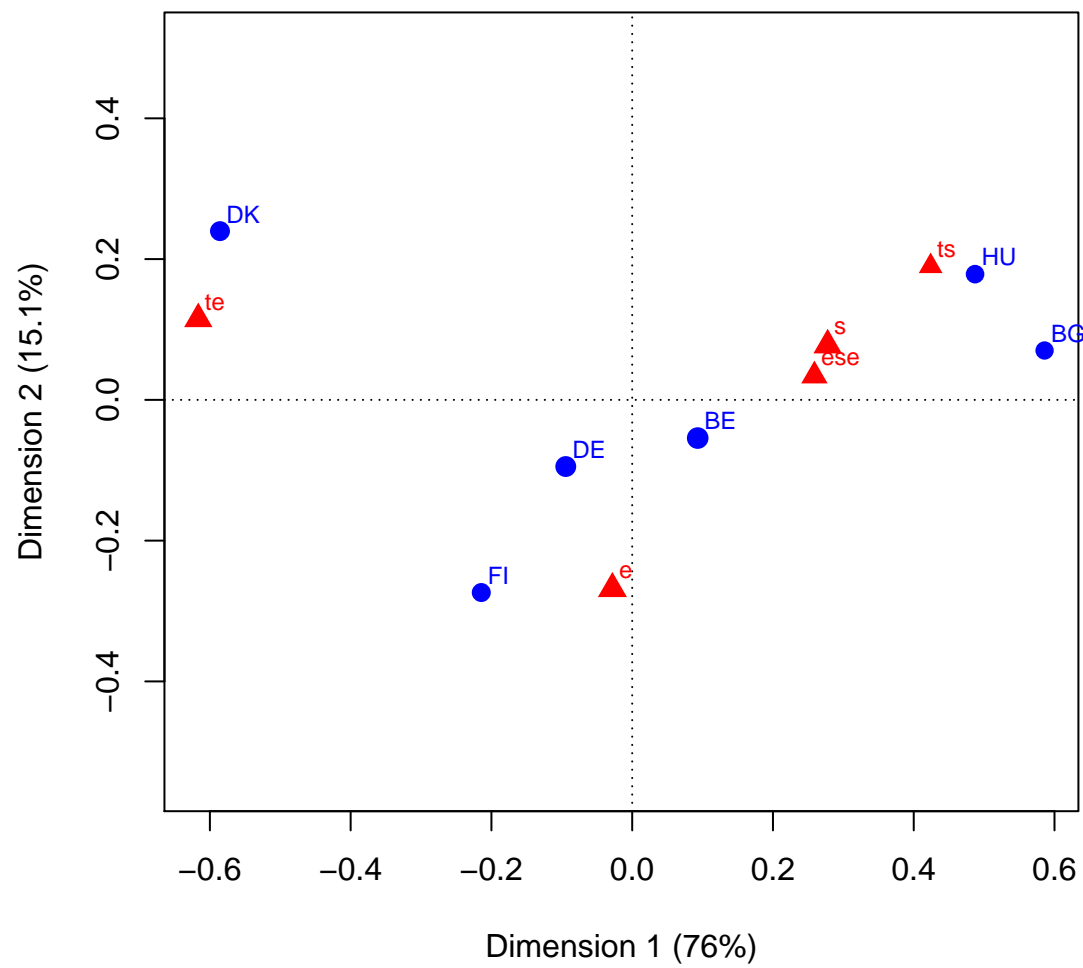
ggplot(apu1b_df, aes(x = variable, y = value, fill = ind)) +
  geom_bar(position = "fill", stat = "identity") +
  scale_y_continuous(labels = percent_format())
```



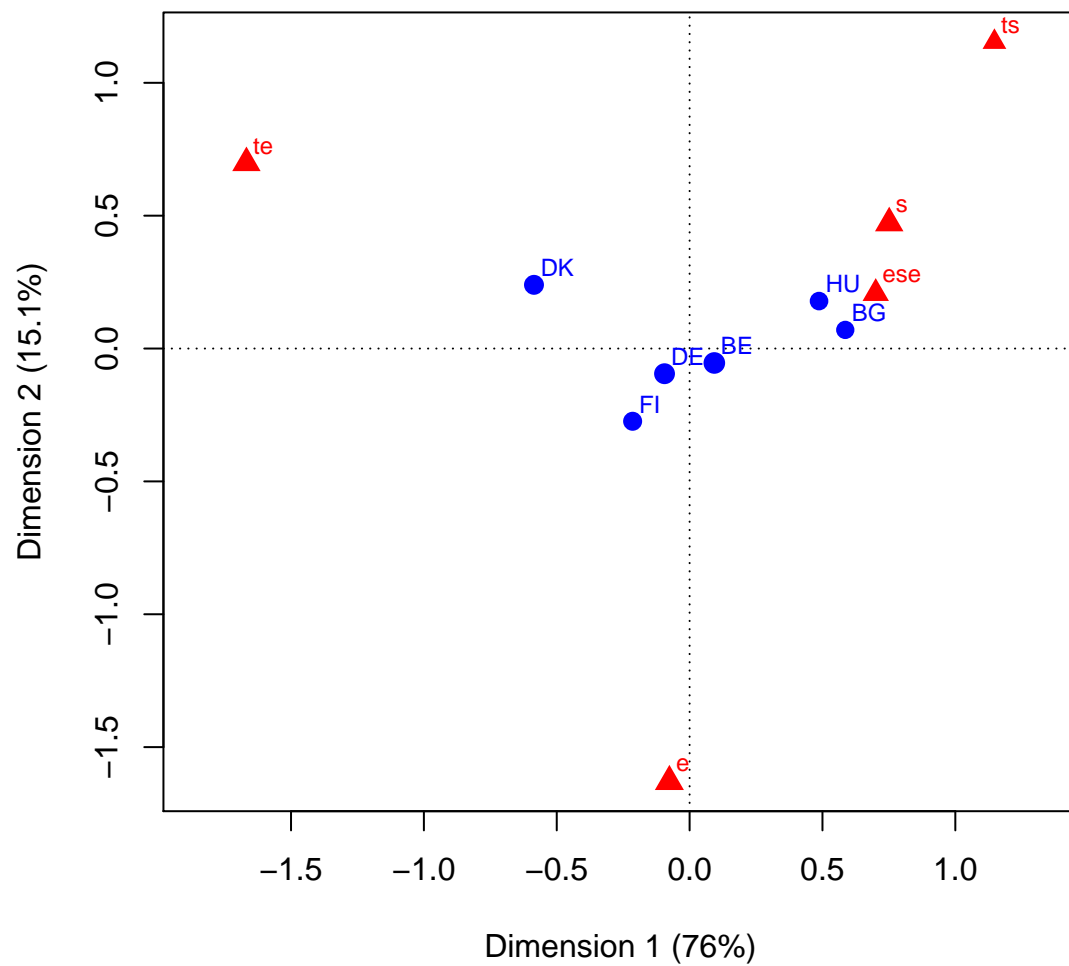
```
#str(apu2_df)
#apu2_df
#apu1_df
#apu1b_df
```

Ensimmäinen korrespondenssianalyysi

