G Luku 1 Yksinkertainen korrespondenssianalyysi

Jussi Hirvonen

versio 1.5.7, tulostettu 2020-05-04

Sisällys

Versiot - vanha Galku - 5.6.2019 versio 1.5.1 Uusi Galku - 2.2.2020 versio 1.5.5, 4.2.2020 versio 1.5.6, 24.2.2020 versio 1.5.7

Siivotaan datan käsittelyn koodilohkot, kopiodaan mahdollisesti hyödylliset koodipätkät tiedostoon siivous1.R (30.1.2020).

Uudet datan luku- ja muunnosskriptit (treeni2-projektista), korjaillaan virheitä ja editoidaan koodia.(31.1.2020)

(2.2.20) Toimii johdattelevaan esimerkkiin asti, myös PDF-tulostus. Kuvien

otsikot vähän mitä sattuu, ja ´profiilikuviin asti maa-muuttujan järjestys “väärä”, ts. eri kuin vanhemmissa versoissa. Korjattu, lisättiin johdattelevan esimerkin dataan myös maakoodi jossa Saksan ja Belgian jako (V3).

(4.2.20) Versio 1.5.6 - Galku toimii loppuun asti, tarkistettava ja editoitava. Poistetaan tarpeetonta tekstiä, vanha koodi voi jäädä selvästi merkittynä.

(24.2.20) Versio 1.5.7. Pieniä ja isompiakin korjailuja, koodin siistimistä jne. (27.3.20) Muutetaan hieman karttojen koodilohkoja, html-tulosteessa kuvasuhde 1 mutta pdf-tulosteessa ei. (8.4.20)

HISTORIAA

6.8.2018 versio 1.0

Siistitään -> 12.8.2018 versio 1.05

Kommentit ja korjaukset -> 4.9.2018 versio 1.1

puuttuva riviprofiilikuva, siistimmät interaktiomuuttujien koodaukset, ensimmäinen “pinottu taulu” - analyysi -> 19.9.2018 versio 1.2

25.9.2018 siistitään datan käsittelyä; ei huomioida puuttuvan tiedon tarkempaa koodausta (read_spss - funktion user_na = TRUE asetus)

1.10.2018 Versio 1.3

Muutokset tarkemmin Readme.md - tiedostossa.

Uusi jakso yksinkertaisen CA:n laajennuksille, joissa otetaan analyysiin useampia muuttujia “pinoamalla” ja/tai yhdistämällä taulkoita. Tässä jaksossa otetaan myös käyttöön isompi aineisto (enemmän maita ja muuttujia). Siisti koodipätkä täydentävien muuttujien lisäämiseen.

3.10.2018 Versio 1.4

Siistitään pois turhat datan listaukset. Aineiston rajaaminen selkeäksi. Ensin kuusi maata, sitten 27 (Espanja pois). Valitaan myös muuttujat, jotta käsiteltävän datan listaukset ovat järkevämpiä. Aineistossa esim. Espanjan ja muutaman Unkarin poikkeavien vastausvaihtoehtojen vastaukset ovat omina muuttujina, ja niiden arvo muille havainnoille on NAP (Not applicaple). Samoin paljon maakohtaisia muuttujia, esim. koulutustaso. Mukaan otetaan vain kv-vertailuihin kelpaavat muuttujat, muutama sellainen on myös aineistoon rakennettu. Jätetään pois kaikki perhesuhteisiin liittyvät kysymykset (esim. kotitöiden jakaminen) ja taustatiedot (esim. rahankäyttö, puolison eri tiedot jne.), koska muuten jouduttaisiin miettimään miten näiden osalta käsitellään perheettömiä. Muutamia muuttujia otetaan mukaan (lasten lkm jne.).

8.10.2018

Datan valinta. Data-jaksossa aluksi, voi miettiä siirtääkö esimerkki-lukuun ja “pinotut taululut” - luvun alkuun kuvailut. Tavallaan siistiä, jos alussa lyhyesti.

10.10.2018

Maiden ja muuttujien valinta. TOPBOT halutaan mukaan, joten USA ja GB on jätettävä pois. Muuttuja on kuitenkin hankala, usealla maalla puuttuva tieto yli 10 prosentissa, ja muutamalla nolla tai ihan muutamia. Pohditaan aikanaan. **5.112.18**Puuttuvat tiedot ovat puuttuvia, ei voi mitään. Jos vähän ja selviä virheitä (ikä, sukupuoli), voidaan pudottaa havainnot.Muuten mukaan, periaatteessa.

Data-jaksosta siirretään aineiston laajentamisen yhteyteen laajemman muuttujajoukon deskriptiiviset tarkastelu. Taulukko muuttujakuvauksesta jää data-lukuun. **5.12.18** Puuttuneisuuden taulukointia on, mutta siisti NA-taulukko puuttu.

11.10.2018 Versio 1.4

- paperitulosteessa v1.3 kommentteja karttoihin ja ca:n numeerisiin tuloksiin, samoin muuttujalistauksiin.
- paperitulosteessa v1.4 samoin, ja puuttuneisuuden taulukointeja

11.10.2018 aloitetaan versio 1.5 - pieniä muutoksia ja kommentteja, aloitetaan uusi versio 1.51 5.12.2018

6.12.2018 1.5.1 - as_factor - funktio käyttöön; testaillaan miten toimii kun (a) user_na - arvoja ei lueta ja (b) puuttuvat ovat mukana.

Muistilista:

1. Taulukot ja kuvat luvusta 2. alkaen eivät ole "bookdown-muodossa". CA-tulokset on tulostettu siiteinä taulukoina Bookdown-demo - dokumentissa. Voi tulostaa myös ca-outputin. Ominaisarvojen taulukko keskeneräinen, samoin "scree plot" kuvana puuttuu.
2. Osa kuvista (esim. profiilikuva) pitää varmaan tulostaa pdf-muodossa ja ottaa capaper-dokkariin include_graphics - funktiolla.
3. Puuttuvia tai mahdollisesti lisättäviä taulukoita (nämä saa ca-funktion tuloksista suoraan)

- khii2 - etäisyydet riveille ja sarakkeille - on tulostettu ilman muotoiluja (11.10.18)
- massoilla painotetut khii2-etäisyyden keskiarvorivistä/sarakkeesta?

4. Kuvissa vielä hiottavaa, pdf-kuvia lisäilty img-hakemistoon.
5. Data-tiedostojen nimeäminen (27.12.18)

ISSP2012*.data - täysi aineisto

ISSP2012*jh1.data - valikoitu aineisto (maat, muuttujat)

ISSP2012*esim1.dat - muuttujien muunnoksia ja uusia muuttujia; analyyseissä käytettävä data, tarkenne dat.

6. kasitteet1.rmd - taulukko käsitteistä ja tärkeimmistä ISSP-dokumenteista

Historiaa (11.10.18)

Vanhoja kommentteja

- kirjastot/paketit ladataan jokaisessa Rmd-dokumentissa
- bib-formaatin viitetietokantaa tullaan kokeilemaan
- kuvasuhde (aspect ratio) edelleen epäselvä juttu! Mutta näyttää PDF-tulosteessa olevan ok.
- Datan käsittely ja hallinta +SPSS:n sallima kolme puuttuvan tiedon koodia saadaan mukaan read_spss-funktion (haven) parametrilla USER_NA = TRUE (mutta tarkistettava!) (25.4.18)
  - faktoreita ei ainakaan toistaiseksi muuteta ordinaaliasteikolle, CA ei tästä välitä
  - pidetään muuttujien ja tiedostojen nimeäminen selkeänä, tarkistetaan aika ajoin
- Taulukot: lisättiin riviprosentti- ja sarakeprosenttitaulut (25.4.18), kuva riviprofiileista puuttu vielä (15.5.2018)
- Datan esittelyssä on turhaa välitulostusta, ja samoin vähän muuallakin. Html on helpompi lukea, kun koodi on oletuksena piilossa
- PDF-tulosteessa koodi pääsääntöisesti näkyy toistaiseksi
- kokeiluja CA-karttojen tulostamiseen (a) suoraan koodilla ja (b) r-grafiikkaikkunasta tallennetun pdf-kuvan avlla. Paras toistaiseksi (a), jätin kokeilu näkyviin. Analyysit R:n grafiikkaikkunassa, jotta asp=1, ja tulkintaa varten voi tallentaa PDF-muodossa.
- rakenteeseen muutoksia (näkyvät sisällysluettelossa), ei erillistä teorialiitettä vaan sopivina annoksina. Lukuun 3 perusasiat, kaavat, määritelmät
- tehdään käsitetaulukko (kirjoittamista varten)
- 20.5.2018 (a) tulkita-osuuteen karttakuvia ja ca-tulokset (b) siistimpi taulukoiden tulostus löytyi (c) kaavaliite laajeni (dispo-haarassa)
- 23.5.2018 lisätään dataan toinen maa-muuttuja maa2, ikäluokkamuuttuja age_cat ja iän ja sukupuolen vuorovaikutusmuuttuja ga.
- 24.5.2018 lisättiin ca-kartta, jossa Saksan ja Belgian ositteet ja summarivit täydentävinä (passiivisina)

---

1 Data

edit 30.1.20 Siivotaan, luodaan faktori-muuttujat heti alussa koko datalle. Uusi G1_1_data_fct1.Rmd tekee muunnokset.

Historiaa

edit tässä luvussa on paljon siistittävää, mutta data on ok. (13.5.2018). **edit** capaper - dokumentissa parempi uusi jäsentely (4.9.2018) **edit** ISSP-datan perustietoa dokumentissa ISSP_data1.docx (4.9.2018) **edit 24.9.18** Poistettiin turhaa, uusi versio tiedostosta (G1_1_data1.Rmd -> G1_1_data2.Rmd).

1.1 Luvun 1 tavoitteet

Datan esittely ja kuvailut - tämä luku täysin uusiksi (24.9.18)

10.10.2018 maat ja muuttujat valittu.

1. Eksploratiivinen ja graafinen menetelmä tarvitseen aineiston, hankalaa esitellä jollain synteettisellä esimerkkiaineistolla. **edit** Eksp&graaf menetelmät määriteltävä johdantoluvussa. Esimerkkiaineistoja (synteettisiä kuten smoke, myös muita) on mm. ca - paketissa.
2. CA (ja MCA) sopivat isojen moniulotteisten ja mutkikkaiden aineistojen analyysiin, siksi iso aineisto. Samalla analyysiä voi laajentaa moneen suuntaan. **V** Benzecri:"kun data menee miljoonaan suuntaan".
3. Aineiston esittely, laajan kyselytutkimusaineiston tyypilliset ominaisuudet
4. Laadukkaan ja hyvin dokumentoidun aineiston edut
5. Huom! CA sopii ja sitä on käytetty myös hyvin toisen tyyppisiin aineistoihin (esim. ekologia ja biologia, arkeologia, kielen tutkimus)

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

luvun pitäisi olla mahdollisimman lyhyt (5.12.18)

Hieman historiaa datasta, sosiaalisesti määräytyneet sukupuoliroolit (gender) tutkimusaiheena neljässä ISSP:n kyselytutkimuksessa. *

Tärkeät linkit

Toimivat html-tulosteessa, PDFtiedostoissa saa toimimaan (vaati tarkat formatoinnit Rmd-koodissa).

www.issp.org, tutkimushankkeen historiaa. Löytyy myös bibliografia tutkimuksista, joissa aineistoja on käytetty.

www.gesis.org - tutkimuksen "sihteeristö", dokumentaatio ja datat.

data ja dokumentaatio (selattavissa): zacat.gesis.org

edit tässä järkevä viite ISSP - dataan 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 Alla myös suora linkki

Linkitys dokumentteihin on hankalaa

- monta portaalia, joista pääsee monien organisaationimien taakse
- tästä lyhyt selostus
- tärkeimmät linkit ISSP-tutkimuksen "kotisivu" ja selkeä **muuttujakuvaukset ja muut tiedot**

- käytännössä linkittäminen “syvälle” johonkin sivustoon tai www-palveluun ei ole järkevää, parempi antaa selkeät viitetiedot ja tiedot organisaatioista. Ne kyllä säilyvät, tai jäljille pääsee.

Edit Refworksiin on kerätty viitteitä, tässä pärjätään kolmen saitin osoitteilla. Voi laittaa taulukon tärkeimmistä dokumenteista, tarvittaessa liitteeksi (tiedostonimet ja kuvaus). Alla linkkejä jotka eivät näy PDF-tulosteessa, lisätty tekstinä.

Aineistot https://dbk.gesis.org/dbksearch/sdesc2.asp?no=5900&db=e2012 **toimii**

[Muuttujakuvaukset ja muut tiedot] (http://zacat.gesis.org/webview/index.jsp?object=http://zacat.gesis.org/obj/fStudy/ZA5900) http://zacat.gesis.org/webview/index.jsp?object=http://zacat.gesis.org/obj/fStudy/ZA5900 **OK - täältä löytyy oikeastaan kaikki!** Suomenkielinen lomake (ZA5900_q_fi-fi.pdf) löytyy helpoiten täältä, samoin muu dokumnetaati tiedostoina. Veppisivuilla kerrotaan, mitä ne dokumentit ovat.

Data ja dokumentit **vie vain aineiston dokumentoinnin etusivulle** https://dbk.gesis.org/dbksearch/sdesc2.asp?no=5900&db=e

Käyttöehdot: **GESIS-palvelun datan yleiset käyttöehdot, viittauskäytännöt**

Havaintojen lukumäärät voi tarkistaa täältä http://zacat.gesis.org/webview/index.jsp?object=http://zacat.gesis.org/obj/fStudy/ZA5900. **Dokumentointisivusto/katalogi, jossa helppo navigoida** zacat.gesis.org.

Dokumentointi on hyvin tarkka, tiedot löytyvät haastattelumenetelmista (parerilomake, tietokoneavusteinen haastattelu, jne), maakohtaisten taustamuuttujien harmonisoinnista maittain, otantamenetelmistä jne. Esittelen vain aineiston tärkeimmät rajaukset. MOnitorointiraportti kertoo puuttuneisuuden määrän, otantamenetlmät jne maittain. “Code book” kertoo muuttujien määritelmät sekä yhteisille että maakohtaisille muuttujille. Kaikista muuttujista on taulukko maittain. Lisätään vielä raportti kyselylomakkeen laadinnasta ja linkki Yhteiskunnalliseen tietoarkistoon.

```
issp_docname <- c("Variable Report", "Study Monitoring Report","Basic Questionnaire",
                  "Contents of ISSP 2012 module", "Questionnaire Development")
issp_docdesc <- c("Perusdokumentti, muuttujien kuvaukset ja taulukot",
                  "tiedokeruun toteutus eri maissa",
                  "Maittain sovellettava kyselylomake", "substanssikysymykset taulukkona",
                  "kyselylomakkeen laatiminen")
issp_docfile <- c("ZA5900_cdb.pdf", "ZA5900_mr.pdf", "ZA5900_bq.pdf","ZA5900_overview.pdf",
                  "ssoar-2014-scholz_et_al-ISSP_2012_Family_and_Changing.pdf")


col_isspdocs <- c("dokumentti","sisältö","tiedosto")
```

```
# colnames(ISSPdocsT.df) <- col_isspdocs
# Vanha df-koodi
# ISSPdocsT.df <- data_frame(issp_docname, issp_docdesc, issp_docfile)
# knitr::kable(ISSPdocsT.df, booktab=TRUE)

# varoituksia data_framen käytöstä, toimisiko tibble()? (21.2.20)
ISSPdocsT.tbl <- tibble(issp_docname, issp_docdesc, issp_docfile)
colnames(ISSPdocsT.tbl) <- col_isspdocs
knitr::kable(ISSPdocsT.tbl, booktab = TRUE)
```

dokumentti	sisältö	tiedosto
Variable Report	Perusdokumentti, muuttujien kuvaukset ja taulukot	ZA5900_cdb.pdf
Study Monitoring Report	tiedokeruun toteutus eri maissa	ZA5900_mr.pdf
Basic Questionnaire	Maittain sovellettava kyselylomake	ZA5900_bq.pdf
Contents of ISSP 2012 module	substanssikysymykset taulukkona	ZA5900_overview.pdf
Questionnaire Development	kyselylomakkeen laatiminen	ssoar-2014-scholz_et_

1.3 Substanssimuuttujat, taustamuuttujat, muut

zxy capaper - dokumentissa uusi jäsentely (4.9.2018)

zxy Aineiston luonne: maakohtaisesti eri tavoin kerätty data, jossa pyritään yhtenäisiin käytäntöihin ja tietosisältöihin. Silti myös substanssikysymyksissä eroja, isoja ja pienempiä. Näin vain on, en pohdi miksi. Ei ole mitenkään ainutlaatuista. Aineiston editoinnissa ja tiedonkeruun suunnittelussa on nähty paljon vaivaa vertailukelpoisuuden vuoksi. Tästä esimerkkejä, esim. “mitä puoluetta äänestit”.

zxy yksi kappale: Aineitoa on harmonisoitu, kysymyksiä hiottu, vertailukelpoisuuteen on pontevasti pyritty. Silti eroja löytyy, osa ymmärrettäviä (lisäkysymykset jne) ja osa ei (Espanja!). Tällaista on kansainvälisen kyselytutkimuksen data.

Paremipi muotoilu: Varsinaiset substanssimuuttujat eli kyselylomakkeet on koitettu hioa mahdollisimman yhdenmukaisiksi. Silti pieniä eroja löytyy, ja isojakin (Espanja on pudottanut neutraalin “en samaa enkä eri mieltä” - vaihtoehdon pois, ja Unkarissakin on muutamat vastausvaihtoehdot valittu omalla tyylillä). Taustamuuttujissa on pyritty samaan, ja aineistoon on myös rakennettu kansainvälisesti vertailukelpoisia muuttujia kansallisesti kerätyistä tiedoista. Näitä ovat erityisesti tuloihin liittyvät tiedot, ja mone muutkin. Muuttujat jakautuvat substanssi- ja taustamuuttujiin, ja taustamuuttujista monet tiedot on kerätty kansallisiin ainiestossa maan kirjantunnisteella alkaviin muuttujiin.

zxy HUOM! Dataa ei ole kerätty vain kansainvälisiin vertailuhihin! Sitä voi ja ehkä pitäisikin analysoida maa kerrallaan, ja vertailla näitä tuloksia. (#**V** Blasiuksen artikkeli, jossa arvioidaan yhden ISSP-tutkimuksen vertailukelpoisuutta.

Kysymykset eivät kovin hyvin näytä toimivan samalla tavalla eri maissa.)

1.4 Aineiston rajaaminen

1. Eurooppa ja samankaltaiset maat (25)

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

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 **(28)** - Espanja, Iso-Britannia, USA pois -> **25 maata (11.10.18)**

Espanja jätettiin pois, koska siellä kysymyksissä jätettiin pois neutraali vaihtoehto ("en puolesta enkä vastaan / en osaa sanoa"). USA ja GB pois koska kiinnostava TOPBOT-muuttuja puuttuu (puuttui 11.10.18, sittemmin USA:n ainestoa on täydennetty).

(24.2.20) Aineistosta valittiin ensin joukko suhteellisen samankaltaisia kehittyneitä teollisuusmaita. Sitten valittiin osa kysymyksistä, ja vielä suppeampi valikoima kiinnostavia taustamuuttujia. Muutama maa pudotettiin pois tämän valinnan jälkeen.

3. kaikki havainnot, joissa on puuttuvia tietoja.

Johdattelevassa esimerkissä on kolme muuttujaa, ei ongelma, aika vähän puuttuvia.

Isomman 25 aineiston osalta tarkistetaan, mitä "listwise deletion" saa aikaan. Aineisto pienenee nopeasti, ja vaikeasti hahmotettavalla tavalla. Tämä erävastauskato ei ole tutkielman ydinaihe, mutta laajemman aineiston käytössä se täytyy ottaa huomioon. Yksikkövastauskatoa ei käsitellä, tutkimuksen toteutuksen raporteissa on kerrottu tarkemmin miten kyselyn toteuttajat ovat tämän huomioineet. 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 maakohtaiset painomuuttujat, mutta ei painoja maiden vertailuun. Vastausprosentit (response rate) vaihtelevat maittain, kts. monitoring report. (**edit** toistoa! 24.2.20)

CA:n eräs etu on se, että muuttujien oletetaan olevan luokitteluasteikon (nominaaliasteikon) muuttujia, ja puuttuva havainto on yksi luokka lisää. Puuttuvat havainnot otetaan mukaan laajemmassa aineistossa myös siksi, että CA ja MCA edellyttävt yleensä useamman muuttujan analyyseissä sitä. Jokaisen kahden muuttujan parittaisen ristiintaulukoinnin reunajakaumien pitää olla samoja.

4. Datan hallinta - reproducible research- periaate

edit 24.2.20 Vanhoja perusideoita

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

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

zxy 3.10.18

R-spesifiä: R-koodissa tarkemmin, kaikki yksityiskohdat.

Kun SPSS-tiedosto luetaan R:n data frame - tiedostoksi, mukana tulee myös metadata. Uusien muuttujien luonnissa tai data-formaatin vaihtuessa (esim. matriisiksi, taulukoksi jne) metadata katoaa. Siksi muuttujien tyyppimuunnokset (yleensä faktorointi) tallennetaan uusiksi muuttujiksi, metatieto säilyy vanhassa muuttujassa.

Helposti toistettava tutkimus: polku alkuperäisestä datasta analyysien dataan selkeä (ja lyhyt jos mahdollista).

Puuttuva tieto voidaan koodata monella tavalla (ei halua vastata jne), ja SPSS (datan jakelutiedosto) sallii kolme koodia puuttuville tiedoille. Ne voi lukea R-dataan, mutta puuttuneisuutta ei tässä työssä tutkita sen tarkemmin. Detaljit R-koodissa (haven-paketin read_spss-funktion user_na -optio, ei käytetä tässä).

Tiedostonimistä (10.10.18, 30.1.20, 11.2.20)

ISSP2012.*data - täysi aineisto, luetaan SPSS-tiedostosta ISSP2012*jh1.data - valittu osa aineistosta (maat, muuttujat) ISSP2012*.jh1.dat - valittu osa aineistosta, luotu uusia muuttujia ja muunnettu muuttujia. Alkuperäiset muuttujat säilytetään, voi aina tarkistaa ja verrata. ISSP2012esim1, 2 jne, tarkenne .dat rajattuja aineistoja joissa uusia muuttujia ja muuttujien nimiä. Näitä luodaan analyysin eri vaiheissa.

zxy R-koodiin jätetään myös tarkistuksia yms. joita ei raportoida tässä, samoin niiden tuloksia. Voiko R-koodi olla fingelskaa? Olkoon toistaiseksi.

