Yksinkertainen korrespondenssianalyysi - kahden luokittelumuuttujan 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 eteisyyksinllä origoon. Rivipisteiden suhteelliset etäisyydet, samoin sarakepisteiden, mutta ei muut.

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 tosiasiassa 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
```

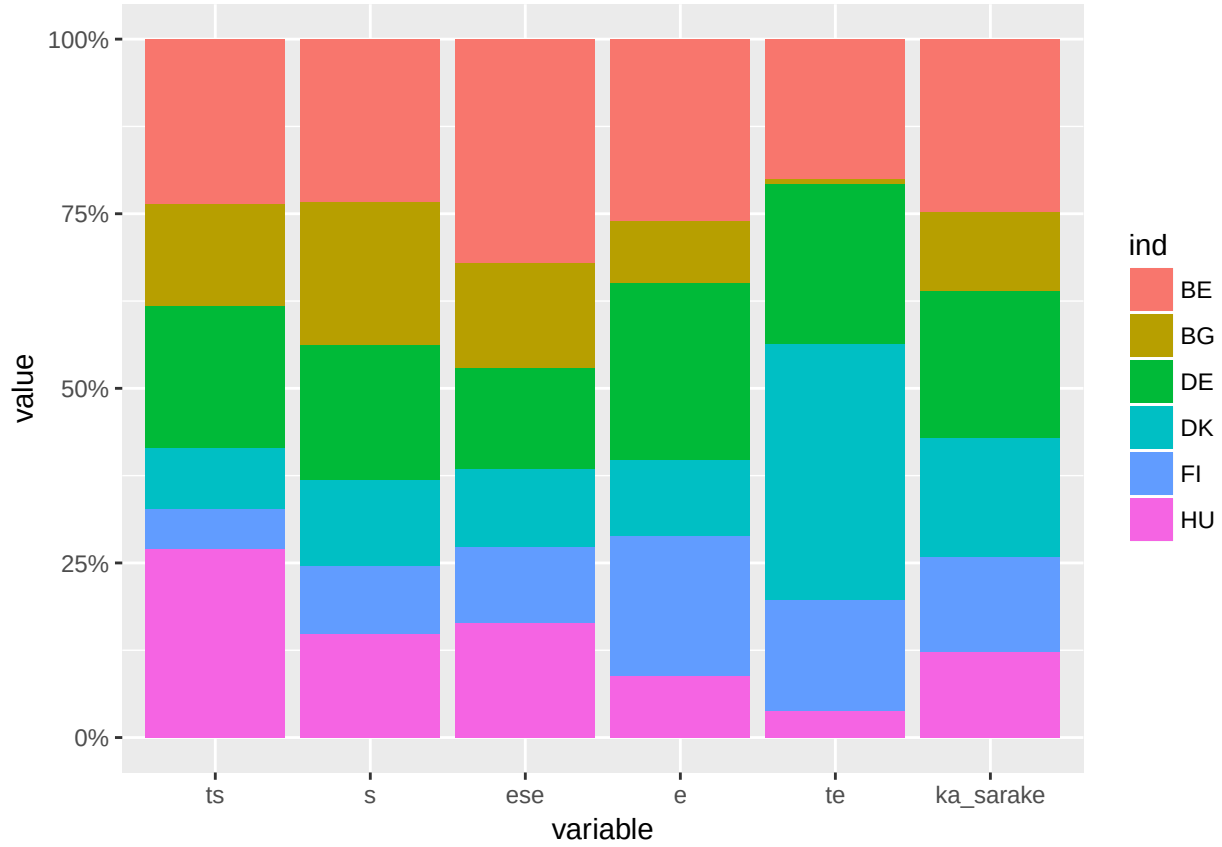
Taulukoissa on kuuden maan vastausten jakauma kysymykseen “Alle kouluikäinen lapsi todennäköisesti kärsii, jos hänen äitinsä käy työssä”. Taulukko on pieni, mutta havaintoja on melko paljon (N=8143). Alemman suhteellisten frekvenssien taulukon rivejä voi verrata toisiinsa ja alimpaan (“Total”“) keskimääräiseen 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. Näistähän riippuvuuden rakenteet näkee ilmankin, jos on tarpeeksi nokkela. 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
##https://stackoverflow.com/questions/9563368/create-stacked-barplot-where-each-stack-is-scaled-to-sum-
#
# 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())
```