```
#simpleCA1 <- ca(~maa + V6,ISSP2012esim1.dat) suoritetaan ennen värikuvaa, tuloksia tarvitaan siinä!
#symmetrinen kartta - asp=1 - optio ei toimi?
plot(simpleCA1, map = "symmetric", mass = c(TRUE,TRUE))
```



```
#asymmetrinen kartta - rivit pc ja sarakkeet sc
plot(simpleCA1, map = "rowprincipal", mass = c(TRUE,TRUE))
```



```
#str(simpleCA1)
```

2.2 Korrespondenssianalyysin käsitteet

1. Profilit
2. Massat
3. Profiilien etäisyydet

EDIT: kaavaliitteessä (LaTeX) on kirjoitettu valmiiksi - en vielä lisää (25.8.18)

3 Tulkinnan perusteita

```
library(rgl)
library(ca)
library(haven)
library(dplyr)
library(knitr)
library(tidyverse)
library(lubridate)
library(rmarkdown)
library(ggplot2)
library(furniture)
library(likert)
library(scales) # G_1_2 - kuva
library(reshape2) # G_1_2 - kuva
```

Luvussa syvennetään esimerkin tulkinnan perusteita. Miksi symmetrinen kartta on yleensä paras vaihtoehto, siksi se oletusarvoisesti esitetäänkin. Milloin voi käyttää vaihtoehtoisia esitystapoja? *Ydinluku*.

Esimerkkiaineistossa tulee jo pohdittavaa, Guttman (arc, horseshoe) - efekti, ratkaisun dimensiot jne.

4 Yksinkertaisen korrespondenssianalyysin laajennuksia

Korrespondenssianalyysi sallii rivien tai sarakkeiden yhdistelyn tai “jakamisen”. Tämä onnistuu esimerkkiaineistossa lisäämällä rivejä eli jakamalla eri maiden vastauksia useampaan ryhmään.

Sen avulla voi myös tarkastella ja vertailla erilaisia ryhmien välisiä tai ryhmien sisäisiä (within groups - between groups) eroja hieman. Teknisesti yksinkertaista korrespondenssianalyysiä sovelletaan muokattuun matriisiin. Datamatriisi rakennetaan useammasta alimatriisista, joko “pinoamalla” osamatriiseja (stacked matrices) tai muodostamalla symmetrinen lohkomatriisi (ABBA).

Lisätään esimerkkidataan uusia muuttujia, vastaajan luokitelut ikä ja sukupuoli.

4.1 Lisämuuttujat: ikäluokka ja sukupuoli

Luodaan aineistoon ikä- ja sukupuolimuuttujat

4.2 ABBA

4.3 Päällekkäiset matriisit (stacked matrices)