DATA RAJAAMISTA - maat(5.10.2018)

```
# Aineiston rajaamisen kolme vaihetta (10.2018)
#
# TIEDOSTOJEN NIMEÄMINEN
#
# R-datatiedostot .data - tarkenteella ovat osajoukkoja koko ISSP-datasta ISSP2012.data
# R-datatiedostot .dat - tarkenteella: mukana alkuperäisten muuttujien muunnoksia
# (yleensä as_factor), alkuperäisissä muuttujissa mukana SPSS-tiedoston metadata.
#
# Luokittelumuuttujan tyyppi on datan lukemisen jälkeen yleensä merkkijono (char)
# ja haven_labelled.
#
# Muutetaan R-datassa ordinaali- tai  nominaaliasteikon muuttujat haven-paketin
# as_factor - funktiolla faktoreiksi. R:n faktorityypin muuttujille voidaan tarvittaessa
```

```
# määritellä järjestys, toistaiseksi niin ei tehdä (25.9.2018).
#
# Muunnetun muuttujan rinnalla säilytetään SPSS-tiedostosta luettu muuttja, metatiedot säilg
# alkuperäisessä.
#
# R-datatiedostot joiden nimen loppuosa on muotoa *esim1.dat: käytetään analyyseissä
#
# 1. VALITAAN MAAT (25) -> ISSP2012jh1a.data. Muuttujat koodilohkossa datasel_vars1
#
# kolme maa-muuttujaa datassa. V3 erottelee joidenkin maiden alueita, V4 on koko
# maan koodi ja C_ALPHAN on maan kaksimerkkinen tunnus.
#
# 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 ainestoa täydennettiin, koska siinä oli puutteita. Jako ei siis ole oleellin
# mutta muuut ovat. Tähdellä merkityt maat valitaan johdattelevaan esimerkkiin.
#
# Maat (25)
#
# 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
```

```r
# 578 NO-Norway
# 616 PL-Poland
# 620 PT-Portugal
# 643 RU-Russia
# 703 SK-Slovakia
# 705 SI-Slovenia
# 752 SE-Sweden
# 756 CH-Switzerland
# 826 GB-Great Britain and/or United Kingdom - jätetään pois jotta saadaan TOPBOT
#                          -muuttuja mukaan (top-bottom self-placement) .(9.10.18)
# 840 US-United States - jätetään pois, jotta saadaan TOPBOT-muuttuja mukaan.(10.10.18)
#
# 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
#
# Unkari (348) toistaiseksi mukana, mutta joissain kysymyksissä myös Unkarilla on
# poikkeavia vastausvaihtoehtoja(HU_V18, HU_V19,HU_V20). Jos näitä muuttujia käytetään,
# Unkari on parempi jättää pois.
#
#
# (25.4.2018) user_na
# haven-paketin read_spss - funktiolla voi r-tiedostoon lukea myös SPSS:n sallimat kolme
# (yleensä 7, 8, 9) tarkempaa koodia puuttuvalle tiedolle.
# "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
#

ISSP2012jh.data <- read_spss("data/ZA5900_v4-0-0.sav") #luetaan alkuperäinen data R- dataks

#str(ISSP2012jh.data)

incl_countries25 <- c(36, 40, 56,100, 124, 191, 203, 208, 246, 250, 276, 348, 352,
                      372, 428, 440, 528, 578, 616, 620, 643, 703, 705, 752, 756)

#str(ISSP2012jh.data)
#str(ISSP2012jh.data) #61754 obs. of  420 variables - kaikki

ISSP2012jh1a.data <- filter(ISSP2012jh.data, V4 %in% incl_countries25)
```

```
#head(ISSP2012jh1a.data)
#str(ISSP2012jh1a.data) #34271 obs. of  420 variables, Espanja ja Iso-Britannia
#                       pois (9.10.2018)
# str(ISSP2012jh1a.data) # 32969 obs. of  420 variable, Espanja Iso-Britannia,
#                        USA pois (10.10.2018)
#
# names() # muuttujen nimet
# Maakohtaiset muuttujat (kun on poikettu ISSP2012 - vastausvaihtoehdoista tms.)
# on aineistossa eroteltu maatunnus-etuliitteellä (esimerkiksi ES_V7).
# Demografisissa ja muissa taustamuuttujissa suuri osa tiedoista on kerätty maa-
# kohtaisilla lomakkeilla. Vertailukelpoiset muuttujat on konstruoitu niistä.
# Muuttujia on 420, vain osa yhteisiä kaikille maille.
```

DATAN RAJAAMISTA - MUUTTUJAT (5.10.2018)

SPSS-tiedostosta saadaan luettua haven-paketin read_spss-funktiolla paljon metatietoja.

```
# 2. VALITAAN MUUTTUJAT  -> ISSP2012jh1b.data. Maat valittu koodilohkossa datasel_country1
#
#
# Muuttujat on luokiteltu dokumentissa ZA5900_overview.pdf
# https://zacat.gesis.org/webview/index.jsp?object=http://zacat.gesis.org/obj/fStudy/ZA5900
# Study Description -> Other Study Description -> Related Materials
#
#

# METADATA

metavars1 <- c("V1", "V2", "DOI")

#MAA - maakoodit ja maan kahden merkin tunnus

countryvars1 <- c("V3","V4","C_ALPHAN")

# SUBSTANSSIMUUTTUJAT - Attitudes towards family and gender roles (9)
#
# Yhdeksän kysymystä (lyhennetyt versiot, englanniksi), vastausvaihtoehdot Q1-Q2
#
# 1 = täysin samaa mieltä, 2 = samaa mieltä, 3 = ei samaa eikä eri mieltä,
# 4 = eri mieltä, 5 = täysin eri mieltä
#
# Q1a Working mother can have warm relation with child
# Q1b Pre-school child suffers through working mother
# Q1c Family life suffers through working mother
# Q1d Women's preference: home and children
```

```r
# Q1e Being housewife is satisfying
#
# Q2a Both should contribute to household income
# Q2b Men's job is earn money, women's job household
#
# Q3a Should women work: Child under school age
# Q3b Should women work: Youngest kid at school
# 1= kokopäivätyö, 2 = osa-aikatyö, 3 = pysyä kotona, 8 = en osaa sanoa (can't choose), 9 =
#
# Kysymysten Q3a ja Q3b eos-vastaus ei ole sama kuin "en samaa enkä eri  mieltä" (ns. neutr
# vaihtoehto), mutta kieltäytymisiä jne. (koodi 9) on aika vähän. Kolmessa
# maassa ne on yhdistety:
# (8 Can't choose, CA:can't choose+no answer, KR:don't know+refused, NL:don't know).
# Kun SPSS-tiedostosta ei ole tuotu puuttuvan tiedon tarkempaa luokittelua,
# erottelua ei voi tehdä.
#
#
#

substvars1 <- c("V5","V6","V7","V8","V9","V10","V11","V12","V13") # 9 muuttujaa

# Nämä yhteiset muuttujat pois (maaspesifien muuttujien lisäksi) :
#
# "V14","V15","V16",  "V17","V18","HU_V18","V19","HU_V19","V20","HU_V20","V21",
# "V28","V29","V30","V31","V32","V33",# "V34", "V35", "V36", "V37", "V38", "V39",
# "V40", "V41", "V42", "V43", "V44", "V45", "V46", "V47", "V48", "V49", "V50",
# "V51", "V52", "V53", "V54", "V55", "V56", "V57", "V58", "V59", "V60", "V61",
# "V62", "V63", "V64", "V65", "V65a","V66", "V67"
#
#
# DEMOGRAFISET JA MUUT TAUSTAMUUTTUJAT (8)
#
# AGE, SEX
#
# DEGREE - Highest completed degree of education: Categories for international comparison.
# Slightly re-arranged subset of ISCED-97
#
# 0 No formal education
# 1 Primary school (elementary school)
# 2 Lower secondary (secondary completed does not allow entry to university: obligatory sch
# 3 Upper secondary (programs that allow entry to university or programs that allow to entr
#   other ISCED level 3 programs - designed to prepare students for direct entry into the l
# 4 Post secondary, non-tertiary (other upper secondary programs toward labour market or te
# 5 Lower level tertiary, first stage (also technical schools at a tertiary level)
# 6 Upper level tertiary (Master, Dr.)
```

```
# 9 No answer, CH: don't know
# Yhdistelyt?
#
# MAINSTAT - main status: Which of the following best describes your current situation?
#
# 1 In paid work
# 2 Unemployed and looking for a job, HR: incl never had a job
# 3 In education
# 4 Apprentice or trainee
# 5 Permanently sick or disabled
# 6 Retired
# 7 Domestic work
# 8 In compulsory military service or community service
# 9 Other
# 99 No answer
# Armeijassa tai yhdyskuntapalvelussa muutamia, muutamissa maissa.Kategoriassa 9
# on hieman väkeä. Yhdistetään 8 ja 9. Huom! Esim Puolassa ei yhtään eläkeläistä
# eikä kategoriaa 9, Saksassa ei ketään kategoriassa 9.
#
# TOPBOT - Top-Bottom self-placement (10 pt scale)
#
# "In our society, there are groups which tend to be towards the top and groups
# which tend to be towards the bottom. Below is a scale that runs
# from the top to the bottom. Where would you put yourself on this scale?"
# Eri maissa hieman erilaisia kysymyksiä.
#
# HHCHILDR - How many children in household: children between [school age] and
# 17 years of age
#
# 0 No children
# 1 One child
# 2 2 children
# 21 21 children
# 96 NAP (Code 0 in HOMPOP)
# 97 Refused
# 99 No answer
#
# Voisi koodata dummymuuttujaksi lapsia (1) - ei lapsia (0).
# Ranskan datassa on erittäin iso osa puuttuvia tietoja ( "99"", n. 20 %), myös
# Austarlialla aika paljon. Sama tilanne myös muissa perheen kokoon liittyvissä
# kysymyksissä.
#
# MARITAL - Legal partnership status
#
# What is your current legal marital status?
```

```
# The aim of this variable is to measure the current 'legal' marital status '.
# PARTLIV - muuttujassa on 'de facto' - tilanteen tieto parisuhteesta
#
# 1 Married
# 2 Civil partnership
# 3 Separated from spouse/ civil partner (still legally married/ still legally
#   in a civil partnership)
# 4 Divorced from spouse/ legally separated from civil partner
# 5 Widowed/ civil partner died
# 6 Never married/ never in a civil partnership, single
# 7 Refused
# 8 Don't know
# 9 No answer
#
# URBRURAL - Place of living: urban - rural
#
# 1 A big city
# 2 The suburbs or outskirts of a big city
# 3 A town or a small city
# 4 A country village
# 5 A farm or home in the country
# 7 Other answer
# 9 No answer
# 1 ja 2 vaihtelevat aika paljon maittain, parempi laskea yhteen. Unkarista puuttuu
# jostain syystä kokonaan vaihtoehto 5.  Vaihotehdon 7 on valinnut vain 4 vastaajaa Ranskas
# Yhdistetään 1 ja 2 = city, 3 = town, rural= 4, 5, 7
#

bgvars1 <- c( "SEX","AGE","DEGREE", "MAINSTAT", "TOPBOT", "HHCHILDR", "MARITAL", "URBRURAL")

#Valitaan muuttujat

jhvars1 <- c(metavars1,countryvars1, substvars1,bgvars1)

#jhvars1
ISSP2012jh1b.data <- select(ISSP2012jh1a.data, all_of(jhvars1))

# laaja aineisto - mukana havainnot joissa puuttuvia tietoja
# hauska detalji URBRURAL - muuttujan metatiedoissa viite jonkun työaseman hakemistoon
# str(ISSP2012jh1b.data) #32969 obs. of  23 variables
#
# SUBSTANSSIMUUTTUJAT
#
# $ V5      : 'haven_labelled' num  5 1 2 2 1 NA 2 4 2 2 ...
#  ..- attr(*, "label")= chr "Q1a Working mom: warm relationship with children as a not work
```

```
#  ..- attr(*, "labels")= Named num  0 1 2 3 4 5 8 9
#
# ISSP2012jh1b.data$V5 näyttää tarkemmin rakenteen
#
# glimpse(ISSP2012jh1b.data)
# str(ISSP2012jh1b.data) # 32969 obs. of  23 variables

# Poistetaan havainnot, joissa ikä (AGE) tai sukupuolitieto puuttuu (5.7.2019)

ISSP2012jh1c.data <- filter(ISSP2012jh1b.data, (!is.na(SEX) & !is.na(AGE)))

str(ISSP2012jh1c.data) # 32823 obs. of  23 variables, 32969-32823 = 146
```

```
## tibble [32,823 x 23] (S3: tbl_df/tbl/data.frame)
##  $ V1      : 'haven_labelled' num [1:32823] 5900 5900 5900 5900 5900 5900 5900 5900 5900
##   ..- attr(*, "label")= chr "GESIS Data Archive Study Number"
##   ..- attr(*, "labels")= Named num 5900
##   .. ..- attr(*, "names")= chr "GESIS Data Archive Study Number ZA5900"
##  $ V2      : chr [1:32823] "4.0.0 (2016-11-23)" "4.0.0 (2016-11-23)" "4.0.0 (2016-11-23)"
##   ..- attr(*, "label")= chr "GESIS Archive Version"
##   ..- attr(*, "format.spss")= chr "A25"
##   ..- attr(*, "display_width")= int 26
##  $ DOI     : chr [1:32823] "doi:10.4232/1.12661" "doi:10.4232/1.12661" "doi:10.4232/1.126
##   ..- attr(*, "label")= chr "Digital Object Identifier"
##   ..- attr(*, "format.spss")= chr "A50"
##   ..- attr(*, "display_width")= int 26
##  $ V3      : 'haven_labelled' num [1:32823] 36 36 36 36 36 36 36 36 36 36 ...
##   ..- attr(*, "label")= chr "Country/ Sample ISO 3166 Code (see V4 for codes for whole na
##   ..- attr(*, "labels")= Named num [1:45] 32 36 40 100 124 152 156 158 191 203 ...
##   .. ..- attr(*, "names")= chr [1:45] "AR-Argentina" "AU-Australia" "AT-Austria" "BG-Bulg
##  $ V4      : 'haven_labelled' num [1:32823] 36 36 36 36 36 36 36 36 36 36 ...
##   ..- attr(*, "label")= chr "Country ISO 3166 Code (see V3 for codes for the sample)"
##   ..- attr(*, "labels")= Named num [1:41] 32 36 40 56 100 124 152 156 158 191 ...
##   .. ..- attr(*, "names")= chr [1:41] "AR-Argentina" "AU-Australia" "AT-Austria" "BE-Belg
##  $ C_ALPHAN: chr [1:32823] "AU" "AU" "AU" "AU" ...
##   ..- attr(*, "label")= chr "Country Prefix ISO 3166 Code - alphanumeric"
##   ..- attr(*, "format.spss")= chr "A20"
##   ..- attr(*, "display_width")= int 22
##  $ V5      : 'haven_labelled' num [1:32823] 5 1 2 2 1 NA 2 4 2 2 ...
##   ..- attr(*, "label")= chr "Q1a Working mom: warm relationship with children as a not wo
##   ..- 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 no
##  $ V6      : 'haven_labelled' num [1:32823] 1 5 4 4 4 NA 4 3 4 3 ...
##   ..- attr(*, "label")= chr "Q1b Working mom: Preschool child is likely to suffer"
##   ..- 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 no
##  $ V7      : 'haven_labelled' num [1:32823] 3 5 2 4 4 NA 4 2 4 2 ...
##   ..- attr(*, "label")= chr "Q1c Working woman: Family life suffers when woman has full-t
##   ..- 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 no
##  $ V8      : 'haven_labelled' num [1:32823] 3 5 5 2 4 NA 4 5 4 5 ...
##   ..- attr(*, "label")= chr "Q1d Working woman: What women really want is home and kids"
##   ..- 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 no
##  $ V9      : 'haven_labelled' num [1:32823] 3 1 2 3 4 NA 2 4 4 1 ...
##   ..- attr(*, "label")= chr "Q1e Working woman: Being housewife is as fulfilling as worki
##   ..- 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 no
##  $ V10     : 'haven_labelled' num [1:32823] 1 3 4 2 2 NA 2 5 2 1 ...
##   ..- attr(*, "label")= chr "Q2a Both should contribute to household income"
##   ..- 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 no
##  $ V11     : 'haven_labelled' num [1:32823] 3 5 4 4 4 NA 2 5 4 1 ...
##   ..- attr(*, "label")= chr "Q2b Men's job earn money, women's job look after home"
##   ..- 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 no
##  $ V12     : 'haven_labelled' num [1:32823] 3 NA NA 2 2 NA 2 NA 2 2 ...
##   ..- attr(*, "label")= chr "Q3a Should women work: Child under school age"
##   ..- attr(*, "labels")= Named num [1:6] 1 2 3 6 8 9
##   .. ..- attr(*, "names")= chr [1:6] "Work full-time" "Work part-time" "Stay at home" "TW
##  $ V13     : 'haven_labelled' num [1:32823] 2 NA 2 1 2 NA 2 NA 2 2 ...
##   ..- attr(*, "label")= chr "Q3b Should women work: Youngest kid at school"
##   ..- attr(*, "labels")= Named num [1:6] 1 2 3 6 8 9
##   .. ..- attr(*, "names")= chr [1:6] "Work full-time" "Work part-time" "Stay at home" "TW
##  $ SEX     : 'haven_labelled' num [1:32823] 1 2 2 2 2 1 2 1 2 2 ...
##   ..- attr(*, "label")= chr "Sex of Respondent"
##   ..- attr(*, "labels")= Named num [1:3] 1 2 9
##   .. ..- attr(*, "names")= chr [1:3] "Male" "Female" "No answer"
##  $ AGE     : 'haven_labelled' num [1:32823] 58 59 40 20 72 68 64 57 45 71 ...
##   ..- attr(*, "label")= chr "Age of respondent"
##   ..- attr(*, "labels")= Named num [1:6] 15 16 17 18 102 999
##   .. ..- attr(*, "names")= chr [1:6] "15 years" "16 years" "17 years" "18 years" ...
##  $ DEGREE  : 'haven_labelled' num [1:32823] 2 5 5 3 2 NA NA 6 5 6 ...
##   ..- attr(*, "label")= chr "Highest completed degree of education: Categories for interv
##   ..- attr(*, "labels")= Named num [1:8] 0 1 2 3 4 5 6 9
##   .. ..- attr(*, "names")= chr [1:8] "No formal education" "Primary school (elementary sc
##  $ MAINSTAT: 'haven_labelled' num [1:32823] 6 6 3 1 6 5 6 2 1 5 ...
##   ..- attr(*, "label")= chr "Main status"
##   ..- attr(*, "labels")= Named num [1:10] 1 2 3 4 5 6 7 8 9 99
##   .. ..- attr(*, "names")= chr [1:10] "In paid work" "Unemployed and looking for a job, P
##  $ TOPBOT  : 'haven_labelled' num [1:32823] 3 7 8 NA 7 2 7 NA 10 6 ...
```

```
##   ..- attr(*, "label")= chr "Top-Bottom self-placement"
##   ..- attr(*, "labels")= Named num [1:14] 0 1 2 3 4 5 6 7 8 9 ...
##   .. ..- attr(*, "names")= chr [1:14] "Not available: GB,US" "Lowest, Bottom, 01" "02" "0
##  $ HHCHILDR: 'haven_labelled' num [1:32823] NA NA 3 1 0 NA 0 0 1 NA ...
##   ..- attr(*, "label")= chr "How many children in household: children between [school age
##   ..- attr(*, "labels")= Named num [1:7] 0 1 2 21 96 97 99
##   .. ..- attr(*, "names")= chr [1:7] "No children" "One child" "2 children" "21 children"
##  $ MARITAL : 'haven_labelled' num [1:32823] 6 1 1 6 1 6 1 1 1 NA ...
##   ..- attr(*, "label")= chr "Legal partnership status"
##   ..- attr(*, "labels")= Named num [1:9] 1 2 3 4 5 6 7 8 9
##   .. ..- attr(*, "names")= chr [1:9] "Married" "Civil partnership" "Separated from spouse
##  $ URBRURAL: 'haven_labelled' num [1:32823] 1 1 1 NA 1 2 NA 2 2 NA ...
##   ..- attr(*, "label")= chr "Place of living: urban - rural"
##   ..- attr(*, "labels")= Named num [1:7] 1 2 3 4 5 7 9
##   .. ..- attr(*, "names")= chr [1:7] "A big city" "The suburbs or outskirts of a big city
##  - attr(*, "notes")= chr [1:45] "document Plan File: /Users/marcic/Desktop/old/GPS2011 sa
```

```
# ISSP2012jh1c.data %>% summary() %>% kable()
```

Metatietojen (3) ja maa-muuttujien (3) lisäksi aineistossa on seitsemäntoista muuttujaa. Yhdeksän muuttujaa ovat ns. substanssikysymysten vastauksia, joilla luodataan asenteita sukupuolirooleihin ja perhearvoihin. Taustamuuttujia on kahdeksan.