Riviprofiilit

```
# riviprofiilit ja keskiarvorivi - aika väärin piirretty 30.4.2018
# kokeillaan vähän simppelimmin
apu2_df <- as.data.frame(apu1)
apu2_df <- rbind(apu2_df, ka_rivi = colSums(apu2_df))
apu2_df
```

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

```
#str(apu2_df)
#typeof(apu2_df) # what is it?
#class(apu2_df) # what is it? (sorry)
#storage.mode(apu2_df) # what is it? (very sorry)
#length(apu2_df) # how long is it? What about two dimensional objects?
#attributes(apu2_df)
apu2_perc <- apply(apu2_df,1,function(x){x/sum(x)})
apu2_perc
```

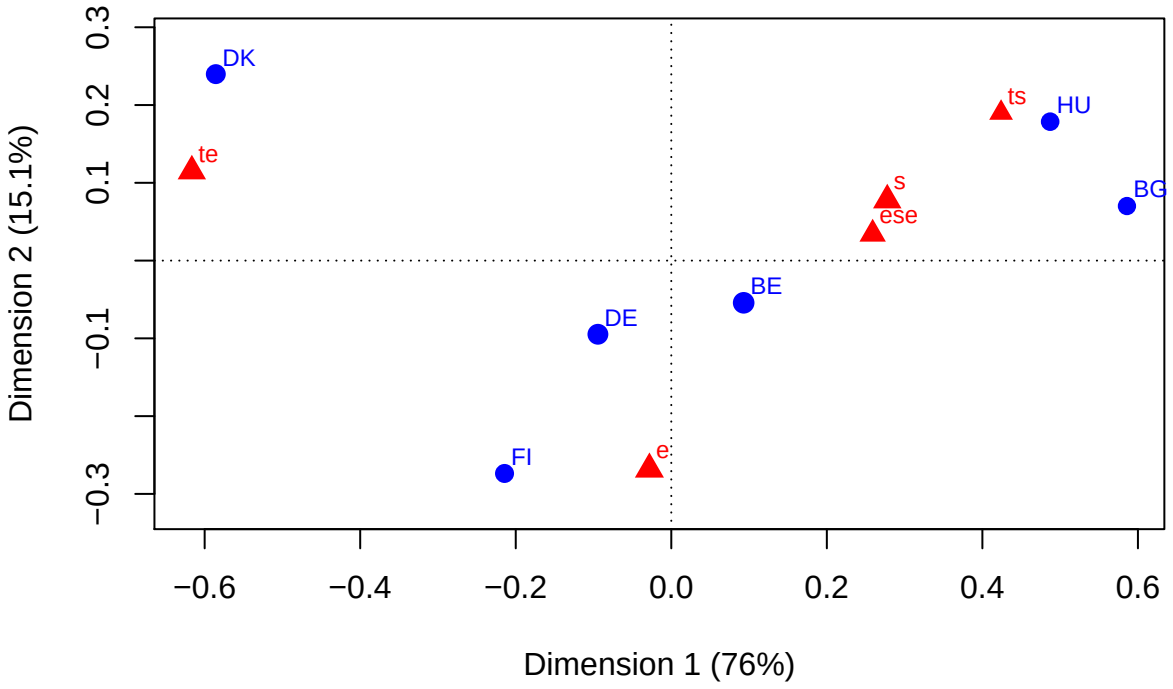
```
##             BE         BG         DE         DK         FI         HU
## ts  0.09488326 0.12812161 0.09626604 0.05043228 0.04234234 0.21965898
## s   0.22404372 0.42888165 0.21878646 0.17146974 0.16936937 0.28886660
## ese 0.21758569 0.22258415 0.11551925 0.10951009 0.13423423 0.22567703
## e   0.27421759 0.20629750 0.31388565 0.16714697 0.38108108 0.19057172
## te  0.18926975 0.01411509 0.25554259 0.50144092 0.27297297 0.07522568
##        ka_rivi
## ts  0.09947194
## s   0.23762741
## ese 0.16787425
## e   0.26096033
## te  0.23406607
```

```
##muokataan 'long data' - muotoon
#apu2b_df <- melt(cbind(apu2_df, ind = rownames(apu2_df)), id.vars = c('ind'))
#
#
#ggplot(apu2b_df, aes(x = variable, y = value, fill = ind)) +
#         geom_bar(position = "fill", stat ="identity") +
#         #coord_flip() +
#         scale_y_continuous(labels = percent_format())
```