Yhdeksän kysymystä (lyhennetyt versiot, englanniksi), vastausvaihtoehdot

Vastausvaihtoehdot:

1 = täysin samaa mieltä, 2 = samaa mieltä, 3 = ei samaa eikä eri mieltä, 4 = eri mieltä, 5 = täysin eri mieltä

Q1a Working mother can have warm relation with child Q1b Pre-school child suffers through working mother Q1c Family life suffers through working mother Q1d Women's preference: home and children Q1e Being housewife is satisfying Q2a Both should contribute to household income Q2b Men's job is earn money, women's job household

Q3a Should women work: Child under school age Q3b Should women work: Youngest kid at school

Vastausvaihtoehdot: "Work full-time" "Work part-time" "Stay at home", "Can't choose" 1 = W, 2 = w, 3 = H, NA = 6,8,9 ei tässä eriteltynä. 6 on Taiwanin oma vastausvaihtoehto, 8 = en osaa sanoa ja 9 = no answer.

```
# Muuttuja taulukkona - karkea tapa

tabVarnames <- c(substvars1,bgvars1) # muuttujanimet muuttujille

# Kysymysten lyhyet versiot englanniksi
```

```r
tabVarDesc <- c("Q1a Working mother can have warm relation with child ",
                "Q1b Pre-school child suffers through working mother",
                "Q1c Family life suffers through working mother",
                "Q1d Women's preference: home and children",
                "Q1e Being housewife is satisfying",
                "Q2a Both should contribute to household income",
                "Q2b Men's job is earn money, women's job household",
                "Q3a Should women work: Child under school age",
                "Q3b Should women work: Youngest kid at school",
                "Respondents age ",
                "Respondents gender",
                "Highest completed degree of education: Categories for international compari
                "Main status: work, unemployed, in education...",
                "Top-Bottom self-placement (10 pt scale)",
                "How many children in household: children between [school age] and 17 years
                "Legal partnership status: married, civil partership...",
                "Place of living: urban - rural"
              )
#tabVarDesc

# Taulukko

# luodaan df - varoitus: data_frame() is deprecated, use tibble" (4.2.20),
# vaihdetaan tibbleen (21.2.20)

# jhVarTable1.df <- data_frame(tabVarnames,tabVarDesc) OLD
jhVarTable1.tbl <- tibble(tabVarnames,tabVarDesc)
cols_jhVarTable1 <- c("muuttuja","kysymyksen tunnus, lyhennetty kysymys")
colnames(jhVarTable1.tbl) <- cols_jhVarTable1
str(jhVarTable1.tbl)
```

```
## tibble [17 x 2] (S3: tbl_df/tbl/data.frame)
##  $ muuttuja                            : chr [1:17] "V5" "V6" "V7" "V8" ...
##  $ kysymyksen tunnus, lyhennetty kysymys: chr [1:17] "Q1a Working mother can have warm re
```

```r
# Suomalaiset pitkät kysymykset
vastf1 <- c("Q1a Työssäkäyvä äiti pystyy luomaan lapsiinsa aivan yhtä lämpimän
            ja turvallisen suhteen kuin äiti, joka ei käy työssä")

vastf2 <- c("Q1b Alle kouluikäinen lapsi todennäköisesti kärsii, jos hänen äitinsä käy työss
vastf3 <- c("Q1c Kaiken kaikkiaan perhe-elämä kärsii, kun naisella on kokopäivätyö.")
vastf4 <- c("Q1d On hyvä käydä töissä mutta tosiasiassa useimmat naiset haluavat
            ensisijaisesti kodin ja lapsia.")
vastf5 <- c("Q1e Kotirouvana oleminen on aivan yhtä antoisaa kuin ansiotyön tekeminen.")
vastf6 <- c("Q2a Sekä miehen että naisen tulee osallistua perheen toimeentulon hankkimiseen.
vastf7 <- c("Q2b Miehen tehtävä on ansaita rahaa; naisen tehtävä on huolehtia kodista ja per
```

```
vastf8 <- c("Q3a Millä tavoin naisten pitäisi mielestäsi käydä työssä seuraavissa tilanteiss
            Kun perheessä on alle kouluikäinen lapsi")
vastf9 <- c("Q3b Millä tavoin naisten pitäisi mielestäsi käydä työssä seuraavissa tilanteiss
            Kun nuorin lapsi on aloittanut koulunkäynnin")

tabVarDesc_fi <- c(vastf1,vastf2,vastf3,vastf4,vastf5,vastf6,vastf7, vastf8,vastf9)
#tabVarDesc_fi
tabVarnames_subst <- c(substvars1)
# jhVarTable1_fi.df <- data_frame(tabVarnames_subst,tabVarDesc_fi) OLD
jhVarTable1_fi.tbl <- tibble(tabVarnames_subst,tabVarDesc_fi)
cols_jhVarTable1 <- c("muuttuja","Kysymyksen tunnus, suomenkielisen lomakkeen kysymys")
colnames(jhVarTable1_fi.tbl) <- cols_jhVarTable1

# TAULUKODEN TULOSTUS

#  kable(booktab = T) # booktab = T gives us a pretty APA-ish table
# Lyhyet kysymykset englanniksi

# OLD
# knitr::kable(jhVarTable1.df, booktab=TRUE,
#                 fig.cap="ISSP2012:Työelämä ja perhearvot - valitut muuttujat")
knitr::kable(jhVarTable1.tbl, booktab = TRUE,
               fig.cap = "ISSP2012:Työelämä ja perhearvot - valitut muuttujat")
```

muuttuja	kysymyksen tunnus, lyhennetty kysymys
V5	Q1a Working mother can have warm relation with child
V6	Q1b Pre-school child suffers through working mother
V7	Q1c Family life suffers through working mother
V8	Q1d Women's preference: home and children
V9	Q1e Being housewife is satisfying
V10	Q2a Both should contribute to household income
V11	Q2b Men's job is earn money, women's job household
V12	Q3a Should women work: Child under school age
V13	Q3b Should women work: Youngest kid at school
SEX	Respondents age
AGE	Respondents gender
DEGREE	Highest completed degree of education: Categories for international comparison
MAINSTAT	Main status: work, unemployed, in education...
TOPBOT	Top-Bottom self-placement (10 pt scale)
HHCHILDR	How many children in household: children between [school age] and 17 years of age
MARITAL	Legal partnership status: married, civil partership...
URBRURAL	Place of living: urban - rural

```
# Suomen lomakkeen kysymykset (löytyy myös kuva lomakkeen sivusta)

# OLD
# knitr::kable(jhVarTable1_fi.df, booktab=TRUE,
#              fig.cap="ISSP2012: suomenkielisen lomakkeen kysymykset")

knitr::kable(jhVarTable1_fi.tbl, booktab = TRUE,
             fig.cap = "ISSP2012: suomenkielisen lomakkeen kysymykset")
```

muuttuja	Kysymyksen tunnus, suomenkielisen lomakkeen kysymys
V5	Q1a Työssäkäyvä äiti pystyy luomaan lapsiinsa aivan yhtä lämpimän ja turvallisen suhteen kuin äiti, joka ei käy työssä
V6	Q1b Alle kouluikäinen lapsi todennäköisesti kärsii, jos hänen äitinsä käy työssä.
V7	Q1c Kaiken kaikkiaan perhe-elämä kärsii, kun naisella on kokopäivätyö.
V8	Q1d On hyvä käydä töissä mutta tosiasiassa useimmat naiset haluavat ensisijaisesti kodin ja lapsia.
V9	Q1e Kotirouvana oleminen on aivan yhtä antoisaa kuin ansiotyön tekeminen.
V10	Q2a Sekä miehen että naisen tulee osallistua perheen toimeentulon hankkimiseen.
V11	Q2b Miehen tehtävä on ansaita rahaa; naisen tehtävä on huolehtia kodista ja perheestä.
V12	Q3a Millä tavoin naisten pitäisi mielestäsi käydä työssä seuraavissa tilanteissa? Kun perheessä on alle kouluikäinen lapsi
V13	Q3b Millä tavoin naisten pitäisi mielestäsi käydä työssä seuraavissa tilanteissa? Kun nuorin lapsi on aloittanut koulunkäynnin

```
# Taulukot voivat olla hankalia eristyisesti PDF-tulostuksessa, jos ne ovat
# monimutkaisia tai solujen "koot" (merkkiä/solu) vaihtelevat paljon.

# Kokeillaan taulukoiden yhdistämistä, jos aikaa jää. Ei luultavasti kannata, kun halutaan
# html-tulostus samalla koodilla (26.12.18).
```

Tarkemmat kysymysten muotoilut poikkeavat tietysti hieman eri maiden välillä. Suomen lomakkeet täydelliset kysymykset voi tarkista tiedostosta ZA5900_q_fi-fi.pdf, löytyy zcat-sivustolta. Tarkemmat kuvaukset lähes tuhatsivuisessa koodikirjassa ZA5900_cdb.pdf (**refworks-viite pitäisi löytyä**, ja ISSP dokumentit kerrotaan luvun alussa).

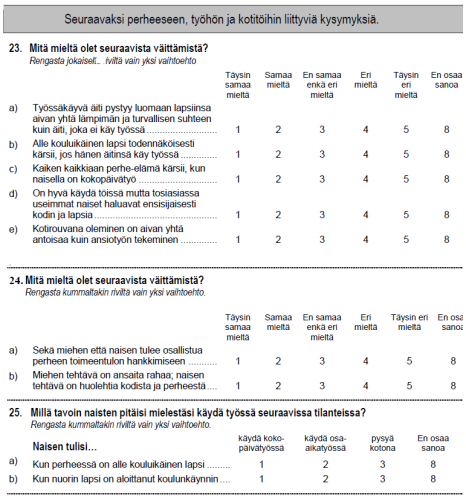
Bookdown-versiossa taulukot omiksi koodilohkoiksi, ja fig.caption - optiolla taulukon otsikko.

Kysymyslomakkeen kuva, vai kuva liitteisiin? **Liitteisiin.**

```
knitr::include_graphics('img/substvar_fi_Q1Q2.png')
```

1.5 Datan valinnan vaiheet ja puuttuvat tiedot

edit 24.2.20 Toistoa



Seuraavaksi perheeseen, työhön ja kotitöihin liittyviä kysymyksiä.

23. Mitä mieltä olet seuraavista väittämistä?
Rengasta jokaiseli... tvältä vain yksi vaihtoehto

		Täysin samaa mieltä	Samaa mieltä	En samaa enkä eri mieltä	Eri mieltä	Täysin eri mieltä	En osaa sanoa
a)	Työssäkäyvä äiti pystyy luomaan lapsiinsa aivan yhtä lämpimän ja turvallisen suhteen kuin äiti, joka ei käy työssä	1	2	3	4	5	8
b)	Alle kouluikäinen lapsi todennäköisesti kärsii, jos hänen äitinsä käy työssä	1	2	3	4	5	8
c)	Kaiken kaikkiaan perhe-elämä kärsii, kun naisella on kokopäivätyö	1	2	3	4	5	8
d)	On hyvä käydä töissä mutta tosiasiassa useimmat naiset haluavat ensisijaisesti kodin ja lapsia	1	2	3	4	5	8
e)	Kotirouvana oleminen on aivan yhtä antoisaa kuin ansiotyön tekeminen	1	2	3	4	5	8

24. Mitä mieltä olet seuraavista väittämistä?
Rengasta kummaltakin rivltä vain yksi vaihtoehto.

		Täysin samaa mieltä	Samaa mieltä	En samaa enkä eri mieltä	Eri mieltä	Täysin eri mieltä	En osa sanoa
a)	Sekä miehen että naisen tulee osallistua perheen toimeentulon hankkimiseen	1	2	3	4	5	8
b)	Miehen tehtävä on ansaita rahaa; naisen tehtävä on huolehtia kodista ja perheestä	1	2	3	4	5	8

25. Millä tavoin naisten pitäisi mielestäsi käydä työssä seuraavissa tilanteissa?
Rengasta kummaltakin rivltä vain yksi vaihtoehto.

	Naisen tulisi...	käydä koko-päivatyössä	käydä osa-aikatyössä	pysyä kotona	En osaa sanoa
a)	Kun perheessä on alle kouluikäinen lapsi	1	2	3	8
b)	Kun nuorin lapsi on aloittanut koulunkäynnin	1	2	3	8

Kuva 1: Suomen lomake

ks Perusasiat havaintojen puuttellisuudesta kyselytutkimusissa. Yksikkövastauskato (unit non-response), eräsvastauskato (item non-response). Mitä on raportoitava, kun käytetään valmista aineistoa? Erävastauskatoa analysoidaan, kun käytetään kaikkia valittuja muuttujia.

Yksikkövastauskato on otettu vaihtelevasti huomioon, kun kyselyn toteuttaja on editoinut ja tarkastanut datan. Eri maiden datassa on (mutta ei aina!) mukana painot mm. vastauskadon oikaiksemiseen **Viittet - tekninen raportti**. Myös selaimella voi zcat-sivustolla tutkailla kysymyksittäin.

Datakatalogi-dokumentista näkee vastausten jakauman jokaisen kysymyksen osalta, myös puuttuvien tietojen tarkemman koodauksen.

1. Valitaan 25 maata ja muuttujat
2. Johdattelevissa esimerkeissä valitaan kuusi maata ja kolme muuttujaa. Jätetään pois kaikki havainnot (vastaukset) joissa on puuttuvia tietoja ("listwise deletion")
3. Kun laajempi aineisto otetaan käyttöön, joudutaan pohtimaan miten puuttuvia havaintoja käsitellään. Jos kyse on selvistä virheistä (esim. haastateltavan ikä puuttu) havainnot jätetään pois, muuten mietitään.

Miten puuttuvia tietoja (erävastuskato, havainnossa puuttu joku tieto) käsitellään?

edit Tämä on vähän hämärää, ehkä pois? (30.1.20)

1. Miksi tieto puuttuu, mitä "puuttuva tieto" tarkoittaa?

Joissain kysymyksissä (V12, V13) puuttuvaksi tiedoksi kirjautuu vastaus ("en

osaa sanoa") "ei vastausta" - vaihtoehdon lisäksi. Nämä mukaan.

Ikä ja sukupuoli: ilmeinen virhe, joten jätetään havainnot pois (näitä ei ole paljon).

2. Puuttuvien tietojen jakauma?

edit 24.2.20) Kun laajempi aineisto ja puuttuvat arvot otetaan mukaan analyysiin loppuluvuissa, vilkaistaa pikaisesti erävastauskadon rakennetta.

3. Onko puuttuvia tietoja tasaisesti eri maissa, vai vaihteleeko niiden suhteellinen osuus?

4. Onko joissain tai jossain maassa huomattava määrä puuttuvia tietoja?

5. Onko puuttuvia tietoja paljon vai vähän?

Tarkemmin puuttuneisuutta ei analysoida. Esimerkkejä löytyy (MG, CAiP ja "vihreä kirja"). Kaksi R-pakettia, joilla pikaisesti vilkaistaan dataa, ei vielä mukana tässä (24.2.20). **edit** Viite!

Koko aineistossa (valitut 25 maata) kysymyksen Q1b (muuttuja V6) vastauksista puuttuvia tietoja on 3,5 prosenttia (1219/34271). **Huom:** kun pudotetaan havainnot joilta SEX tai AGE puuttuu, N = 32823. On oikea määrä (5.7.2019, kts. treeni2- projekti, Data_iso1.R).

edit Vanhoja koodilohkoja, olkoon toistaiseksi mukana (11.2.20)

Puuttuvien tietojen tarkempi koodaus ISSP-datassa:

0: Not applicapble (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 spesific variables, all the other countries are coded NAP.

GESIS adds 'Not available' for variables, which in singe countries may not have been conducted for whatever reason."

1.6 Perusmuunnokset ISSP2012 - datalle

Datatiedosto on ISSP2012jh1.data, ja luokittelumuuttujat muunnetaan R:n factor- muuttujaksi.

Jokaisesta muuttujasta on kaksi versiota, toisessa puuttuvat tiedot ovat R:n "NA"- arvoja ja toisessa "NA"-arvo on eksplisiittinen muuttuja ("missing").

Substanssimuuttujien luokkien tunnukset (faktorilabelit, levels?) muutetaan graafisiin analyyseihin sopivan lyhyiksi. Taustamuuttujien luokittelua ja luokkien tunnuksia pohditaan, kun ne otetaan käyttöön.

TODO 30.1.20 Tarkistukset, varmistukset jne. (24.2.20) Lisätty muutama testi, paljon välitulostuksia joita voi tarvittaessa kommentoida koodista pois.

TODO 2.2.20 Muunnetaanko muuttujan maa (C_ALPHAN as_factor) järjestys heti samaksi kuin C_ALPHAN? Nyt tehdään G1_2_johdesim.Rmd:ssä. (24.2.20) tehty, muunnetaan heti alussa. Käytännössä kaikenlaisia korjailuja joutuu tekemään myös analyyseissä käytettävissä R-datoissa.

TODO 3.2.20 Aluejaon maakoodi V3 mukaan, pohditaan järjestykset jne luvussa G1_2_johdesim.Rmd. (24.2.20) tehty, järjesteään myös uusi muuttuja C_ALPHAN-järjestykseen.

1.6.1 Vaihe 1 - muuttujat joissa ei ole puuttuvia tietoja

Aineistosta on jätetty pois ne havainnot, joissa ikä (AGE) tai sukupuoli (SEX) on puuttuva tieto. Aika paljon tarkistuksia, kolmen maa-muuttujaa järjestetään C_ALPHAN - muuttujan järjestykseen. Ikä-muuttuja säilyy numeerisena. Ensimmäiseen faktori-tyypin muuttujaan jää tyhjänä luokkana puuttuva tieto, luokka poistetaan.

```
# VAIHE 1 - muuttujat joissa ei ole puuttuvia tietoja

# vaihe 1.1 haven_labelled ja chr -> as_factor

ISSP2012jh1d.dat <- ISSP2012jh1c.data %>%
    mutate(maa = as_factor(C_ALPHAN), # ei puuttuvia, ei tyhjiä leveleitä
           maa3 = as_factor(V3),  # maakoodi, jossa aluejako joillan mailla
           sp1 = as_factor(SEX), # ei puuttuvia, tyhjä level "no answer" 999
         )


# C_ALPHAN - maa - maa3 tarkistuksia

# V3
# "Pulma" on järjestys. C_ALPHAN ("chr") on aakkosjärjestyksessä, kun luodaan
# maa = as_factor(C_ALPHAN) järjestys muuttuu (esiintymisjärjestys datassa?)
# maa3 muunnetaan maakoodista (haven_labelled' num), jonka

str(ISSP2012jh1d.dat$maa) #Country Prefix ISO 3166 Code - alphanumeric
```

```
##  Factor w/ 25 levels "AU","AT","BG",..: 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, "label")= chr "Country Prefix ISO 3166 Code - alphanumeric"
```

```
# attributes(ISSP2012jh1d.dat$maa) # ei tyhiä levels-arvoja, 25 levels
# ISSP2012jh1d.dat$maa %>% fct_unique()
# ISSP2012jh1d.dat$maa %>% fct_count() # summary kertoo samat tiedot (20.2.20)
# sum(is.na(ISSP2012jh1d.dat$maa)) # ei puuttuvia tietoja
ISSP2012jh1d.dat$maa %>% summary() # mukana vain valitut 25 maata
```

```
##   AU   AT   BG   CA   HR   CZ   DK   FI   FR   HU   IS   IE   LV   LT   NL   NO
## 1557 1182 1003  953  997 1804 1403 1171 2409 1012 1172 1166 1000 1187 1315 1444
##   PL   RU   SK   SI   SE   CH   BE   DE   PT
## 1115 1525 1128 1034 1059 1237 2192 1761  997
```

```
str(ISSP2012jh1d.dat$maa3)  #"Country/ Sample ISO 3166 Code
```

```
##  Factor w/ 45 levels "AR-Argentina",..: 2 2 2 2 2 2 2 2 2 2 ...
##  - attr(*, "label")= chr "Country/ Sample ISO 3166 Code (see V4 for codes for whole natio
```

```
                            #(see V4 for codes for whole nation states)"
                            # 29 levels
str(ISSP2012jh1d.dat$V3)
```

```
##  'haven_labelled' num [1:32823] 36 36 36 36 36 36 36 36 36 36 ...
##  - attr(*, "label")= chr "Country/ Sample ISO 3166 Code (see V4 for codes for whole natio
##  - attr(*, "labels")= Named num [1:45] 32 36 40 100 124 152 156 158 191 203 ...
##   ..- attr(*, "names")= chr [1:45] "AR-Argentina" "AU-Australia" "AT-Austria" "BG-Bulgari
```

```
# attributes(ISSP2012jh1d.dat$maa3) # ei tyhiä levels-arvoja, 29 levels
# sum(is.na(ISSP2012jh1d.dat$maa3)) # nolla ei ole puuttuva tieto! (3.2.20)
# ISSP2012jh1d.dat$maa3 %>% fct_unique()
# ISSP2012jh1d.dat$maa3 %>% fct_count()
# Vain näissä on jaettu maan havainnot (3.2.20)
#
# [38] BE-FLA-Belgium/ Flanders
# [39] BE-WAL-Belgium/ Wallonia
# [40] BE-BRU-Belgium/ Brussels
# [41] DE-W-Germany-West
# [42] DE-E-Germany-East
# [43] PT-Portugal 2012: first fieldwork round (main sample)
# [44] PT-Portugal 2012: second fieldwork round (complementary sample)

# ISSP2012jh1d.dat$maa3 %>% fct_count() #miksi ei tulosta mitään? (3.2.2020)

# ISSP2012jh1d.dat$maa3 %>% summary()
# ISSP2012jh1d.dat$maa3 %>% fct_unique()
# maa3: 25 maata, havaintojen määrä. Poisjätetyissä havaintoja 0.
# glimpse(ISSP2012jh1d.dat$maa3)
# head(ISSP2012jh1d.dat$maa3)
# length(levels(ISSP2012jh1d.dat$maa3))
```

```
# C_ALPHAN alkuperäinen järjestys, maa aakkosjärjestyssä  (2.2.20)
#
# Huom1: Myös merkkijonomuuttujaa C_ALPHAN tarvitaan jatkossa.
#
# Huom2: kun dataa rajataan, on tarkistettava ja tarvittaessa poistettava
# "tyhjät" R-factor - muuttujan "maa" luokat (3.2.2020)


# vaihe 1.2 tyhjät luokat (levels) pois faktoreista

ISSP2012jh1d.dat <- ISSP2012jh1d.dat %>%
    mutate(sp = fct_drop(sp1),
           maa3 = fct_drop(maa3))
# Poistetaan maa3-muuttujan tyhjät luokat (3.2.20)

#  maa3 - tarkistuksia

# str(ISSP2012jh1d.dat$maa3)  # 29 levels

attributes(ISSP2012jh1d.dat$maa3) #
```

```
## $levels
##  [1] "AU-Australia"
##  [2] "AT-Austria"
##  [3] "BG-Bulgaria"
##  [4] "CA-Canada"
##  [5] "HR-Croatia"
##  [6] "CZ-Czech Republic"
##  [7] "DK-Denmark"
##  [8] "FI-Finland"
##  [9] "FR-France"
## [10] "HU-Hungary"
## [11] "IS-Iceland"
## [12] "IE-Ireland"
## [13] "LV-Latvia"
## [14] "LT-Lithuania"
## [15] "NL-Netherlands"
## [16] "NO-Norway"
## [17] "PL-Poland"
## [18] "RU-Russia"
## [19] "SK-Slovakia"
## [20] "SI-Slovenia"
## [21] "SE-Sweden"
## [22] "CH-Switzerland"
## [23] "BE-FLA-Belgium/ Flanders"
```

```
## [24] "BE-WAL-Belgium/ Wallonia"
## [25] "BE-BRU-Belgium/ Brussels"
## [26] "DE-W-Germany-West"
## [27] "DE-E-Germany-East"
## [28] "PT-Portugal 2012: first fieldwork round (main sample)"
## [29] "PT-Portugal 2012: second fieldwork round (complementary sample)"
##
## $class
## [1] "factor"
##
## $label
## [1] "Country/ Sample ISO 3166 Code (see V4 for codes for whole nation states)"
```

```
#sum(is.na(ISSP2012jh1d.dat$maa3)) # nolla ei ole puuttuva tieto! (3.2.20)
# ISSP2012jh1d.dat$maa3 %>% summary()
# ISSP2012jh1d.dat$maa3 %>% fct_unique()
ISSP2012jh1d.dat$maa3 %>% fct_count() # miksi ei tulosta? Tulostaa komentoriviltä!
```

f	n
AU-Australia	1557
AT-Austria	1182
BG-Bulgaria	1003
CA-Canada	953
HR-Croatia	997
CZ-Czech Republic	1804
DK-Denmark	1403
FI-Finland	1171
FR-France	2409
HU-Hungary	1012
IS-Iceland	1172
IE-Ireland	1166
LV-Latvia	1000
LT-Lithuania	1187
NL-Netherlands	1315
NO-Norway	1444
PL-Poland	1115
RU-Russia	1525
SK-Slovakia	1128
SI-Slovenia	1034
SE-Sweden	1059
CH-Switzerland	1237
BE-FLA-Belgium/ Flanders	1090
BE-WAL-Belgium/ Wallonia	543
BE-BRU-Belgium/ Brussels	559
DE-W-Germany-West	1205
DE-E-Germany-East	556

f	n
PT-Portugal 2012: first fieldwork round (main sample)	894
PT-Portugal 2012: second fieldwork round (complementary sample)	103

```
str(ISSP2012jh1d.dat$C_ALPHAN)
```

```
##  chr [1:32823] "AU" "AU" "AU" "AU" "AU" "AU" "AU" "AU" "AU" "AU" "AU" "AU" ...
##  - attr(*, "label")= chr "Country Prefix ISO 3166 Code - alphanumeric"
##  - attr(*, "format.spss")= chr "A20"
##  - attr(*, "display_width")= int 22
```

```
attributes(ISSP2012jh1d.dat$C_ALPHAN)
```

```
## $label
## [1] "Country Prefix ISO 3166 Code - alphanumeric"
##
## $format.spss
## [1] "A20"
##
## $display_width
## [1] 22
```

```
ISSP2012jh1d.dat %>% tableX(C_ALPHAN, maa)
```