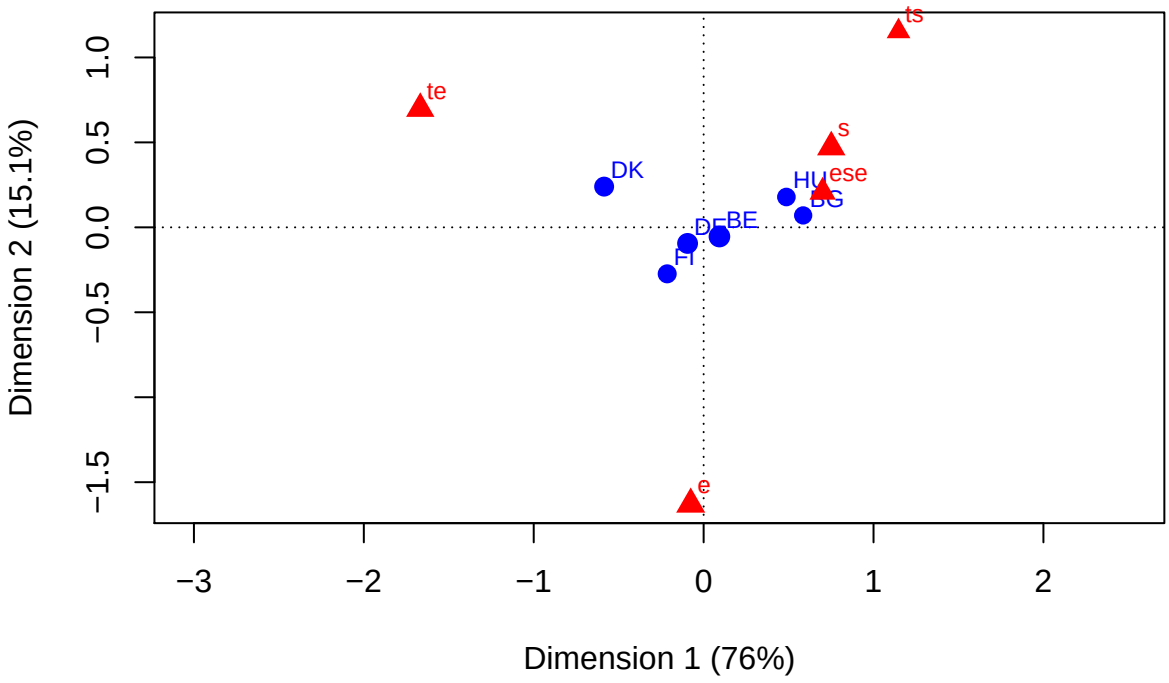
Ensimmäinen korrespondenssianalyysi

```
#simpleCA1 <- ca(~maa + V6,ISSP2012esim1.dat) suoritetaan ennen värikuvaa, tuloksia tarvitaan siinä!
#symmetrinen kartta - asp=1 - optio ei toimi? Tilapäinen fiksi alla (12.5.2018)

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)
#kuvasuhde saadaan tilapäisesti ratkaistua omalla tulostusikkunalla komentoriviltä X11() ja #tulostusko
```

Yritetään tuoda tähän pdf-muodossa kuvatiedosto, jossa kuvasuhde on oikea. Nämä toiminevat vain pdf-tulostuksessa.

[Alt text]1CAmap_sy.pdf

Ja toinen tapa

```
#include_graphics[width=8]{1CAmap_sy.pdf}
```

Korrespondenssianalyysin käsitteet

1. Profiilit
2. Massat
3. Profiilien etäisyydet

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