C_ALPHAN/maa	AU	AT	BG	CA	HR	CZ	DK	FI	FR	HU	IS	IE	LV
AT	0	1182	0	0	0	0	0	0	0	0	0	0	0
AU	1557	0	0	0	0	0	0	0	0	0	0	0	0
BE	0	0	0	0	0	0	0	0	0	0	0	0	0
BG	0	0	1003	0	0	0	0	0	0	0	0	0	0
CA	0	0	0	953	0	0	0	0	0	0	0	0	0
CH	0	0	0	0	0	0	0	0	0	0	0	0	0
CZ	0	0	0	0	0	1804	0	0	0	0	0	0	0
DE	0	0	0	0	0	0	0	0	0	0	0	0	0
DK	0	0	0	0	0	0	1403	0	0	0	0	0	0
FI	0	0	0	0	0	0	0	1171	0	0	0	0	0
FR	0	0	0	0	0	0	0	0	2409	0	0	0	0
HR	0	0	0	0	997	0	0	0	0	0	0	0	0
HU	0	0	0	0	0	0	0	0	0	1012	0	0	0
IE	0	0	0	0	0	0	0	0	0	0	0	1166	0
IS	0	0	0	0	0	0	0	0	0	0	1172	0	0
LT	0	0	0	0	0	0	0	0	0	0	0	0	0
LV	0	0	0	0	0	0	0	0	0	0	0	0	100
NL	0	0	0	0	0	0	0	0	0	0	0	0	0
NO	0	0	0	0	0	0	0	0	0	0	0	0	0
PL	0	0	0	0	0	0	0	0	0	0	0	0	0

C_ALPHAN/maa	AU	AT	BG	CA	HR	CZ	DK	FI	FR	HU	IS	IE	LV
PT	0	0	0	0	0	0	0	0	0	0	0	0	0
RU	0	0	0	0	0	0	0	0	0	0	0	0	0
SE	0	0	0	0	0	0	0	0	0	0	0	0	0
SI	0	0	0	0	0	0	0	0	0	0	0	0	0
SK	0	0	0	0	0	0	0	0	0	0	0	0	0
Total	1557	1182	1003	953	997	1804	1403	1171	2409	1012	1172	1166	100

```
ISSP2012jh1d.dat %>% tableX(C_ALPHAN, maa3)
```

C_ALPHAN/maa3	AU-Australia	AT-Austria	BG-Bulgaria	CA-Canada	HR-Croatia	CZ-Czech Re
AT	0	1182	0	0	0	0
AU	1557	0	0	0	0	0
BE	0	0	0	0	0	0
BG	0	0	1003	0	0	0
CA	0	0	0	953	0	0
CH	0	0	0	0	0	0
CZ	0	0	0	0	0	1804
DE	0	0	0	0	0	0
DK	0	0	0	0	0	0
FI	0	0	0	0	0	0
FR	0	0	0	0	0	0
HR	0	0	0	0	997	0
HU	0	0	0	0	0	0
IE	0	0	0	0	0	0
IS	0	0	0	0	0	0
LT	0	0	0	0	0	0
LV	0	0	0	0	0	0
NL	0	0	0	0	0	0
NO	0	0	0	0	0	0
PL	0	0	0	0	0	0
PT	0	0	0	0	0	0
RU	0	0	0	0	0	0
SE	0	0	0	0	0	0
SI	0	0	0	0	0	0
SK	0	0	0	0	0	0
Total	1557	1182	1003	953	997	1804

```
ISSP2012jh1d.dat %>% tableX(maa, maa3)
```

maa/maa3	AU-Australia	AT-Austria	BG-Bulgaria	CA-Canada	HR-Croatia	CZ-Czech Republic
AU	1557	0	0	0	0	0
AT	0	1182	0	0	0	0
BG	0	0	1003	0	0	0
CA	0	0	0	953	0	0
HR	0	0	0	0	997	0
CZ	0	0	0	0	0	1804
DK	0	0	0	0	0	0
FI	0	0	0	0	0	0
FR	0	0	0	0	0	0
HU	0	0	0	0	0	0
IS	0	0	0	0	0	0
IE	0	0	0	0	0	0
LV	0	0	0	0	0	0
LT	0	0	0	0	0	0
NL	0	0	0	0	0	0
NO	0	0	0	0	0	0
PL	0	0	0	0	0	0
RU	0	0	0	0	0	0
SK	0	0	0	0	0	0
SI	0	0	0	0	0	0
SE	0	0	0	0	0	0
CH	0	0	0	0	0	0
BE	0	0	0	0	0	0
DE	0	0	0	0	0	0
PT	0	0	0	0	0	0
Total	1557	1182	1003	953	997	1804

```
ISSP2012jh1d.dat %>% tableX(V3, maa3)
```

V3/maa3	AU-Australia	AT-Austria	BG-Bulgaria	CA-Canada	HR-Croatia	CZ-Czech Republic	D
36	1557	0	0	0	0	0	0
40	0	1182	0	0	0	0	0
100	0	0	1003	0	0	0	0
124	0	0	0	953	0	0	0
191	0	0	0	0	997	0	0
203	0	0	0	0	0	1804	0
208	0	0	0	0	0	0	1
246	0	0	0	0	0	0	0
250	0	0	0	0	0	0	0
348	0	0	0	0	0	0	0
352	0	0	0	0	0	0	0
372	0	0	0	0	0	0	0
428	0	0	0	0	0	0	0

V3/maa3	AU-Australia	AT-Austria	BG-Bulgaria	CA-Canada	HR-Croatia	CZ-Czech Republic	D
440	0	0	0	0	0	0	0
528	0	0	0	0	0	0	0
578	0	0	0	0	0	0	0
616	0	0	0	0	0	0	0
643	0	0	0	0	0	0	0
703	0	0	0	0	0	0	0
705	0	0	0	0	0	0	0
752	0	0	0	0	0	0	0
756	0	0	0	0	0	0	0
5601	0	0	0	0	0	0	0
5602	0	0	0	0	0	0	0
5603	0	0	0	0	0	0	0
27601	0	0	0	0	0	0	0
27602	0	0	0	0	0	0	0
62001	0	0	0	0	0	0	0
62002	0	0	0	0	0	0	0
Total	1557	1182	1003	953	997	1804	1

```
# sp, sp1, SEX - tarkistuksia

ISSP2012jh1d.dat$sp %>% fct_count()
```

f	n
Male	14789
Female	18034

```
ISSP2012jh1d.dat$sp %>% fct_count()
```

f	n
Male	14789
Female	18034

```
ISSP2012jh1d.dat %>% tableX(SEX,sp1)
```

SEX/sp1	Male	Female	No answer	Total
1	14789	0	0	14789
2	0	18034	0	18034
Total	14789	18034	0	32823

```
ISSP2012jh1d.dat %>% tableX(SEX,sp)
```

SEX/sp	Male	Female	Total
1	14789	0	14789
2	0	18034	18034
Total	14789	18034	32823

```
ISSP2012jh1d.dat %>% tableX(sp1,sp)
```

sp1/sp	Male	Female	Total
Male	14789	0	14789
Female	0	18034	18034
No answer	0	0	0
Total	14789	18034	32823

```
# vaihe 1.3 uudet "faktorilabelit"
ISSP2012jh1d.dat <- ISSP2012jh1d.dat %>%
    mutate(sp =
                fct_recode(sp,
                           "m" = "Male",
                           "f" = "Female")
            )

# Tarkistuksia

ISSP2012jh1d.dat$sp %>% fct_unique()
```

```
## [1] m f
## Levels: m f
```

```
ISSP2012jh1d.dat$sp %>% fct_count()
```

f	n
m	14789
f	18034

```
ISSP2012jh1d.dat$sp %>% summary()
```

```
##     m     f
## 14789 18034
```

		[1]	[2]
[1]AGE	[1]AGE	1.00	
[2]ika	[2]ika	1.00	1.00

```
# AGE -> ika
# AGE----
ISSP2012jh1d.dat$ika <- ISSP2012jh1d.dat$AGE

# Tarkistuksia
attributes(ISSP2012jh1d.dat$ika) # tyhjä level "No answer
```

```
## $label
## [1] "Age of respondent"
##
## $labels
##  15 years  16 years  17 years  18 years 102 years No answer
##        15        16        17        18       102       999
##
## $class
## [1] "haven_labelled"
```

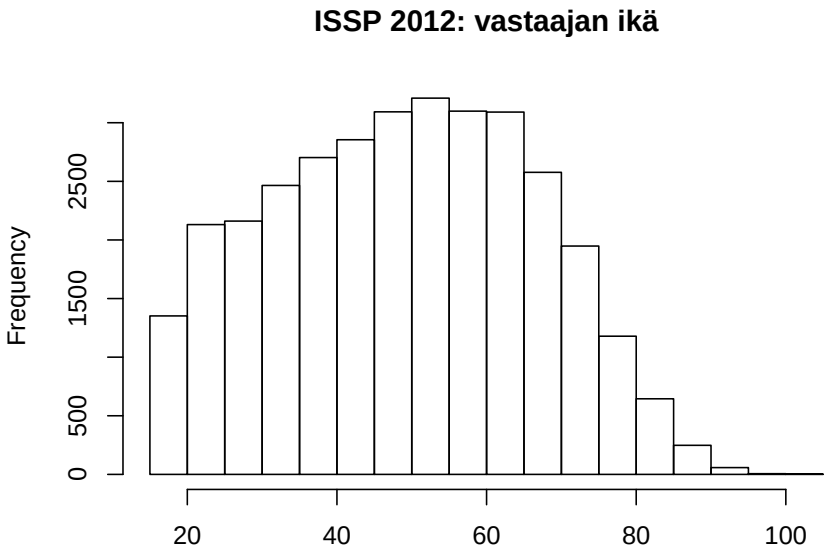
```
ISSP2012jh1d.dat$ika %>% summary()
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
15	36	50	49.51607	63	102

```
ISSP2012jh1d.dat %>%
tableC(AGE, ika,cor_type = "pearson", na.rm = FALSE, rounding = 5,
       output = "text", booktabs = TRUE, caption = NULL, align = NULL,
       float = "htb") %>% kable()
```

```
## N = 32823
## Note: pearson correlation (p-value).
```

```
ISSP2012jh1d.dat$ika %>% hist(main = "ISSP 2012: vastaajan ikä")
```



.

```
# str(ISSP2012jh1d.dat) - tarkistus
```

1.6.2 Vaihe 2

Vaihessa 2 luodaan samalla samalla periaatteella substanssi- ja taustamuuttujille kaksi R-factor- tyypin muuttujaa. Toisessa (esim. Q1a) puuttuva tieto on R-ohjelmiston sisäinen NA-arvo. Toisessa (Q1am) puuttuva tieto on yksi luokittelumuuttujan arvo("missing").

```
# Substanssi- ja taustamuuttujat R-faktoreiksi
ISSP2012jh1d.dat <- ISSP2012jh1d.dat %>%
    mutate(Q1a1 = as_factor(V5), #labels
           Q1b1 = as_factor(V6),
           Q1c1 = as_factor(V7),
           Q1d1 = as_factor(V8),
           Q1e1 = as_factor(V9),
           Q2a1 = as_factor(V10),
           Q2b1 = as_factor(V11),
           Q3a1 = as_factor(V12), #labels = vastQ3_labels (W,w,H)
           Q3b1 = as_factor(V13), #labels = vastQ3_labels
           edu1 = as_factor(DEGREE),
           msta1 = as_factor(MAINSTAT),
           sosta1 = as_factor(TOPBOT),
           nchild1 = as_factor(HHCHILDR),
```

```
            lifsta1 = as_factor(MARITAL),
            urbru1 = as_factor(URBRURAL)

    )

# Muuttujat Q1a1...urbru1 ovat apumuuttujia, joissa on periaatteessa kaikki SPSS-
# tiedostosta siirtyvä metatieto. Poikkeus on SPSS:n kolme tarkentavaa koodia
# puuttuvalle tiedolle, ne saisi mukaan read_spss - parametrin avulla (user_na=TRUE)
#

# Tarkistusksia
# ISSP2012jh1d.dat %>% summary()

ISSP2012jh1d.dat %>%
    select(Q1a1, Q1b1, Q1c1,Q1d1,Q1e1, Q2a1, Q2b1, Q3a1,Q3b1) %>%
    summary()
```

Q1a1	Q1b1	Q1c1	
Agree :12352	Disagree :9003	Disagree :8706	
Strongly agree :11116	Agree :8389	Agree :8263	
Disagree : 4074	Neither agree nor disagree:5949	Neither agree nor disagree:6000	Nei
Neither agree nor disagree: 3382	Strongly disagree :5547	Strongly disagree :5960	
Strongly disagree : 1051	Strongly agree :2747	Strongly agree :2838	
(Other) : 0	(Other) : 0	(Other) : 0	
NA's : 848	NA's :1188	NA's :1056	

```
ISSP2012jh1d.dat %>%
    select(edu1,msta1, sosta1, nchild1, lifsta1, urbru1) %>%
    summary()
```

edu1
Lower secondary (secondary completed does not allow entry to university: obligatory school) :7811
Upper secondary (programs that allows entry to university :7115
Post secondary, non-tertiary (other upper secondary programs toward labour market or technical formatio
Lower level tertiary, first stage (also technical schools at a tertiary level) :5147
Upper level tertiary (Master, Dr.) :4762
(Other) :2022
NA's : 308

```
# Substanssimuuttujat - ristiintaulukoinnit riittävät (6.2.20)

# ISSP2012jh1d.dat$Q1a1 %>% fct_count()
```

```
# ISSP2012jh1d.dat$Q1b1 %>% fct_count()
# ISSP2012jh1d.dat$Q1c1 %>% fct_count()
# ISSP2012jh1d.dat$Q1d1 %>% fct_count()
# ISSP2012jh1d.dat$Q1e1 %>% fct_count()
# ISSP2012jh1d.dat$Q2a1 %>% fct_count()
# ISSP2012jh1d.dat$Q2b1 %>% fct_count()
# ISSP2012jh1d.dat$Q3a1 %>% fct_count()
#ISSP2012jh1d.dat$Q3b1 %>% fct_count()

# Taustamuuttujat - ristiintaulukoinnit riittävät (6.2.20)

# ISSP2012jh1d.dat$edu1 %>% fct_count()
# ISSP2012jh1d.dat$msta1 %>% fct_count()
# ISSP2012jh1d.dat$sosta1 %>% fct_count()
# ISSP2012jh1d.dat$nchild1 %>% fct_count()
# ISSP2012jh1d.dat$lifsta1 %>% fct_count()
# ISSP2012jh1d.dat$urbru1 %>% fct_count()
```

Taustamuuttujien luokitteluja (esim. luokkien yhdistäminen) pohditaan tarkemmin, kun muuttujat otetaan käyttöön.

Poistetaan muuuttujista luokittelumuuttujien arvot, joissa ei ole havaintoja. Näitä tyhjiä luokkia siirtyy SPSS-tiedostosta haven_labelled -luokan tietohin.

```
# Poistetaan tyhjät luokat muuttujista

ISSP2012jh1d.dat <- ISSP2012jh1d.dat %>%
    mutate(Q1a = fct_drop(Q1a1),
           Q1b = fct_drop(Q1b1),
           Q1c = fct_drop(Q1c1),
           Q1d = fct_drop(Q1d1),
           Q1e = fct_drop(Q1e1),
           Q2a = fct_drop(Q2a1),
           Q2b = fct_drop(Q2b1),
           Q3a = fct_drop(Q3a1),
           Q3b = fct_drop(Q3b1),
           edu = fct_drop(edu1),
           msta = fct_drop(msta1),
           sosta = fct_drop(sosta1),
           nchild = fct_drop(nchild1),
           lifsta = fct_drop(lifsta1),
           urbru = fct_drop(urbru1)

    )
# Tarkistuksia 1
```

```
ISSP2012jh1d.dat %>% summary()
```

V1	V2	DOI	V3	V4	C_ALPHAN	
Min. :5900	Length:32823	Length:32823	Min. : 36	Min. : 36.0	Length:32823	
1st Qu.:5900	Class :character	Class :character	1st Qu.: 208	1st Qu.:203.0	Class :character	1
Median :5900	Mode :character	Mode :character	Median : 428	Median :276.0	Mode :character	M
Mean :5900	NA	NA	Mean : 4063	Mean :362.1	NA	
3rd Qu.:5900	NA	NA	3rd Qu.: 705	3rd Qu.:578.0	NA	3
Max. :5900	NA	NA	Max. :62002	Max. :756.0	NA	
NA	NA	NA	NA	NA	NA	

```
ISSP2012jh1d.dat %>%
    select(Q1a, Q1b, Q1c, Q1d, Q1e,Q2a,Q2b,Q3a, Q3b) %>%
    str()
```

```
## tibble [32,823 x 9] (S3: tbl_df/tbl/data.frame)
##  $ Q1a: Factor w/ 5 levels "Strongly agree",..: 5 1 2 2 1 NA 2 4 2 2 ...
##   ..- attr(*, "label")= chr "Q1a Working mom: warm relationship with children as a not wo
##  $ Q1b: Factor w/ 5 levels "Strongly agree",..: 1 5 4 4 4 NA 4 3 4 3 ...
##   ..- attr(*, "label")= chr "Q1b Working mom: Preschool child is likely to suffer"
##  $ Q1c: Factor w/ 5 levels "Strongly agree",..: 3 5 2 4 4 NA 4 2 4 2 ...
##   ..- attr(*, "label")= chr "Q1c Working woman: Family life suffers when woman has full-t
##  $ Q1d: Factor w/ 5 levels "Strongly agree",..: 3 5 5 2 4 NA 4 5 4 5 ...
##   ..- attr(*, "label")= chr "Q1d Working woman: What women really want is home and kids"
##  $ Q1e: Factor w/ 5 levels "Strongly agree",..: 3 1 2 3 4 NA 2 4 4 1 ...
##   ..- attr(*, "label")= chr "Q1e Working woman: Being housewife is as fulfilling as worki
##  $ Q2a: Factor w/ 5 levels "Strongly agree",..: 1 3 4 2 2 NA 2 5 2 1 ...
##   ..- attr(*, "label")= chr "Q2a Both should contribute to household income"
##  $ Q2b: Factor w/ 5 levels "Strongly agree",..: 3 5 4 4 4 NA 2 5 4 1 ...
##   ..- attr(*, "label")= chr "Q2b Men's job earn money, women's job look after home"
##  $ Q3a: Factor w/ 3 levels "Work full-time",..: 3 NA NA 2 2 NA 2 NA 2 2 ...
##   ..- attr(*, "label")= chr "Q3a Should women work: Child under school age"
##  $ Q3b: Factor w/ 3 levels "Work full-time",..: 2 NA 2 1 2 NA 2 NA 2 2 ...
##   ..- attr(*, "label")= chr "Q3b Should women work: Youngest kid at school"
```

```
ISSP2012jh1d.dat %>%
    select(Q1a1, Q1b1, Q1c1, Q1d1, Q1e1,Q2a1,Q2b1,Q3a1, Q3b1) %>%
    str()
```

```
## tibble [32,823 x 9] (S3: tbl_df/tbl/data.frame)
##  $ Q1a1: Factor w/ 8 levels "NAP: ES","Strongly agree",..: 6 2 3 3 2 NA 3 5 3 3 ...
##   ..- attr(*, "label")= chr "Q1a Working mom: warm relationship with children as a not wo
##  $ Q1b1: Factor w/ 8 levels "NAP: ES","Strongly agree",..: 2 6 5 5 5 NA 5 4 5 4 ...
##   ..- attr(*, "label")= chr "Q1b Working mom: Preschool child is likely to suffer"
##  $ Q1c1: Factor w/ 8 levels "NAP: ES","Strongly agree",..: 4 6 3 5 5 NA 5 3 5 3 ...
```

```
##   ..- attr(*, "label")= chr "Q1c Working woman: Family life suffers when woman has full-t
##  $ Q1d1: Factor w/ 8 levels "NAP: ES","Strongly agree",..: 4 6 6 3 5 NA 5 6 5 6 ...
##   ..- attr(*, "label")= chr "Q1d Working woman: What women really want is home and kids"
##  $ Q1e1: Factor w/ 8 levels "NAP: ES","Strongly agree",..: 4 2 3 4 5 NA 3 5 5 2 ...
##   ..- attr(*, "label")= chr "Q1e Working woman: Being housewife is as fulfilling as worki
##  $ Q2a1: Factor w/ 8 levels "NAP: ES","Strongly agree",..: 2 4 5 3 3 NA 3 6 3 2 ...
##   ..- attr(*, "label")= chr "Q2a Both should contribute to household income"
##  $ Q2b1: Factor w/ 8 levels "NAP: ES","Strongly agree",..: 4 6 5 5 5 NA 3 6 5 2 ...
##   ..- attr(*, "label")= chr "Q2b Men's job earn money, women's job look after home"
##  $ Q3a1: Factor w/ 6 levels "Work full-time",..: 3 NA NA 2 2 NA 2 NA 2 2 ...
##   ..- attr(*, "label")= chr "Q3a Should women work: Child under school age"
##  $ Q3b1: Factor w/ 6 levels "Work full-time",..: 2 NA 2 1 2 NA 2 NA 2 2 ...
##   ..- attr(*, "label")= chr "Q3b Should women work: Youngest kid at school"
```

```
ISSP2012jh1d.dat %>%
    select(edu, msta, sosta, nchild,lifsta, urbru) %>%
    str()
```

```
## tibble [32,823 x 6] (S3: tbl_df/tbl/data.frame)
##  $ edu   : Factor w/ 7 levels "No formal education",..: 3 6 6 4 3 NA NA 7 6 7 ...
##   ..- attr(*, "label")= chr "Highest completed degree of education: Categories for interr
##  $ msta  : Factor w/ 9 levels "In paid work",..: 6 6 3 1 6 5 6 2 1 5 ...
##   ..- attr(*, "label")= chr "Main status"
##  $ sosta : Factor w/ 10 levels "Lowest, Bottom, 01",..: 3 7 8 NA 7 2 7 NA 10 6 ...
##   ..- attr(*, "label")= chr "Top-Bottom self-placement"
##  $ nchild: Factor w/ 11 levels "No children",..: NA NA 4 2 1 NA 1 1 2 NA ...
##   ..- attr(*, "label")= chr "How many children in household: children between [school age
##  $ lifsta: Factor w/ 6 levels "Married","Civil partnership",..: 6 1 1 6 1 6 1 1 1 NA ...
##   ..- attr(*, "label")= chr "Legal partnership status"
##  $ urbru : Factor w/ 5 levels "A big city","The suburbs or outskirts of a big city",..: 1
##   ..- attr(*, "label")= chr "Place of living: urban - rural"
```

```
ISSP2012jh1d.dat %>%
    select(edu1, msta1, sosta1, nchild1,lifsta1, urbru1) %>%
    str()
```

```
## tibble [32,823 x 6] (S3: tbl_df/tbl/data.frame)
##  $ edu1   : Factor w/ 8 levels "No formal education",..: 3 6 6 4 3 NA NA 7 6 7 ...
##   ..- attr(*, "label")= chr "Highest completed degree of education: Categories for interr
##  $ msta1  : Factor w/ 10 levels "In paid work",..: 6 6 3 1 6 5 6 2 1 5 ...
##   ..- attr(*, "label")= chr "Main status"
##  $ sosta1 : Factor w/ 14 levels "Not available: GB,US",..: 4 8 9 NA 8 3 8 NA 11 7 ...
##   ..- attr(*, "label")= chr "Top-Bottom self-placement"
##  $ nchild1: Factor w/ 14 levels "No children",..: NA NA 4 2 1 NA 1 1 2 NA ...
##   ..- attr(*, "label")= chr "How many children in household: children between [school age
##  $ lifsta1: Factor w/ 9 levels "Married","Civil partnership",..: 6 1 1 6 1 6 1 1 1 NA ...
##   ..- attr(*, "label")= chr "Legal partnership status"
```

```
##  $ urbru1 : Factor w/ 7 levels "A big city","The suburbs or outskirts of a big city",..:
##   ..- attr(*, "label")= chr "Place of living: urban - rural"
```

```
# Tarkistuksia 2 - ristiintaulukointi Q1a/Q1am riittää (6.2.20)

# Substanssimuuttujat

# ISSP2012jh1d.dat %>% tableX(Q1a,Q1a1)
# ISSP2012jh1d.dat %>% tableX(Q1b,Q1b1)
# ISSP2012jh1d.dat %>% tableX(Q1c,Q1c1)
# ISSP2012jh1d.dat %>% tableX(Q1d,Q1d1)
# ISSP2012jh1d.dat %>% tableX(Q1e,Q1e1)
# ISSP2012jh1d.dat %>% tableX(Q2a,Q2a1)
# ISSP2012jh1d.dat %>% tableX(Q2b,Q2b1)
# ISSP2012jh1d.dat %>% tableX(Q3a,Q3a1)
# ISSP2012jh1d.dat %>% tableX(Q3b,Q3b1)

# Taustamuuttujat

# ISSP2012jh1d.dat %>% tableX(edu,edu1)
# ISSP2012jh1d.dat %>% tableX(msta,msta1)
# ISSP2012jh1d.dat %>% tableX(sosta,sosta1)
# ISSP2012jh1d.dat %>% tableX(nchild,nchild1)
# ISSP2012jh1d.dat %>% tableX(lifsta,lifsta1)
# ISSP2012jh1d.dat %>% tableX(urbru,urbru1)
```

Luodaan uusi muuttuja, jossa puuttuva tieto (NA) on mukana luokittelumuuttujan uutena arvona (“missing”).

```
# Uusi muuttuja, jossa NA-arvot ovat mukana muuttujan uutena luokkana. Muuttujat
# nimetään Q1a -> Q1am.

ISSP2012jh1d.dat <- ISSP2012jh1d.dat %>%
    mutate(Q1am = fct_explicit_na(Q1a, na_level = "missing"),
           Q1bm = fct_explicit_na(Q1b, na_level = "missing"),
           Q1cm = fct_explicit_na(Q1c, na_level = "missing"),
           Q1dm = fct_explicit_na(Q1d, na_level = "missing"),
           Q1em = fct_explicit_na(Q1e, na_level = "missing"),
           Q2am = fct_explicit_na(Q2a, na_level = "missing"),
           Q2bm = fct_explicit_na(Q2b, na_level = "missing"),
           Q3am = fct_explicit_na(Q3a, na_level = "missing"),
           Q3bm = fct_explicit_na(Q3b, na_level = "missing"),
           edum = fct_explicit_na(edu, na_level = "missing"),
           mstam = fct_explicit_na(msta, na_level = "missing"),
           sostam = fct_explicit_na(sosta, na_level = "missing"),
           nchildm = fct_explicit_na(nchild, na_level = "missing"),
           lifstam = fct_explicit_na(lifsta, na_level = "missing"),
```

```
            urbrum = fct_explicit_na(urbru, na_level = "missing"),
            )
# Tarkistuksia 3

ISSP2012jh1d.dat %>%
    select(Q1am, Q1bm, Q1cm, Q1dm, Q1em, Q2am, Q2bm, Q3am, Q3bm) %>%
    summary()
```

Q1am	Q1bm	Q1cm	
Strongly agree :11116	Strongly agree :2747	Strongly agree :2838	
Agree :12352	Agree :8389	Agree :8263	
Neither agree nor disagree: 3382	Neither agree nor disagree:5949	Neither agree nor disagree:6000	Nei
Disagree : 4074	Disagree :9003	Disagree :8706	
Strongly disagree : 1051	Strongly disagree :5547	Strongly disagree :5960	
missing : 848	missing :1188	missing :1056	

```
ISSP2012jh1d.dat %>%
    select(edum,mstam, sostam,nchildm,lifstam, urbrum) %>%
    summary()
```

edum
Lower secondary (secondary completed does not allow entry to university: obligatory school) :7811
Upper secondary (programs that allows entry to university :7115
Post secondary, non-tertiary (other upper secondary programs toward labour market or technical formatio
Lower level tertiary, first stage (also technical schools at a tertiary level) :5147
Upper level tertiary (Master, Dr.) :4762
Primary school (elementary school) :1531
(Other) : 799

```
ISSP2012jh1d.dat %>%
    select(Q1am, Q1bm, Q1cm, Q1dm, Q1em, Q2am, Q2bm, Q3am, Q3bm) %>%
    str()
```

```
## tibble [32,823 x 9] (S3: tbl_df/tbl/data.frame)
##  $ Q1am: Factor w/ 6 levels "Strongly agree",..: 5 1 2 2 1 6 2 4 2 2 ...
##   ..- attr(*, "label")= chr "Q1a Working mom: warm relationship with children as a not wo
##  $ Q1bm: Factor w/ 6 levels "Strongly agree",..: 1 5 4 4 4 6 4 3 4 3 ...
##   ..- attr(*, "label")= chr "Q1b Working mom: Preschool child is likely to suffer"
##  $ Q1cm: Factor w/ 6 levels "Strongly agree",..: 3 5 2 4 4 6 4 2 4 2 ...
##   ..- attr(*, "label")= chr "Q1c Working woman: Family life suffers when woman has full-t
##  $ Q1dm: Factor w/ 6 levels "Strongly agree",..: 3 5 5 2 4 6 4 5 4 5 ...
##   ..- attr(*, "label")= chr "Q1d Working woman: What women really want is home and kids"
```

```
##  $ Q1em: Factor w/ 6 levels "Strongly agree",..: 3 1 2 3 4 6 2 4 4 1 ...
##   ..- attr(*, "label")= chr "Q1e Working woman: Being housewife is as fulfilling as worki
##  $ Q2am: Factor w/ 6 levels "Strongly agree",..: 1 3 4 2 2 6 2 5 2 1 ...
##   ..- attr(*, "label")= chr "Q2a Both should contribute to household income"
##  $ Q2bm: Factor w/ 6 levels "Strongly agree",..: 3 5 4 4 4 6 2 5 4 1 ...
##   ..- attr(*, "label")= chr "Q2b Men's job earn money, women's job look after home"
##  $ Q3am: Factor w/ 4 levels "Work full-time",..: 3 4 4 2 2 4 2 4 2 2 ...
##   ..- attr(*, "label")= chr "Q3a Should women work: Child under school age"
##  $ Q3bm: Factor w/ 4 levels "Work full-time",..: 2 4 2 1 2 4 2 4 2 2 ...
##   ..- attr(*, "label")= chr "Q3b Should women work: Youngest kid at school"
```

```
ISSP2012jh1d.dat %>%
    select(edum,mstam, sostam,nchildm,lifstam, urbrum) %>%
    str()
```

```
## tibble [32,823 x 6] (S3: tbl_df/tbl/data.frame)
##  $ edum   : Factor w/ 8 levels "No formal education",..: 3 6 6 4 3 8 8 7 6 7 ...
##   ..- attr(*, "label")= chr "Highest completed degree of education: Categories for interr
##  $ mstam  : Factor w/ 10 levels "In paid work",..: 6 6 3 1 6 5 6 2 1 5 ...
##   ..- attr(*, "label")= chr "Main status"
##  $ sostam : Factor w/ 11 levels "Lowest, Bottom, 01",..: 3 7 8 11 7 2 7 11 10 6 ...
##   ..- attr(*, "label")= chr "Top-Bottom self-placement"
##  $ nchildm: Factor w/ 12 levels "No children",..: 12 12 4 2 1 12 1 1 2 12 ...
##   ..- attr(*, "label")= chr "How many children in household: children between [school age
##  $ lifstam: Factor w/ 7 levels "Married","Civil partnership",..: 6 1 1 6 1 6 1 1 1 7 ...
##   ..- attr(*, "label")= chr "Legal partnership status"
##  $ urbrum : Factor w/ 6 levels "A big city","The suburbs or outskirts of a big city",..:
##   ..- attr(*, "label")= chr "Place of living: urban - rural"
```

```
# Taustamuuttuja, puuttuva tieto mukana - ristiintaulkointi riittää (6.2.20)

# ISSP2012jh1d.dat$edum %>% fct_count()
# ISSP2012jh1d.dat$mstam %>% fct_count()
# ISSP2012jh1d.dat$sostam %>% fct_count()
# ISSP2012jh1d.dat$nchildm %>% fct_count()
# ISSP2012jh1d.dat$lifstam %>% fct_count()
# ISSP2012jh1d.dat$urbrum %>% fct_count()

# Substanssimuuttujat, puuttuva tieto mukana  - ristiintaulkointi riittää (6.2.20)

# ISSP2012jh1d.dat$Q1am %>% fct_count()
# ISSP2012jh1d.dat$Q1bm %>% fct_count()
# ISSP2012jh1d.dat$Q1cm %>% fct_count()
# ISSP2012jh1d.dat$Q1dm %>% fct_count()
# ISSP2012jh1d.dat$Q1em %>% fct_count()
# ISSP2012jh1d.dat$Q2am %>% fct_count()
# ISSP2012jh1d.dat$Q2bm %>% fct_count()
```

```
# ISSP2012jh1d.dat$Q3am %>% fct_count()
# ISSP2012jh1d.dat$Q3bm %>% fct_count()
```

Lopuksi luodaan uuden "faktorilabelit" substanssimuuttujille. Graafisessa analyysissä kuviin on saatava mukaan kaikki oleellinen, mutta ei mitään sen lisäksi. Näitä muuttujan arvojen tunnuksia muokataan tarvittaessa.

```
# Vaihe 2.4.1

# Viisi vastausvaihtoehtoa - ei eksplisiittistä NA-tietoa("missing")
# Q3a - Q3b kolme vastausvaihtoehtoa

ISSP2012jh1d.dat <- ISSP2012jh1d.dat %>%
    mutate(Q1a = fct_recode(Q1a,
                        "S" = "Strongly agree",
                        "s" = "Agree",
                        "?" = "Neither agree nor disagree",
                        "e" = "Disagree",
                        "E"= "Strongly disagree"),
            Q1b = fct_recode(Q1b,
                       "S" = "Strongly agree",
                       "s" = "Agree",
                       "?" = "Neither agree nor disagree",
                       "e" = "Disagree",
                       "E" = "Strongly disagree"),
           Q1c = fct_recode(Q1c,
                            "S" = "Strongly agree",
                            "s" = "Agree",
                            "?" = "Neither agree nor disagree",
                            "e" = "Disagree",
                            "E" = "Strongly disagree"),
           Q1d = fct_recode(Q1d,
                            "S" = "Strongly agree",
                            "s" = "Agree",
                            "?" = "Neither agree nor disagree",
                            "e" = "Disagree",
                            "E" = "Strongly disagree"),
           Q1e = fct_recode(Q1e,
                            "S" = "Strongly agree",
                            "s" = "Agree",
                            "?" = "Neither agree nor disagree",
                            "e" = "Disagree",
                            "E" = "Strongly disagree"),
          Q2a = fct_recode(Q2a,
                            "S" = "Strongly agree",
                            "s" = "Agree",
```

```
                            "?" = "Neither agree nor disagree",
                            "e" = "Disagree",
                            "E" = "Strongly disagree" ),
          Q2b = fct_recode(Q2b,
                            "S" = "Strongly agree",
                            "s" = "Agree",
                            "?" = "Neither agree nor disagree",
                            "e" = "Disagree",
                            "E" = "Strongly disagree"),
          Q3a = fct_recode(Q3a,
                           "W" = "Work full-time",
                           "w" = "Work part-time",
                           "H" = "Stay at home" ),
          Q3b = fct_recode(Q3b,
                            "W" = "Work full-time",
                            "w" = "Work part-time",
                            "H" = "Stay at home" )
                        )


# Tarkistuksia 1
ISSP2012jh1d.dat %>%
    select(Q1a, Q1b, Q1c, Q1d, Q1e, Q2a, Q2b, Q3a, Q3b) %>%
    summary()
```

Q1a	Q1b	Q1c	Q1d	Q1e	Q2a	Q2b	Q3a	C
S :11116	S :2747	S :2838	S :2818	S :3357	S :11305	S :2704	W : 5373	W
s :12352	s :8389	s :8263	s :7672	s :8342	s :13464	s :5164	w :15655	w
? : 3382	? :5949	? :6000	? :7403	? :7841	? : 5039	? :6109	H : 8367	H
e : 4074	e :9003	e :8706	e :7863	e :7267	e : 1929	e :9210	NA's: 3428	NA'
E : 1051	E :5547	E :5960	E :5016	E :3462	E : 403	E :8917	NA	
NA's: 848	NA's:1188	NA's:1056	NA's:2051	NA's:2554	NA's: 683	NA's: 719	NA	

```
# Vaihe 2.4.2 - muuttujassa eksplisiittinen NA-tieto
ISSP2012jh1d.dat <- ISSP2012jh1d.dat %>%
    mutate(Q1am = fct_recode(Q1am,
                            "S" = "Strongly agree",
                            "s" = "Agree",
                            "?" = "Neither agree nor disagree",
                            "e" = "Disagree",
                            "E" = "Strongly disagree",
                            "P" = "missing"),
           Q1bm = fct_recode(Q1bm,
                            "S" = "Strongly agree",
```

```
                    "s" = "Agree",
                    "?" = "Neither agree nor disagree",
                    "e" = "Disagree",
                    "E" = "Strongly disagree",
                    "P" = "missing"),
         Q1cm = fct_recode(Q1cm,
                    "S" = "Strongly agree",
                    "s" = "Agree",
                    "?" = "Neither agree nor disagree",
                    "e" = "Disagree",
                    "E" = "Strongly disagree",
                    "P" = "missing"),
         Q1dm = fct_recode(Q1dm,
                    "S" = "Strongly agree",
                    "s" = "Agree",
                    "?" = "Neither agree nor disagree",
                    "e" = "Disagree",
                    "E" = "Strongly disagree",
                    "P" = "missing"),
         Q1em = fct_recode(Q1em,
                    "S" = "Strongly agree",
                    "s" = "Agree",
                    "?" = "Neither agree nor disagree",
                    "e" = "Disagree",
                    "E" = "Strongly disagree",
                    "P" = "missing"),
         Q2am = fct_recode(Q2am,
                     "S" = "Strongly agree",
                     "s" = "Agree",
                     "?" = "Neither agree nor disagree",
                     "e" = "Disagree",
                     "E" = "Strongly disagree",
                     "P" = "missing"),
         Q2bm = fct_recode(Q2bm,
                    "S" = "Strongly agree",
                    "s" = "Agree",
                    "?" = "Neither agree nor disagree",
                    "e" = "Disagree",
                    "E" = "Strongly disagree",
                    "P" = "missing"),
         Q3am = fct_recode(Q3am,
                    "W" = "Work full-time",
                    "w" = "Work part-time",
                    "H" = "Stay at home",
                    "P" = "missing"),
```

```
            Q3bm = fct_recode(Q3bm,
                              "W" = "Work full-time",
                              "w" = "Work part-time",
                              "H" = "Stay at home",
                              "P" = "missing")
                )

# Tarkistuksia 4

ISSP2012jh1d.dat %>%
    select(Q1am, Q1bm, Q1cm, Q1dm, Q1em, Q2am, Q2bm, Q3am, Q3bm) %>%
    summary()
```

Q1am	Q1bm	Q1cm	Q1dm	Q1em	Q2am	Q2bm	Q3am	Q3bm
S:11116	S:2747	S:2838	S:2818	S:3357	S:11305	S:2704	W: 5373	W:13722
s:12352	s:8389	s:8263	s:7672	s:8342	s:13464	s:5164	w:15655	w:13817
?: 3382	?:5949	?:6000	?:7403	?:7841	?: 5039	?:6109	H: 8367	H: 1762
e: 4074	e:9003	e:8706	e:7863	e:7267	e: 1929	e:9210	P: 3428	P: 3522
E: 1051	E:5547	E:5960	E:5016	E:3462	E: 403	E:8917	NA	NA
P: 848	P:1188	P:1056	P:2051	P:2554	P: 683	P: 719	NA	NA

```
# Tarkistuksia 5

# Substanssimuuttuja

ISSP2012jh1d.dat %>%
    tableX(Q1a,Q1am)
```

Q1a/Q1am	S	s	?	e	E	P	Total
1	11116	0	0	0	0	0	11116
2	0	12352	0	0	0	0	12352
3	0	0	3382	0	0	0	3382
4	0	0	0	4074	0	0	4074
5	0	0	0	0	1051	0	1051
Missing	0	0	0	0	0	848	848
Total	11116	12352	3382	4074	1051	848	32823

```
ISSP2012jh1d.dat %>%
    tableX(Q1b,Q1bm)
```

Q1b/Q1bm	S	s	?	e	E	P	Total
1	2747	0	0	0	0	0	2747
2	0	8389	0	0	0	0	8389
3	0	0	5949	0	0	0	5949
4	0	0	0	9003	0	0	9003
5	0	0	0	0	5547	0	5547
Missing	0	0	0	0	0	1188	1188
Total	2747	8389	5949	9003	5547	1188	32823

```
ISSP2012jh1d.dat %>%
    tableX(Q1c,Q1cm)
```

Q1c/Q1cm	S	s	?	e	E	P	Total
1	2838	0	0	0	0	0	2838
2	0	8263	0	0	0	0	8263
3	0	0	6000	0	0	0	6000
4	0	0	0	8706	0	0	8706
5	0	0	0	0	5960	0	5960
Missing	0	0	0	0	0	1056	1056
Total	2838	8263	6000	8706	5960	1056	32823

```
ISSP2012jh1d.dat %>%
    tableX(Q1d,Q1dm)
```

Q1d/Q1dm	S	s	?	e	E	P	Total
1	2818	0	0	0	0	0	2818
2	0	7672	0	0	0	0	7672
3	0	0	7403	0	0	0	7403
4	0	0	0	7863	0	0	7863
5	0	0	0	0	5016	0	5016
Missing	0	0	0	0	0	2051	2051
Total	2818	7672	7403	7863	5016	2051	32823

```
ISSP2012jh1d.dat %>%
    tableX(Q1e,Q1em)
```

Q1e/Q1em	S	s	?	e	E	P	Total
1	3357	0	0	0	0	0	3357
2	0	8342	0	0	0	0	8342
3	0	0	7841	0	0	0	7841
4	0	0	0	7267	0	0	7267

Q1e/Q1em	S	s	?	e	E	P	Total
5	0	0	0	0	3462	0	3462
Missing	0	0	0	0	0	2554	2554
Total	3357	8342	7841	7267	3462	2554	32823

```
ISSP2012jh1d.dat %>%
    tableX(Q2a,Q2am)
```

Q2a/Q2am	S	s	?	e	E	P	Total
1	11305	0	0	0	0	0	11305
2	0	13464	0	0	0	0	13464
3	0	0	5039	0	0	0	5039
4	0	0	0	1929	0	0	1929
5	0	0	0	0	403	0	403
Missing	0	0	0	0	0	683	683
Total	11305	13464	5039	1929	403	683	32823

```
ISSP2012jh1d.dat %>%
    tableX(Q2b,Q2bm)
```

Q2b/Q2bm	S	s	?	e	E	P	Total
1	2704	0	0	0	0	0	2704
2	0	5164	0	0	0	0	5164
3	0	0	6109	0	0	0	6109
4	0	0	0	9210	0	0	9210
5	0	0	0	0	8917	0	8917
Missing	0	0	0	0	0	719	719
Total	2704	5164	6109	9210	8917	719	32823

```
ISSP2012jh1d.dat %>%
    tableX(Q3a,Q3am)
```

Q3a/Q3am	W	w	H	P	Total
1	5373	0	0	0	5373
2	0	15655	0	0	15655
3	0	0	8367	0	8367
Missing	0	0	0	3428	3428
Total	5373	15655	8367	3428	32823

```
ISSP2012jh1d.dat %>%
    tableX(Q3b,Q3bm)
```

Q3b/Q3bm	W	w	H	P	Total
1	13722	0	0	0	13722
2	0	13817	0	0	13817
3	0	0	1762	0	1762
Missing	0	0	0	3522	3522
Total	13722	13817	1762	3522	32823

```
ISSP2012jh1d.dat %>%  # tableX muotoilee taulukkoa!
    tableX(Q3am,Q3a)
```

Q3am/Q3a	1	2	3	Missing	Total
W	5373	0	0	0	5373
w	0	15655	0	0	15655
H	0	0	8367	0	8367
P	0	0	0	3428	3428
Total	5373	15655	8367	3428	32823

```
ISSP2012jh1d.dat$Q3a %>% levels()
```

```
## [1] "W" "w" "H"
```

```
ISSP2012jh1d.dat$Q3am %>% levels()
```

```
## [1] "W" "w" "H" "P"
```

```
# Taustamuuttujat

ISSP2012jh1d.dat %>%
    tableX(edu, edum)
```

edu/edum	No formal education	Primary school (elementary school)	Lower secondary (secondary comple
1	491	0	0
2	0	1531	0
3	0	0	7811
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	0
Missing	0	0	0
Total	491	1531	7811

```
ISSP2012jh1d.dat %>%
    tableX(msta, mstam)
```

msta/mstam	In paid work	Unemployed and looking for a job, HR: incl never had a job	In education	A
1	17967	0	0	0
2	0	1769	0	0
3	0	0	1763	0
4	0	0	0	1
5	0	0	0	0
6	0	0	0	0
7	0	0	0	0
8	0	0	0	0
9	0	0	0	0
Missing	0	0	0	0
Total	17967	1769	1763	1

```
ISSP2012jh1d.dat %>%
    tableX(sosta, sostam)
```

sosta/sostam	Lowest, Bottom, 01	02	03	04	05	06	07	08	09	Highest, Top, 1
1	562	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	442
2	0	866	0	0	0	0	0	0	0	0
3	0	0	2221	0	0	0	0	0	0	0
4	0	0	0	3346	0	0	0	0	0	0
5	0	0	0	0	6798	0	0	0	0	0
6	0	0	0	0	0	6889	0	0	0	0
7	0	0	0	0	0	0	5778	0	0	0
8	0	0	0	0	0	0	0	3477	0	0
9	0	0	0	0	0	0	0	0	667	0
Missing	0	0	0	0	0	0	0	0	0	0
Total	562	866	2221	3346	6798	6889	5778	3477	667	442

```
ISSP2012jh1d.dat %>%
    tableX(nchild,nchildm)
```

nchild/nchildm	No children	One child	2 children	3	4	5	6	7	8	18	21 children	miss
1	24102	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	1	0	0
11	0	0	0	0	0	0	0	0	0	0	1	0
2	0	4378	0	0	0	0	0	0	0	0	0	0
3	0	0	2643	0	0	0	0	0	0	0	0	0

nchild/nchildm	No children	One child	2 children	3	4	5	6	7	8	18	21 children	miss
4	0	0	0	598	0	0	0	0	0	0	0	0
5	0	0	0	0	117	0	0	0	0	0	0	0
6	0	0	0	0	0	20	0	0	0	0	0	0
7	0	0	0	0	0	0	13	0	0	0	0	0
8	0	0	0	0	0	0	0	7	0	0	0	0
9	0	0	0	0	0	0	0	0	3	0	0	0
Missing	0	0	0	0	0	0	0	0	0	0	0	940
Total	24102	4378	2643	598	117	20	13	7	3	1	1	940

```
ISSP2012jh1d.dat %>%
    tableX(lifsta, lifstam)
```

lifsta/lifstam	Married	Civil partnership	Separated from spouse/ civil partner (still legally married/ still
1	17573	0	0
2	0	1035	0
3	0	0	486
4	0	0	0
5	0	0	0
6	0	0	0
Missing	0	0	0
Total	17573	1035	486

```
ISSP2012jh1d.dat %>%
    tableX(urbru, urbrum)
```

urbru/urbrum	A big city	The suburbs or outskirts of a big city	A town or a small city	A country villa
1	8442	0	0	0
2	0	4386	0	0
3	0	0	9203	0
4	0	0	0	8646
5	0	0	0	0
Missing	0	0	0	0
Total	8442	4386	9203	8646

2 Yksinkertainen korrespondenssianalyysi - kahden luokittelumuuttujan taulukko

Vanhaa jäsennystä

Yksinkertainen esimerkki, yksi kysymys (V6/Q1b) ja kuusi maata ristiintaulukoituna. Johdatteluna aiheeseen esitellään ca-käsitteet profiili, massa ja reunajakauma. Havainnollistetaan rivi- ja sarakeprofiilien vertailua vastaaviin keskiarvoprofiileihin.

Taulukoita tarkastella ensin rivien ja sitten sarakkeiden suhteen. Miten ne poikkeavat keskiarvostaan, miten toisistaan saman kategorian profiilista. Usein taulukoissa muuttujilla on selvästi eri rooli, kuten tässä. Koitan hahmottaa maiden (=aggregoituja yksilöitä) eroja ja yhtäläisyyksiä. Sarakkeiden vertailussa taas näemme, miten muuttujien profiilit poikkeavat keskiarvostaan. Monia riippuvuksia ja poikkeamia näyttäisi olevan. Klassinen ongelma, Pearson ja Fisher. Luokittelumuuttujien yhteys ("korrelaatio") on hankala juttu.

Riippumattomuushypoteesi ja χ^2 - riippumattomuustesti (pieni huomautus - on monta tapaa testata taulukon riippuvuuksia). Riippumattomuushypoteesi ehdollisena todennäköisyytenä reunajakauman suhteen. Riippumattomuustulkinta ei aina päde, jos aggregoidut havainnot/rivi-tai sarakeprofiilit/"samples" MG:n terminologiassa eivät ole riippumattomia. Esimerkki Barentsin merenpohjan lajiston havainnot (lukumäärät, "abundance") öljylauttojen liepeiltä (havainnot ryväksiä).

zxy Tämä puuttuu kaavoista!

Käsitteitä

1. Taulukko

Erityisesti CA, jossa "ranskalaisella terminologialla" käsitellään yksilöiden tai havaintoyksiköiden pilveä ja muuttujien pilveä . Taulukot saadaan yksinkertaisen CA:n tapauksessa aggregoimalla "cloud of individuals".

#V MOOC, LeReoux

2. Kontingenssitaulu (kts. viite, jossa ohje "yhteys aina riviä pitkin"), frekvenssitaulu, ristiintaulukointi

Dataa valitaan, aggregoidaan, ryhmitellään. Aktiivisia valintoja. Blasius emt. "data ei löydy kadulta", ja vaikka siitä ei ole epäilystäkään ISSP-datan tapauksessa, niin siitäkin jatketaan eteenpäin. (**edit 24.2.20** Epäselvä muistiinpano?)

Peruskäsitteiden yksinkertaisessa esityksessä tärkein lähde MG:n CAiP **#V** Siellä tästäkin on sananen: substanssiero usein on.

3. CA:ssa vaikea juttu on (Blasius, "vizualisation - verkkokirja") rivien ja sarakkeiden **tekninen** symmetria. No ei se nyt niin hämäävä ehkä ole,

oleellinen juttu (21.2.20). Kts. myös MG:n didaktiset esittelyt, skaalataan "hajontamittarilla" ja painotetaan massoilla.

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

Muutama versio tiiviiksi kuvaukseksi - toistoa on (10.4.20)

Dimensioiden vähentäminen tärkein asia ("the essence"), pienessä taulossa ei ihan ilmeinen. Esimerkin pienissä taulukoissa on toisaalta helppo katsoa datasta, mistä on kyse. Toinen tavoite on visualisointi, yleensä kaksiulotteisena kuvana (karttana). Kartta on metaforana hieman hankala. Kartalla esitetään kahden pistejoukon ("pilven") projektiot, jotka säilyttävät maksimimäärän alkuperäisen n-ulotteisen pistejoukon hajonnasta (inertiasta). Projektiossa lähekkäin olevat saman pilven pisteet voivat kuitenkin olla n-ulotteisessa pilvessä hyvinkin kaukana toisistaan. Tulkinnassa tärkeitä ovat "ääripäät", ja numeeriset tulokset kertovat kuinka hyvin piste on tasossa esitetty. Pisteiden väliset etäisyydet suhteellisia, ja eri pistejoukkojen välisillä etäisyyksillä ei ole suoraan mitään tulkintaa. Tämä ei oikein vastaa mielikuvaa kartasta, josta helposti näkee kuinka kaukan on Uudenmaan raja.

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".

Akseleiden tulkinta "ääripäiden" kautta ("kontrasti" ?). Huom "ääripää" ei välttämättä Likert-asteikolla tarkoita "äärimielipidettä", vaan se voi tarkoittaa myös selvää tai varmaa mielipidettä.(3.10.18).

Vanha lista - tehty jo

1. Ensimmäinen taulukko: profiilit, massat, keskiarvoprofiilit, khii2 - riippumattomuustesti ja etäisyysmitta
2. Hyvin tiivis esitys CA:n perusideasta, mutta ilman aivan simppeleitä kolmiulotteisia kuvia (niitä on jo).
3. Ensimmäinen symmetrinen kartta, perustulkinta (mitä kuvasta voidaan sanoa, mitä ei)
4. Lyhyt viittaus graafisen esityksen tulkintapulmiin, jotka eivät ole kovin pahoja. CA-kartta kaksoiskuvana (ts. informaatio voidaan palauttaa, skalaaritulo)?

5. Tulkinnan syventäminen - CA-käsitteiden tarkempi esittely

Haaste: käsitteet ja niiden suhteet ovat abstraktien matemaattisten rakenteiden tuloksia (barycentric, sentroidi), ja ne pitää jotenkin johdonmukaisesti pala kerrallaan tuoda esimerkkien kautta tekstiin. Käsittteistä oma Rmd (ja Excel jos osoittautuu kätevämmäksi), kaavaliite Dispo-repossa ja myös Rmd-muodossa.

edit(10.4.20): kaavaliitteen lisäksi voi tekstiin upottaa muutaman r-koodi-esimerkin

Ensimmäinen symmetrinen kartta

Tulkinnat ja yksinkertaisimmat perussäännöt. Dimensiot ja kuinka paljon alkuperäisen taulukon inertiaa saadaan esitettyä kartalla. Sitten asian ydin, akseleiden tulkinta ("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 sarakepisteidet. Mitä tarkoittavat prosentit akseleilla?

Varoitus virhetulkinnasta: ryhmien tunnistaminen rivi, myös pelkästään rivi- tai sarakepisteistä koostuvien ryhmien.

zxy Ja silti tavallaan voi. Sarake- ja rivipisteiden etäisyyksille ei ole suoraa tulkintaa, mutta on "vetovoima" (attraktio) ja "työntövoima" (repulsio). Jos profiilissa sarakemuuttujan osuus on suuri (siis suurempi kuin keskiarvopisteessä, suhteellinen ero), se "ajautuu" lähelle sarekepistettä. MG: "loose ends" - paperi, symmetrinen kuva eräs suurin sekaannuksen lähde. Tätä koitetaan selventää myös MG:n JASA-artikkelissa.

zxy(teoria/historia-jaksoon,104.20).Termi korrespondenssi: "neglected multivariate method" - paperissa käännetty näin englanniksi ransk. termi (Benzecri) rivien ja sarakkeiden "correspondence" eli yhteys/"riippuvuus"/vastaavuus tms.

2.1 Äiti työssä

zxy Perustellaan aineiston valinnan vaiheet. Esimerkiksi otetaan yksi kysymys.

zxy Suhde data-lukuun, siellä pitäisi esitellä aineisto sisällöllisesti. Tässä vain valitan esimerkkiä varten yksi kysymys ja kuusi maata.

Aineisto muuttujat Q1a-Q1e (arvot 1-5, täysin samaa mieltä - täysin eri mieltä) ovat vastauksia ensimmäiseen kysymyspatteriin (kts. lomake).

edit 10.4.20 Muuttujien "suunta" samaksi, jos monta. Laajemman aineiston käsittelyyn tästä huomautus.

(V6/Q1b) Alle kouluikäinen lapsi todennäköisesti kärsii, jos hänen äitinsä käy työssä. V6 muunnetaan uudeksi luokittelumuuttujaksi (R:ssä factor) Q1b. Tämä ei vielä tee kuvista ahtaita kun sarakkeita ja rivejä on vähän. Pudotetaan tarvittaessa turha Q-kirjain pois. Alkuperäisessä muuttujassa metatieto säilyy varmemmin, ja tarkistuksia on helpompi tehdä.

Valitaan esimerkin data edellisessä luvussa luodusta R-datasta ISSP2012jh1d.dat). Ihan yhtä hyvin voisi aina lukea suoraan alkuperäisestä spss-tiedostosta, mutta pidemmässä raportissa tämä on siistimpi tapa (23.3.2019). Kun havaintoja ja maita jätetään pois, uuteen dataan jää tyhjiä luokittelumuuttujien luokkia, ne poistetaan.

```
# UUSI DATA 30.1.20
#
# LUETAAN DATA G1_1_data2.Rmd - tiedostossa, luodaan faktorimuuttujat
# G1_1_data_fct1.Rmd-tiedostossa -> ISSP2012jh1d.dat (df)
# 23 muuttujaa (9 substanssimuuttujaa, 8 taustamuuttujaa, 3 maa-muuttujaa, 3 metadatamuuttu
# 25 maata.
# Poistettu 146 havaintoa, joilla SEX tai AGE puuttuu
# Johdattelevassa esimerkissä kuusi maata, kaksi taustamuuttujaa ja yksi kysymys
# (V6/Q1b)

# Kuusi maata

countries_esim1 <- c(56, 100, 208, 246, 276, 348) #BE,BG,DK,FI,DE,HU
ISSP2012esim3.dat <- filter(ISSP2012jh1d.dat, V4 %in% countries_esim1)
# str(ISSP2012esim3.dat) - pitkä listaus pois (24.2.20)

#neljä maamuuttujaa, kysymys Q1b, ikä ja sukupuoli

vars_esim1 <- c("C_ALPHAN", "V3", "maa","maa3", "Q1b", "sp", "ika")
ISSP2012esim2.dat <- select(ISSP2012esim3.dat, all_of(vars_esim1))

str(ISSP2012esim2.dat) # 8542 obs. of  7 variables
```

```
## tibble [8,542 x 7] (S3: tbl_df/tbl/data.frame)
##  $ C_ALPHAN: chr [1:8542] "BG" "BG" "BG" "BG" ...
##   ..- attr(*, "label")= chr "Country Prefix ISO 3166 Code - alphanumeric"
##   ..- attr(*, "format.spss")= chr "A20"
##   ..- attr(*, "display_width")= int 22
##  $ V3      : 'haven_labelled' num [1:8542] 100 100 100 100 100 100 100 100 100 100 ...
##   ..- attr(*, "label")= chr "Country/ Sample ISO 3166 Code (see V4 for codes for whole na
##   ..- attr(*, "labels")= Named num [1:45] 32 36 40 100 124 152 156 158 191 203 ...
##   .. ..- attr(*, "names")= chr [1:45] "AR-Argentina" "AU-Australia" "AT-Austria" "BG-Bulg
##  $ maa     : Factor w/ 25 levels "AU","AT","BG",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ maa3    : Factor w/ 29 levels "AU-Australia",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ Q1b     : Factor w/ 5 levels "S","s","?","e",..: 3 2 3 4 3 3 4 3 2 3 ...
##  $ sp      : Factor w/ 2 levels "m","f": 2 2 1 2 2 2 1 1 2 1 ...
##  $ ika     : 'haven_labelled' num [1:8542] 64 43 63 31 52 46 51 40 57 64 ...
##   ..- attr(*, "label")= chr "Age of respondent"
##   ..- attr(*, "labels")= Named num [1:6] 15 16 17 18 102 999
##   .. ..- attr(*, "names")= chr [1:6] "15 years" "16 years" "17 years" "18 years" ...
```

```r
# C_ALPHAN: chr, maa: Factor w/ 25

# Poistetaan havainnot, joilla Q1b - muuttujassa puuttuva tieto 'NA'

ISSP2012esim1.dat <- filter(ISSP2012esim2.dat, !is.na(Q1b))

str(ISSP2012esim1.dat) # 8143 obs. of  6 variable
```

```
## tibble [8,143 x 7] (S3: tbl_df/tbl/data.frame)
##  $ C_ALPHAN: chr [1:8143] "BG" "BG" "BG" "BG" ...
##   ..- attr(*, "label")= chr "Country Prefix ISO 3166 Code - alphanumeric"
##   ..- attr(*, "format.spss")= chr "A20"
##   ..- attr(*, "display_width")= int 22
##  $ V3      : 'haven_labelled' num [1:8143] 100 100 100 100 100 100 100 100 100 100 ...
##   ..- attr(*, "label")= chr "Country/ Sample ISO 3166 Code (see V4 for codes for whole na
##   ..- attr(*, "labels")= Named num [1:45] 32 36 40 100 124 152 156 158 191 203 ...
##   .. ..- attr(*, "names")= chr [1:45] "AR-Argentina" "AU-Australia" "AT-Austria" "BG-Bulg
##  $ maa     : Factor w/ 25 levels "AU","AT","BG",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ maa3    : Factor w/ 29 levels "AU-Australia",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ Q1b     : Factor w/ 5 levels "S","s","?","e",..: 3 2 3 4 3 3 4 3 2 3 ...
##  $ sp      : Factor w/ 2 levels "m","f": 2 2 1 2 2 2 1 1 2 1 ...
##  $ ika     : 'haven_labelled' num [1:8143] 64 43 63 31 52 46 51 40 57 64 ...
##   ..- attr(*, "label")= chr "Age of respondent"
##   ..- attr(*, "labels")= Named num [1:6] 15 16 17 18 102 999
##   .. ..- attr(*, "names")= chr [1:6] "15 years" "16 years" "17 years" "18 years" ...
```

```r
# Tarkistuksia - miksi nämä eivät tulosta mitään? (3.2.20)

fct_count(ISSP2012esim1.dat$sp) %>% table1()
```

```
##
## -------------------------
##       Mean/Count (SD/%)
##       n = 2
##  f
##     m 1 (50%)
##     f 1 (50%)
##  n
##       4071.5 (385.4)
## -------------------------
```

```r
fct_count(ISSP2012esim1.dat$Q1b)
```

f	n
S	810
s	1935

f	n
?	1367
e	2125
E	1906

```
fct_count(ISSP2012esim1.dat$maa)
```

f	n
AU	0
AT	0
BG	921
CA	0
HR	0
CZ	0
DK	1388
FI	1110
FR	0
HU	997
IS	0
IE	0
LV	0
LT	0
NL	0
NO	0
PL	0
RU	0
SK	0
SI	0
SE	0
CH	0
BE	2013
DE	1714
PT	0

```
fct_count(ISSP2012esim1.dat$maa3)
```

f	n
AU-Australia	0
AT-Austria	0
BG-Bulgaria	921
CA-Canada	0
HR-Croatia	0

f	n
CZ-Czech Republic	0
DK-Denmark	1388
FI-Finland	1110
FR-France	0
HU-Hungary	997
IS-Iceland	0
IE-Ireland	0
LV-Latvia	0
LT-Lithuania	0
NL-Netherlands	0
NO-Norway	0
PL-Poland	0
RU-Russia	0
SK-Slovakia	0
SI-Slovenia	0
SE-Sweden	0
CH-Switzerland	0
BE-FLA-Belgium/ Flanders	1012
BE-WAL-Belgium/ Wallonia	490
BE-BRU-Belgium/ Brussels	511
DE-W-Germany-West	1167
DE-E-Germany-East	547
PT-Portugal 2012: first fieldwork round (main sample)	0
PT-Portugal 2012: second fieldwork round (complementary sample)	0

```
# Toimivat tarkistukset (3.2.20)
summary(ISSP2012esim1.dat$sp)
```

```
##    m    f
## 3799 4344
```

```
#sp: 3799 + 4344 = 8143
summary(ISSP2012esim1.dat$Q1b)
```

```
##    S    s    ?    e    E
##  810 1935 1367 2125 1906
```

```
#  S      s      ?     e      E
# 810 + 1935 + 1367 + 2125 + 1906 = 8143


# EDELLINEN DATA - havaintojen määrät samat kuin uudella datalla (31.1.20)
#
# 8557 obs. ennen kuin sexagemissing poistettiin, nyt 8542, 8557-8542 = 15
#
```

```
# Poistetaan havainnot joissa puuttuva tieto muuttujassa V6 (Q1b) n = 399
# 8542-399 = 8143

# Tyhjät "faktorilabelit" on poistettava

 ISSP2012esim1.dat <- ISSP2012esim1.dat %>%
     mutate(maa = fct_drop(maa),
            maa3 = fct_drop(maa3)
            )

summary(ISSP2012esim1.dat$maa)
```

```
##   BG   DK   FI   HU   BE   DE
##  921 1388 1110  997 2013 1714
```

```
summary(ISSP2012esim1.dat$maa3)
```

```
##               BG-Bulgaria                DK-Denmark                FI-Finland
##                       921                      1388                      1110
##               HU-Hungary BE-FLA-Belgium/ Flanders BE-WAL-Belgium/ Wallonia
##                       997                      1012                       490
## BE-BRU-Belgium/ Brussels        DE-W-Germany-West        DE-E-Germany-East
##                       511                      1167                       547
```

```
# str(ISSP2012esim1.dat$maa)
# attributes(ISSP2012esim1.dat$maa)

# str(ISSP2012esim1.dat$maa3)
# attributes(ISSP2012esim1.dat$maa3)


ISSP2012esim1.dat %>% tableX(maa, Q1b, type = "count")
```

maa/Q1b	S	s	?	e	E	Total
BG	118	395	205	190	13	921
DK	70	238	152	232	696	1388
FI	47	188	149	423	303	1110
HU	219	288	225	190	75	997
BE	191	451	438	552	381	2013
DE	165	375	198	538	438	1714
Total	810	1935	1367	2125	1906	8143

```
fct_count(ISSP2012esim1.dat$Q1b)
```

f	n
S	810
s	1935
?	1367
e	2125
E	1906

```
# fct_count(ISSP2012esim1.dat$sp)
# fct_unique(ISSP2012esim1.dat$maa)
# fct_count(ISSP2012esim1.dat$maa)
ISSP2012esim1.dat %>% tableX(maa, C_ALPHAN, type = "count")
```

maa/C_ALPHAN	BE	BG	DE	DK	FI	HU	Total
BG	0	921	0	0	0	0	921
DK	0	0	0	1388	0	0	1388
FI	0	0	0	0	1110	0	1110
HU	0	0	0	0	0	997	997
BE	2013	0	0	0	0	0	2013
DE	0	0	1714	0	0	0	1714
Total	2013	921	1714	1388	1110	997	8143

```
# maa3 - siistitään "faktorilabelit" kaksikirjaimisiksi
#
# ISO 3166 Code V3 - maiden jaot
#  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
# Tähän pitäisi päästä
# levels = c("100","208","246","348","5601","5602","5603","27601","27602"),
#  labels = c("BG","DK","FI","HU","bF","bW","bB","dW","dE"))
levels(ISSP2012esim1.dat$maa3)
```

```
## [1] "BG-Bulgaria"              "DK-Denmark"
## [3] "FI-Finland"               "HU-Hungary"
## [5] "BE-FLA-Belgium/ Flanders" "BE-WAL-Belgium/ Wallonia"
## [7] "BE-BRU-Belgium/ Brussels" "DE-W-Germany-West"
## [9] "DE-E-Germany-East"
```

```
ISSP2012esim1.dat <- ISSP2012esim1.dat %>%
        mutate(maa3 =
                fct_recode(maa3,
```

```
                 "BG" = "BG-Bulgaria",
                 "DK" = "DK-Denmark",
                 "FI" = "FI-Finland",
                 "HU" = "HU-Hungary",
                 "bF" = "BE-FLA-Belgium/ Flanders",
                 "bW" = "BE-WAL-Belgium/ Wallonia",
                 "bB" = "BE-BRU-Belgium/ Brussels",
                 "dW" = "DE-W-Germany-West",
                 "dE" = "DE-E-Germany-East")
               )
# tarkistuksia
levels(ISSP2012esim1.dat$maa3)
```

```
## [1] "BG" "DK" "FI" "HU" "bF" "bW" "bB" "dW" "dE"
```

```
# str(ISSP2012esim1.dat$maa3) # 9 levels
summary(ISSP2012esim1.dat$maa3)
```

```
##   BG   DK   FI   HU   bF   bW   bB   dW   dE
##  921 1388 1110  997 1012  490  511 1167  547
```

```
# TÄSSÄ TOISTOA! (4.2.20)

# Muutetaan muuttujien "maa" ja "maa3" arvojen (levels) järjestys samaksi kuin alkuperäisen
# muuttujan C_ALPHAN. Helpomi verrata aikaisempiin tuloksiin.

# maa samaan järjestukseen kuin C_ALPHAN - olisiko aakkosjärjestys?
# tämä vain siksi, että muuten  esimerkin ca-kartta "kääntyy"
# "vanha" maa-muuttuja talteen - ei ehkä tarpeen? (4.2.20)

ISSP2012esim1.dat$maa2 <- ISSP2012esim1.dat$maa # "alkuperäinen" maa talteen

ISSP2012esim1.dat <- ISSP2012esim1.dat %>%
        mutate(maa =
                fct_relevel(maa,
                            "BE",
                            "BG",
                            "DE",
                            "DK",
                            "FI",
                            "HU"))
ISSP2012esim1.dat <- ISSP2012esim1.dat %>%
        mutate(maa3 =
                fct_relevel(maa3,
                        "bF",
                        "bW",
                        "bB",
```

```
                        "BG",
                        "dW",
                        "dE",
                        "DK",
                        "FI",
                        "HU"))

# Tarkistus
ISSP2012esim1.dat %>% tableX(maa2,maa, type = "count")
```

maa2/maa	BE	BG	DE	DK	FI	HU	Total
BG	0	921	0	0	0	0	921
DK	0	0	0	1388	0	0	1388
FI	0	0	0	0	1110	0	1110
HU	0	0	0	0	0	997	997
BE	2013	0	0	0	0	0	2013
DE	0	0	1714	0	0	0	1714
Total	2013	921	1714	1388	1110	997	8143

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

maa/C_ALPHAN	BE	BG	DE	DK	FI	HU	Total
BE	2013	0	0	0	0	0	2013
BG	0	921	0	0	0	0	921
DE	0	0	1714	0	0	0	1714
DK	0	0	0	1388	0	0	1388
FI	0	0	0	0	1110	0	1110
HU	0	0	0	0	0	997	997
Total	2013	921	1714	1388	1110	997	8143

```
str(ISSP2012esim1.dat)
```

```
## tibble [8,143 x 8] (S3: tbl_df/tbl/data.frame)
##  $ C_ALPHAN: chr [1:8143] "BG" "BG" "BG" "BG" ...
##   ..- attr(*, "label")= chr "Country Prefix ISO 3166 Code - alphanumeric"
##   ..- attr(*, "format.spss")= chr "A20"
##   ..- attr(*, "display_width")= int 22
##  $ V3      : 'haven_labelled' num [1:8143] 100 100 100 100 100 100 100 100 100 100 ...
##   ..- attr(*, "label")= chr "Country/ Sample ISO 3166 Code (see V4 for codes for whole na
##   ..- attr(*, "labels")= Named num [1:45] 32 36 40 100 124 152 156 158 191 203 ...
##   .. ..- attr(*, "names")= chr [1:45] "AR-Argentina" "AU-Australia" "AT-Austria" "BG-Bulg
##  $ maa     : Factor w/ 6 levels "BE","BG","DE",..: 2 2 2 2 2 2 2 2 2 2 ...
```

```
##  $ maa3    : Factor w/ 9 levels "bF","bW","bB",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Q1b     : Factor w/ 5 levels "S","s","?","e",..: 3 2 3 4 3 3 4 3 2 3 ...
##  $ sp      : Factor w/ 2 levels "m","f": 2 2 1 2 2 2 1 1 2 1 ...
##  $ ika     : 'haven_labelled' num [1:8143] 64 43 63 31 52 46 51 40 57 64 ...
##   ..- attr(*, "label")= chr "Age of respondent"
##   ..- attr(*, "labels")= Named num [1:6] 15 16 17 18 102 999
##   .. ..- attr(*, "names")= chr [1:6] "15 years" "16 years" "17 years" "18 years" ...
##  $ maa2    : Factor w/ 6 levels "BG","DK","FI",..: 1 1 1 1 1 1 1 1 1 1 ...
```

TODO (1) Taulukot erotettava omiksi koodilohkoiksi bookdowniin. (2) Ikä - maa - taulukko vain tarkistuksiin, ihan liian pitkä.

```
# Taulukoita (31.1.2020) ja tarkistuksia

# toinen maa-muuttuja, jossa Saksan ja Belgian jako
#  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

# Tarkastuksia

# assert_that ehkä tarpeeton - expect_equivalet testaa levelien
# järjestyksen ja määrän (20.2.20)

assert_that(length(levels(ISSP2012esim1.dat$maa)) == 6)
```

```
## [1] TRUE
```

```
assert_that(length(levels(ISSP2012esim1.dat$maa3)) == 9)
```

```
## [1] TRUE
```

```
assert_that(length(levels(ISSP2012esim1.dat$Q1b)) == 5)
```

```
## [1] TRUE
```

```
# expect_ ei anna ok-ilmoitusta, ainoastaan virheilmoituksen? (11.4.20)

expect_equivalent(levels(ISSP2012esim1.dat$maa),
                  c("BE", "BG", "DE", "DK", "FI", "HU"))

expect_equivalent(levels(ISSP2012esim1.dat$maa3),
                  c("bF","bW","bB", "BG","dW","dE","DK", "FI", "HU"))

expect_equivalent(levels(ISSP2012esim1.dat$sp), c("m", "f"))
```

```
expect_equivalent(levels(ISSP2012esim1.dat$Q1b),
                  c("S", "s", "?", "e", "E"))

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

maa/ika	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
BE	0.00	0.00	0.00	0.79	1.29	1.24	1.19	1.54	1.49	1.34	1.44	1.19	1.69	1.34	1.39
BG	0.00	0.00	0.00	1.41	1.09	0.98	0.98	0.76	1.19	0.76	1.19	1.41	0.98	0.98	1.09
DE	0.00	0.00	0.00	1.11	1.69	1.23	1.58	1.40	1.93	1.46	1.58	1.17	1.40	1.23	1.46
DK	0.00	0.00	0.00	1.73	1.30	1.30	2.23	2.52	2.74	1.95	1.15	1.08	1.73	1.37	1.44
FI	0.72	1.80	1.17	1.62	1.08	1.35	1.17	0.63	1.26	1.53	1.35	1.44	1.26	1.17	2.61
HU	0.00	0.00	0.00	0.90	1.20	1.00	0.80	1.91	1.91	1.10	1.50	1.00	1.40	1.30	1.91
All	0.10	0.25	0.16	1.22	1.31	1.20	1.38	1.51	1.78	1.40	1.39	1.20	1.46	1.25	1.61

```
# Riviprofiilit

# ISSP2012esim1.dat %>% tableX(maa,ika,type = "row_perc")
ISSP2012esim1.dat %>% tableX(maa,sp ,type = "row_perc")
```

maa/sp	m	f	Total
BE	47.44	52.56	100.00
BG	40.72	59.28	100.00
DE	48.66	51.34	100.00
DK	49.42	50.58	100.00
FI	42.88	57.12	100.00
HU	47.44	52.56	100.00
All	46.65	53.35	100.00

```
# Kysymyksen Q1b vastaukset

ISSP2012esim1.dat %>% tableX(maa,Q1b,type = "row_perc")
```

maa/Q1b	S	s	?	e	E	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

```
# Kuuluu ehkä vasta seuraavaan jaksoon ? (20.2.20)

ISSP2012esim1.dat %>% tableX(maa3,Q1b,type = "row_perc")
```

maa3/Q1b	S	s	?	e	E	Total
bF	5.04	23.81	25.89	30.83	14.43	100.00
bW	10.82	21.02	18.57	24.08	25.51	100.00
bB	17.03	20.94	16.63	23.87	21.53	100.00
BG	12.81	42.89	22.26	20.63	1.41	100.00
dW	11.40	26.82	11.83	32.13	17.82	100.00
dE	5.85	11.33	10.97	29.80	42.05	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

```
# str(ISSP2012esim1.dat) # 8143 obs. of  7 variable,
# sama kuin vanhassa Galku-koodissa.
```

Taulukot ja kuvat omina koodilohkoina bookdown-versioon

Frekvenssitaulukko

```
# Esimerkki - siisti tuloste (20.2.20)

taulu2 <- ISSP2012esim1.dat %>% tableX(maa, Q1b, type = "count")
knitr::kable(taulu2,digits = 2, booktabs = TRUE,
             caption = "Kysymyksen Q1b vastaukset maittain")
```

Taulukko 51: Kysymyksen Q1b vastaukset maittain

	S	s	?	e	E	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

Riviprosentit

```
taulu3 <- ISSP2012esim1.dat %>% tableX(maa,Q1b,type = "row_perc")
```

```
knitr::kable(taulu3,digits = 2, booktabs = TRUE,
             caption = "Kysymyksen Q1b vastaukset, riviprosentit")
```

Taulukko 52: Kysymyksen Q1b vastaukset, riviprosentit

	S	s	?	e	E	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

Sarakeprosentit

```
taulu4 <- ISSP2012esim1.dat %>% tableX(maa,Q1b,type = "col_perc")

knitr::kable(taulu4,digits = 2, booktabs = TRUE,
             caption = "Kysymyksen Q1b vastaukset, sarakeprosentit")
```

Taulukko 53: Kysymyksen Q1b vastaukset, sarakeprosentit

	S	s	?	e	E	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 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 24.2.20** Lisätty ca-karttoihin versio, jossa maiden painot samat (= 1). Esimerkkilaskelmia CAcalc_1.R.

```
# CA tässä, jotta saadaan rivi- ja sarakeprofiilikuvat

simpleCA1 <- ca(~maa + Q1b,ISSP2012esim1.dat)

# Maiden järjestys kääntää kuvan (1.2.20) - esimerkki on
# vähän kuriositeetti. Kartta voi tietysti "flipata" koordintaattien suhteen ainakin
# neljällä tavalla (? 180 astetta molempien akseleiden ympäri molempiin suuntiin?)
# (18.2.20). Tämän maa2-muuttujaa käyttävän kuvan voi jättää pois (8.4.20)

# simpleCA2 <- ca(~maa2 + Q1b,ISSP2012esim1.dat)

# Oikeastaan maiden vertailussa pitäisi niiden massat skaalata yhtä suuriksi, tässä
# pikainen kokeilu (20.2.20)
# Riviprosentit taulukoksi, nimet sarakkeille ja riveille (ei kovin robustia...)


johdesim1_rowproc.tab <- simpleCA1$N / rowSums(simpleCA1$N)
colnames(johdesim1_rowproc.tab) <- c("S" ,"s" ,"?","e", "E")
rownames(johdesim1_rowproc.tab) <- c("BE", "BG", "DE", "DK", "FI", "HU")

# Miten tibblenä? Ei toimi, ei maa-muuttujaa ollenkaan
# johdesim1_rowproc.tbl <- as_tibble(johdesim1_rowproc.tab)
# str(johdesim1_rowproc.tbl)

# TARKISTUKSIA (20.2.20)
# johdesim1_rowproc.tab
# rowSums(johdesim1_rowproc.tab)
# str(johdesim1_rowproc.tab)


simpleCA3 <- ca(johdesim1_rowproc.tab)


# Kartta piirretään koodilohkossa simpleCAmap1, mikä Rmd-tiedosto? #T(11.4.20)

# Riviprosentit tarkistusta varten
#        S  s   ?   e   E
#BE 9.49    22.40   21.76   27.42   18.93
#BG 12.81   42.89   22.26   20.63   1.41
#DE 9.63    21.88   11.55   31.39   25.55
#DK 5.04    17.15   10.95   16.71   50.14
#FI 4.23    16.94   13.42   38.11   27.30
#HU 21.97   28.89   22.57   19.06   7.52
#
# Ja datan saa leikepöydän kautta, jos on tarve pikatarkistuksiin
```

```
# read <- read.table("clipboard")

# 9.4.2020 CAcalc_1.R - laskentoa ca-funktion tuloksilla (16 objektin lista)
```

TODO 2.2.20

Onko tämä kuva tallennettava kuvatiedostoksi, vai onnistuuko sen tuottaminen Bookdownissa. Ei taida onnistua? (4.9.18)

Sarakeprofiilit, oikea järjestys maa-muuttujan tasoilla. Faktoreiden järjestys voi tuottaa yllätyksiä, kun dataa muokataan ggplot - grafiikaksi.

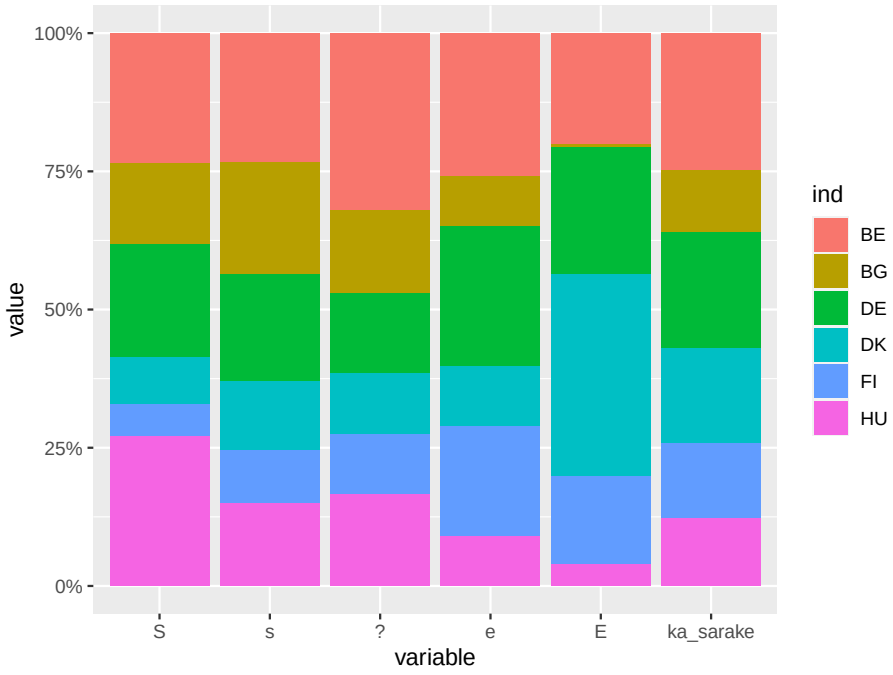
```
#mutkikas kuvan piirto - sarakeprofiilit vertailussa
#ggplot vaatii df-rakenteen ja 'long data' - muotoon
##https://stackoverflow.com/questions/9563368/create-stacked-barplot-where-each-stack-is-sc
#
# käytetään ca - tuloksia
apu1 <- (simpleCA1$N)
colnames(apu1) <- c("S", "s", "?", "e", "E")
rownames(apu1) <- c("BE", "BG", "DE", "DK", "FI", "HU")
apu1_df <- as.data.frame(apu1)
#lasketan rivien reunajakauma
apu1_df$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())
```

```
# apu1_df
# apu1b_df
```

Testaus: maa2, eri järjestys kuin C_ALPHAN (joka oli käytössä vanhemmissa Galku-versioissa) **edit** pois tulosteesta 30.3.20.

TODO 2.2.20 Voisi harkita taulukoiden (rivi- ja sarakeprosentit) sijoittamista kuvien viereen?

```
# riviprofiilit ja keskiarvorivi -  18.9.2018
apu2_df <- as.data.frame(apu1)
apu2_df <- rbind(apu2_df, ka_rivi = colSums(apu2_df))

#apu2_df
#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)
```



Kuva 2: Q1b:Sarakeprofiilit ja keskiarvoprofiili

```
## length(apu2_df) # how long is it? What about two dimensional
## objects?
# attributes(apu2_df)

# temp1 <- cbind(apu2_df, ind = rownames(apu2_df))
# temp1
##muokataan 'long data' - muotoon
apu2b_df <- melt(cbind(apu2_df, ind = rownames(apu2_df)), id.vars = c('ind'))
# str(apu2b_df)
# glimpse(apu2b_df)

#
#ggplot(apu2b_df, aes(x = value, y = ind, fill = variable)) +
#       geom_bar(position = "fill", stat ="identity") +
#       #coord_flip() +
#        scale_x_continuous(labels = percent_format())

#versio2 toimii (18.9.2018)

ggplot(apu2b_df, aes(x = ind, y = value, fill = variable)) +
       geom_bar(position = "fill", stat = "identity") +
       coord_flip() +
        scale_y_continuous(labels = percent_format())
```

Ja sama testaus kuin sarakeprofiilikuvilla. **edit** Pois tulosteesta 30.3.20

Graafinen analyysi ja R

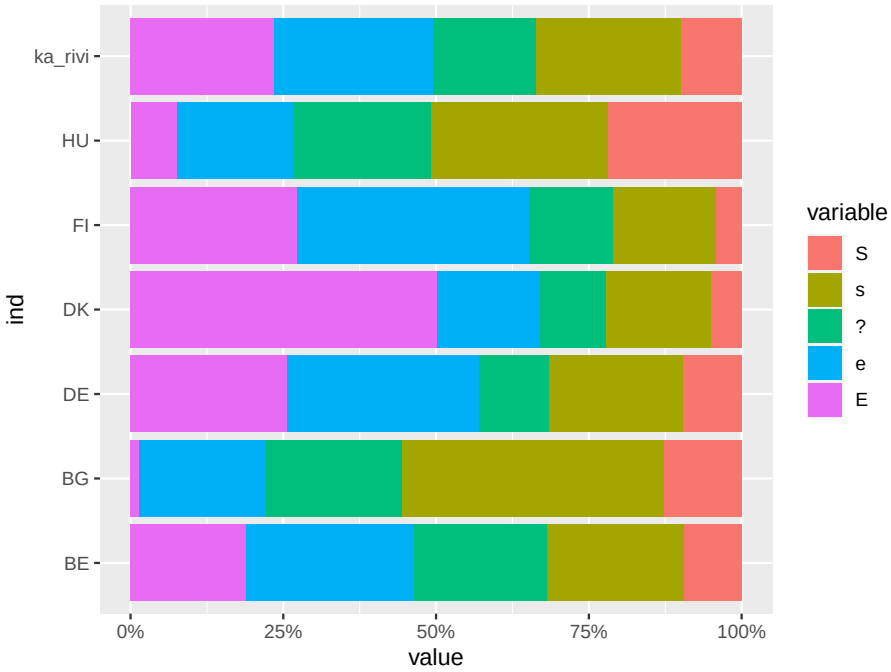
Käytänön neuvoja data-analyysiin, kuulunee tekstiin vai meneekö "ohjelmisto-ympäristö" -liitteeseen? Tärkeä juttu!

Kuvasuhteen saa oikeaksi, kun avaa g-ikkunan (X11()) ja sitten plot. Voi tallentaa pdf-muodossa grafiikkaikkunasta, ja ladata outputiin knitr-vaiheessa. Parempi tulostaa kuvatdsto pdf-ajurilla, jos lopulliseen versioon joutuu näin tekemään (13.5.2018). Tämä voi olla järkevä tapa analyysivaiheessa? Teksti kopsattu alla olevasta koodilohkosta.

Ensimmäinen korrespondenssianalyysi - kokeiluja kuvasuhteen säätämiseksi output- dokumentissa. RStudiossa voi avata komentokehoitteessa grafiikka-ikkunan. Siitä käsin tallennettu pdf-kuva on ladattu alla Rmarkdownin omalla komennolla, kohdistus keskelle. Parhaiten näyttäisi toimivan knitrin funktio, mutta oletuskuvakoolla saa ca-kuvasta näköjään aika lähelle oikeanlaisen ilman mitään temppuja.

zxy Selventäisikö vielä khii2-etäisyyksien taulukko, tai ehkä seuraavassa luvussa? **#V** MG&Blasius, "vihreän kirja", johdanto.

Rivien (1) ja sarakkeiden (2) khii2-etäisyydet keskiarvosta. TODO



Kuva 3: Q1b: riviprofiilit ja keskiarvorivi

19.2.20 Siistiksi taulukoksi. as_tibble antaa varoituksen, mutta toimii (11.4.20).

```
# khii2 - etäisyyksien taulukko
#str(simpleCA1)
#simpleCA1$rowdist
#str(simpleCA1$rowdist)


simpleCA1$rowdist
```

```
## [1] 0.1579735 0.6309909 0.1750128 0.6340627 0.3477331 0.5504040
```

```
rowdist.tbl <- as_tibble(rbind(simpleCA1$rowdist))
```

```
## Warning: The `x` argument of `as_tibble.matrix()` must have column names if `.name_repair
## Using compatibility `.name_repair`.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
colnames(rowdist.tbl) <- simpleCA1$rownames
# rowdist.tbl <- table(simpleCA1$rowdist)
# rowdist.tbl <- names(simpleCA1$rownames)
# str(rowdist.tbl)
# print(rowdist.tbl)

simpleCA1$coldist
```

```
## [1] 0.5246525 0.3248840 0.3078230 0.2721699 0.6271108
```

```
coldist.tbl <- as_tibble(rbind(simpleCA1$coldist))
colnames(coldist.tbl) <- simpleCA1$colnames
# print(coldist.tbl)
knitr::kable(rowdist.tbl,digits = 3, booktabs = TRUE,
             caption = "Rivietäisyydet keskiarvosta (khii2)")
```

Taulukko 54: Rivietäisyydet keskiarvosta (khii2)

BE	BG	DE	DK	FI	HU
0.158	0.631	0.175	0.634	0.348	0.55

```
knitr::kable(coldist.tbl,digits = 3, booktabs = TRUE,
             caption = "Sarake-etäisyydet keskiarvosta (khii2)")
```

Taulukko 55: Sarake-etäisyydet keskiarvosta (khii2)

S	s	?	e	E
0.525	0.325	0.308	0.272	0.627
TODO	Rivi- ja	sarake-	etäisyyd	et (khii2) vaatisivat kommentin. Tässä näytetään,
että muu	ttu			

CA-ratkaisun lähtötieto: suhteelliset frekvenssit (korrespondenssimatriisi P) **edit** Algoritmin lähtötieto (30.3.20)

```
taulu5 <- ISSP2012esim1.dat %>% tableX(maa,Q1b,type = "cell_perc")
knitr::kable(taulu5,digits = 2, booktabs = TRUE,
             caption = "Kysymyksen V6 vastaukset maittain (%)")
```

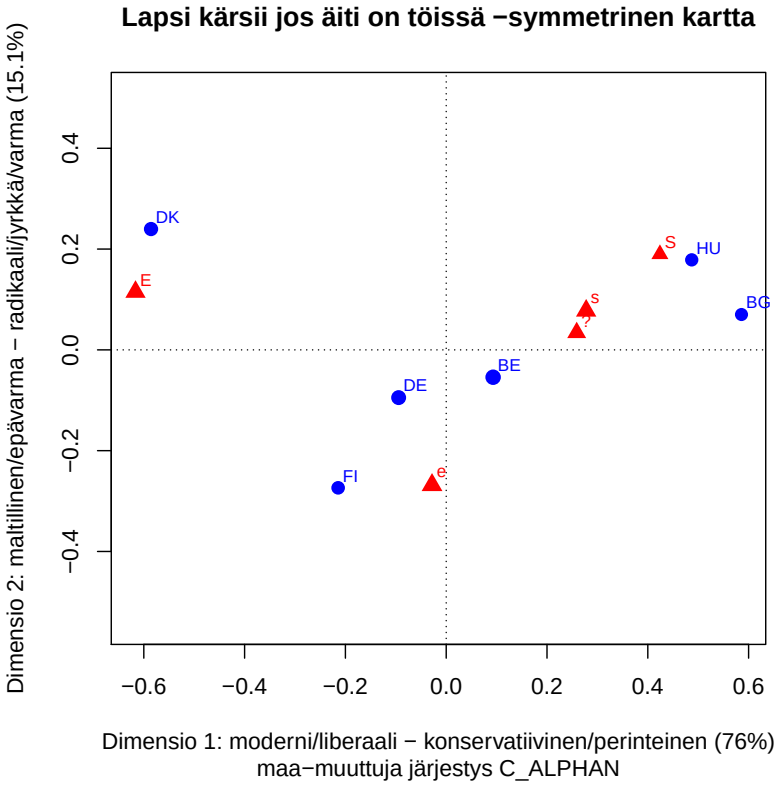
Taulukko 56: Kysymyksen V6 vastaukset maittain (%)

	S	s	?	e	E	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

zxy Tätä ensimmäistä kuvaa on muistiinpanoissa kommentoitu (löytyy printattuna) Kolme karttaa. Kartan kääntyminen ei ole ongelma, mutta vähän kiusallista. Tässä se on seurausmaiden järjestyksestä. Maiden vertailussa on järkevää vakioida niiden massat (kolmas kartta). Massan käsite on CA:n ydinasioita, siksi maiden massat ovat jatkossa mukana. Kartta määräytyy maiden otoskokojen suuruisilla painoilla, mutta ero ei ole kovin suuri.

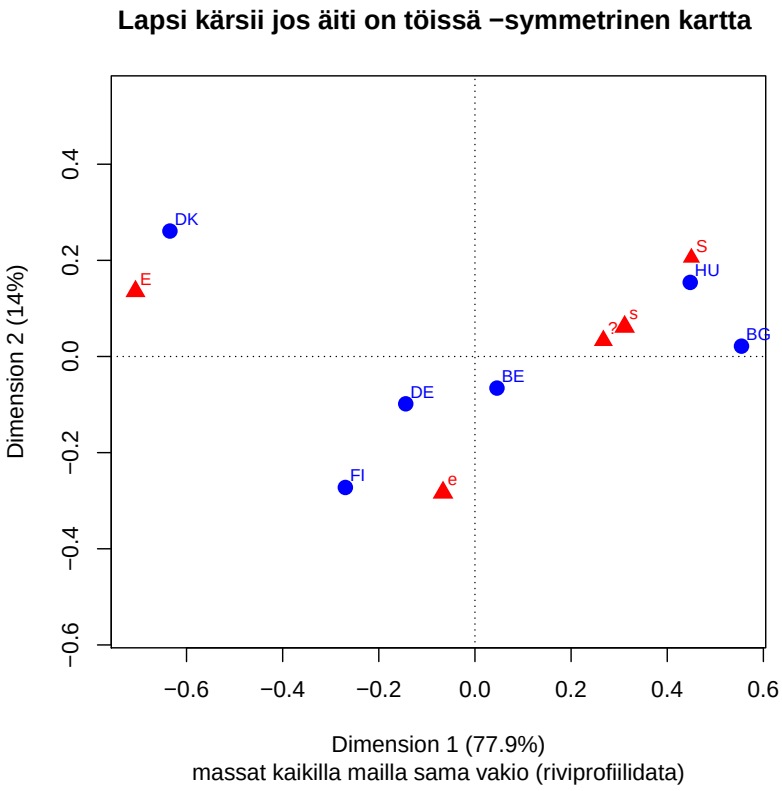
```
#simpleCA1 <- ca(~maa + V6,ISSP2012esim1.dat) suoritetaan ennen värikuvaa, tuloksia tarvitaa
#siinä.

# TODO(11.4.20) fig.cap koodilohkossa tekee kuvasta "kelluvan", ja kuvat numeroidaan.
# Miten plot-komennon kuvaotsikot vaikuttavat?
# Pitäiskö (a) jokaiselle kuvalle oma koodilohko (b) esittää nämä kaksi yhdessä vierekkäin
# Pohditaan kun koodataan capaper-projektia.

# Symmetrinen kartta
# Akselien tekstit "käsityönä" - esimerkki (3.5.2020)

plot(simpleCA1, map = "symmetric", mass = c(TRUE,TRUE),
```

```
     main = "Lapsi kärsii jos äiti on töissä -symmetrinen kartta",
     xlab = "Dimensio 1: moderni/liberaali - konservatiivinen/perinteinen (76%)",
     ylab = "Dimensio 2: maltillinen/epävarma - radikaali/jyrkkä/varma (15.1%)",
     sub = "maa-muuttuja järjestys C_ALPHAN")
```



Kuva 4: Q1b: lapsi kärsii jos äiti on töissä

```
# plot(simpleCA2, map = "symmetric", mass = c(TRUE,TRUE),
#     main = "Lapsi kärsii jos äiti on töissä -symmetrinen kartta ",
#     sub = "maa-muuttuja maa2,järjestys as_factor(C_ALPHAN)")
# Kartta kääntyy x-akselin ympäri - esimerkki faktoroinnin arvaamattomista
# seurauksista (30.3.20)

# par(cex = 0.3) pisteen koko
plot(simpleCA3, map = "symmetric", mass = c(TRUE,TRUE),
main = "Lapsi kärsii jos äiti on töissä -symmetrinen kartta ",
sub = "massat kaikilla mailla sama vakio (riviprofiilidata)")
```



Kuva 5: Q1b: lapsi kärsii jos äiti on töissä

```
#str(simpleCA1)
# 13.5.2018
# kuvasuhteen saa oikeaksi, kun avaa g-ikkunan (X11()) ja sitten plot. Voi tallentaa pdf-mu
# grafiikkaikkunasta, ja ladata outputiin knitr-vaiheessa. Parempi tulostaa kuvatdsto pdf-a
# jos lopulliseen versioon joutuu
# näin tekemään.
# näitä kokeiln chunk-optioissa mutta ei toimineet (out.width = "6", out.hight = "6")
# (13.5.2018), vaan pdf-konversiossa pandoc failed with error 43
```

edit 2.5.2020 Riviprofiilitaulukossa rivimassat ovat vakioita (=1), mutta caratkaisussa skaalautuvat eri arvoksi (vakio).

Näitä karttoja vertaillaan seuraavassa luvussa tarkemmin.

Toinen tapa - kuvatiedoston lataaminen include_graphics - funktiolla. Pitää miettiä mikä on järkevää, dataa tutkaillessa piirretään useita kuvia. PDF-muodossa ne ovat skaalautuvia, kommentteja voi lisätä jne.

Grafiikan hienosäätö on hieman haastavaa: analyysivaiheessa kannattaa tallentaa kuvia RStudion grafiikkaikkunasta pdf-muodossa talteen, graafisessa dataanalyysissä niitä tietenkin syntyy aika paljon. HTML- ja pdf- formaatin kuvat viimeistellään bookdown-ympäristössä.

2.2 Korrespondenssianalyysin käsitteet

1. Profiilit
2. Massat
3. Profiilien etäisyydet (khii2)

zxy Ja tätä “triplettiä” täydentää neljä siitä johdettua käsitettä, viite muistiinpanoissa. **#V** Tässäkin CAiP ja MG2017HY-luentokalvot.

---

3 Tulkinnan perusteita

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.

Asymmetrinen kartta, jossa riviprofiilit ovat pääkomponentti-koordinaateissa ja sarakeprofiilit standardikoordinaateissa.

(1) Sarakkeet ideaalipisteinä, edustavat kuvittellisia maita joissa kaikki ovat vastanneet vain yhdellä tavalla.

(2) Sarakepisteet kaukana origosta, koska skaalattu

(3) Rivipisteet kasautuneet keskiarvopisteen ympärille

(4) Rivi-ja sarakepisteiden suhteelliset sijannit samat kuin symmetrisessä kuvassa

(5) Tässäkin kuvassa pisteen koko kuvaa sen massaa. Sarakkeista "täysin samaa mieltä" (ts) ja "ei samaa eikä eri mieltä" ovat massoiltaan pienimmät.

(6) Pisteiden koko kuvaa rivin tai sarakkeen massaa.

Tarinaa voi tarvittaessa jatkaa, tämä on CA:n hankalin asia. Kaksi koordinaatistoa, ja niiden yhteys.

(7) Asymmetrinen kuva ja akseleiden / dimensioiden tulkinta

Piirretään sama asymmetrinen kartta uudelleen, mutta yhdistetään sarakepisteet keskiarvopisteeseen (sentroidiin) suorilla. Mitä terävämpi on sarakesuoran (vektorin?) ja akselin kulma, sitä enemmän sarake määrittää tätä ulottuvuutta. Jos vektori on lähettä 45 asteen kulmaa, sarake määrittää yhtä paljon molempia ulottuvuuksia.

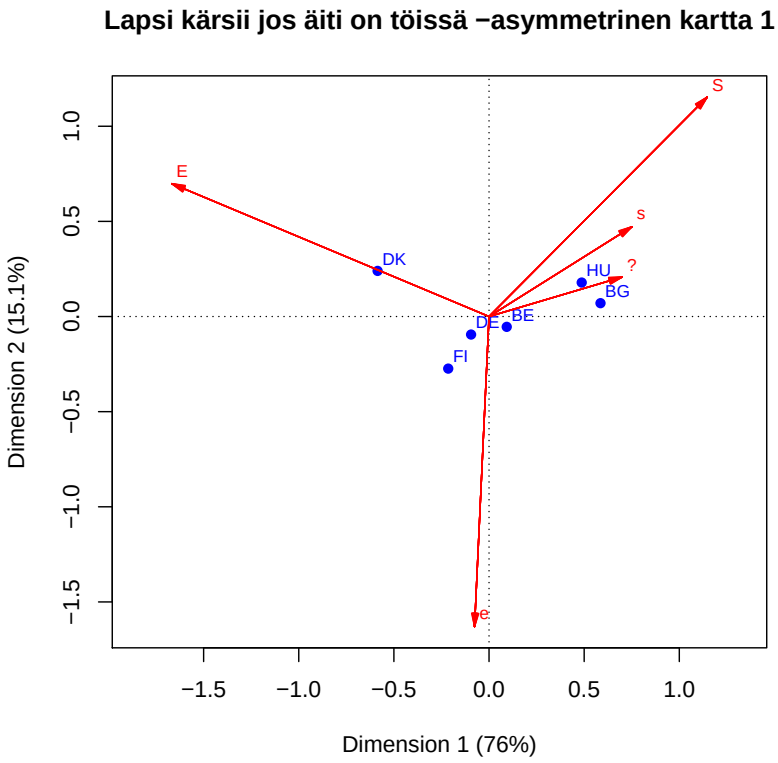
```
# asymmetrinen kartta - rivit pc ja sarakkeet sc
# sarakkeet vektorikuvina
# HUOM! simpleCA1 luodaan G1_2_johdesim.Rmd - tiedostossa
#
# Kuva tiedostoon - ennen plot-komentoa avataan tiedosto
# pdf("img/sCA1asymm1.pdf")
plot(simpleCA1, map = "rowprincipal",
     arrows = c(FALSE,TRUE),
     main = "Lapsi kärsii jos äiti on töissä -asymmetrinen kartta 1" )
```

```
# Kuva tiedostoon - suljetaan
# dev.off()
```

Sarakkeen "Eri mieltä" (e) määrittää toisen ulottuvuuden, jonka voisi tulkita erottelevan "maltilliset" mielipiteen tiukemmista. Sarake "täysin samaa mieltä" (S) määrittää toista ulottuvuutta lähes yhtä paljon kuin ensimmäistä, mutta "täysin eri mieltä" (E) on vasemmalla ja kolme vastausvaihtoehtoa oikealla. Kovin terävästi dimensiot eivät eroa toisistaan?

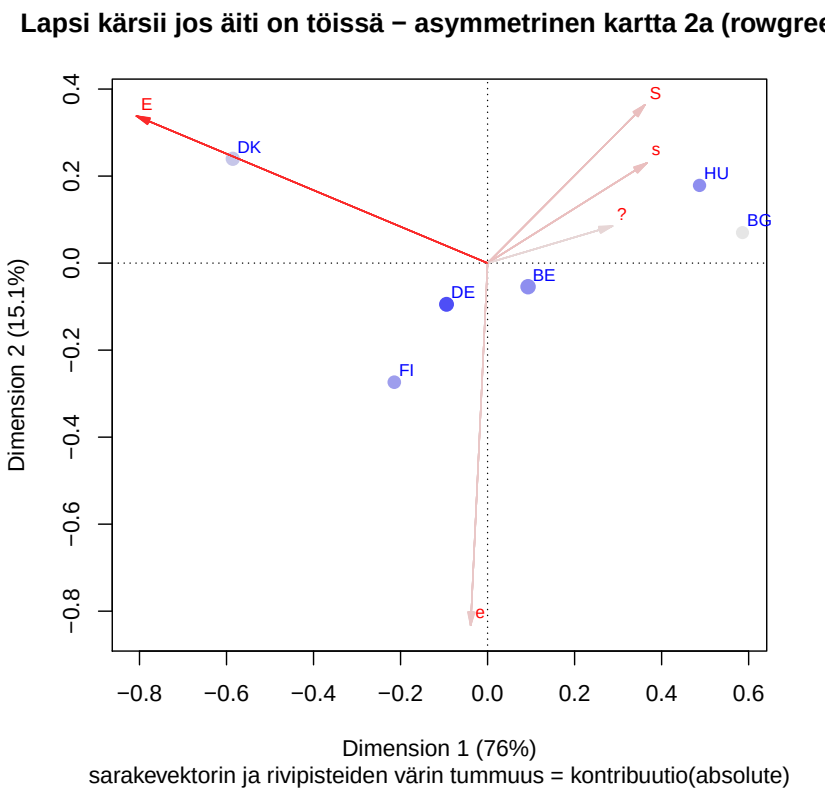
Edit 3.5.20 Selvennä: sarakevektroit ovat standardikoordinaateissa, ideaalipisteitä ("maa jossa kaikki samaa mieltä"). Miksi ne ovat kartalla "reilusti" ykköstä suurempia? Vastaus: ideaalipisteet esitetään rivipisteiden koordinaatistossa - > skaalaus.

```
#X11() komentoriville ja plot-komento
plot(simpleCA1, map = "rowgreen",
     contrib = c("absolute", "absolute"),
     mass = c(TRUE,TRUE),
```



Kuva 6: Q1b: lapsi kärsii jos äiti on töissä

```
     arrows = c(FALSE, TRUE),
     main = "Lapsi kärsii jos äiti on töissä - asymmetrinen kartta 2a (rowgreen)",
     sub = "sarakevektorin ja rivipisteiden värin tummuus = kontribuutio(absolute)")
```



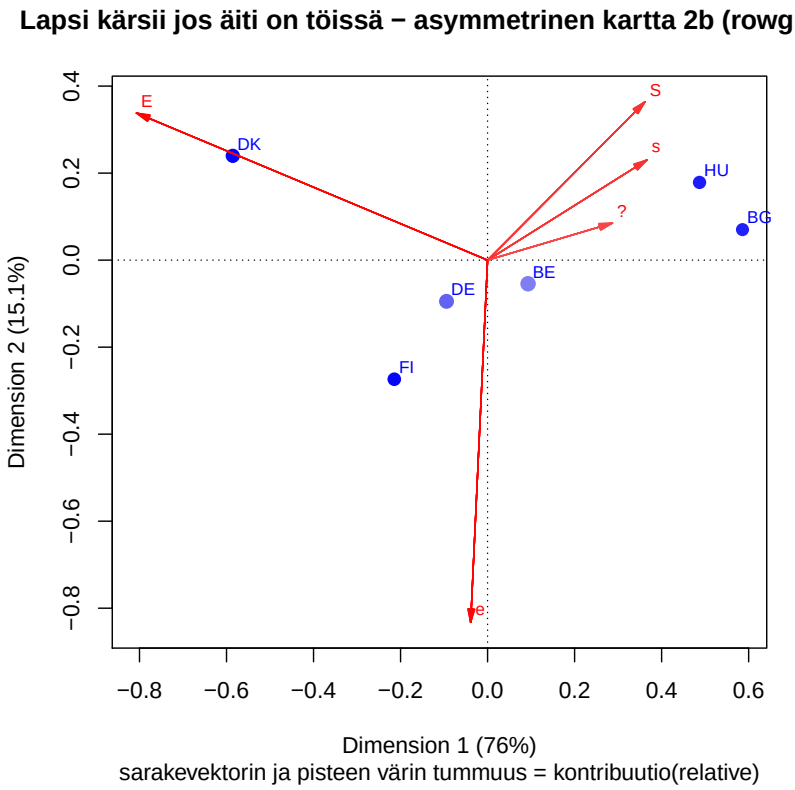
Kuva 7: Q1b: lapsi kärsii jos äiti on töissä

Greenacre (2006, "loose ends -artikkeli") ehdotti asymmetrisessä kuvassa standardikoordinaattien skaalaamista niin, että ne kerrotaan massan neliöjuurella. Tämä skaalaus toimii hyvin pienen ja suuren inertian tapauksessa. Kartoissa pätee sama sääntö kuin muussakin graafisessa data-analyysissä, kuvien on esitettävä oleelliset yhteydet, mutta mielellään vain ne.

Sama kuva, kontribuutiot "relative". **edit 24.2.20** Ero selitettävä!

```
#X11() komentoriville ja plot-komento
plot(simpleCA1, map = "rowgreen",
     contrib = c("relative", "relative"),
     mass = c(TRUE,TRUE),
     arrows = c(FALSE, TRUE),
```

```
     main = "Lapsi kärsii jos äiti on töissä - asymmetrinen kartta 2b (rowgreen)",
     sub = "sarakevektorin ja pisteen värin tummuus = kontribuutio(relative)")
```



Kuva 8: Q1b: lapsi kärsii jos äiti on töissä

Tulkinta: rivipisteiden ortogonaalinen projektio "sarakevektorille"

Asymmetrisessä kartassa 2 pisteiden koko on suhteessa niiden massaan, ja värisävy absoluuttiseen tai suhteelliseen kontribuutioon.

```
# CA:n numeeriset tulokset
# (11.4.20) yhdistä koodilohkoon khi2dist1 (G1_2_johdesim.Rmd, r. 665)
# CA:n numeeristen tulosten käsittelyä myös CAcalc_1.R -skriptissä.

summary(simpleCA1)
```

```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %    cum%   scree plot
```

```
##  1      0.136619  76.0  76.0  ********************
##  2      0.027089  15.1  91.1  ****
##  3      0.010054   5.6  96.7  *
##  4      0.005988   3.3 100.0  *
##         -------- -----
##  Total: 0.179751 100.0
##
##
## Rows:
##     name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 |   BE |  247  465   34 |   93 347  16 |  -54 118  27 |
## 2 |   BG |  113  874  251 |  586 862 284 |   70  12  21 |
## 3 |   DE |  210  584   36 |  -94 291  14 |  -95 293  70 |
## 4 |   DK |  170  996  381 | -586 853 428 |  240 143 362 |
## 5 |   FI |  136 1000   92 | -214 380  46 | -274 620 377 |
## 6 |   HU |  122  889  206 |  487 783 213 |  179 105 144 |
##
## Columns:
##     name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 |    S |   99  784  152 |  424 653 131 |  190 131 132 |
## 2 |    s |  238  788  140 |  278 731 134 |   78  57  53 |
## 3 |      |  168  720   88 |  259 707  82 |   34  12   7 |
## 4 |    e |  261  982  108 |  -28  11   2 | -268 971 693 |
## 5 |    E |  234 1000  512 | -616 966 651 |  115  34 114 |
```

```
# vertailun vuoksi numeeriset tulokset, kun maiden massat vakiot
summary(simpleCA3)
```

```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.167678  77.9  77.9  ********************
##  2      0.030095  14.0  91.9  ***
##  3      0.013206   6.1  98.0  **
##  4      0.004296   2.0 100.0
##         -------- -----
##  Total: 0.215275 100.0
##
##
## Rows:
##     name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 |   BE |  167  295   17 |   46  97   2 |  -66 199  24 |
## 2 |   BG |  167  884  270 |  554 882 306 |   22   1   3 |
## 3 |   DE |  167  718   33 | -144 489  21 |  -98 229  54 |
## 4 |   DK |  167  993  367 | -635 849 400 |  261 144 377 |
```

```
## 5 |   FI |  167  999  114 | -270 494  72 | -272 505 411 |
## 6 |   HU |  167  870  200 |  448 778 199 |  154  92 132 |
##
## Columns:
##     name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 |    S |  105  785  153 |  450 649 127 |  206 135 148 |
## 2 |    s |  250  792  148 |  311 762 145 |   62  30  32 |
## 3 |      |  171  792   73 |  267 780  73 |   33  12   6 |
## 4 |    e |  256  976  103 |  -66  51   7 | -283 925 681 |
## 5 |    E |  218 1000  524 | -706 964 649 |  136  36 133 |
```

```
# Rivi- ja sarake-etäisyydet (keskiarvosta/sentroidista)
# HUOM! Edellisessä jaksossa taulukko rivi- ja sarake-etäisyyksistä. Tuskin
# kannattaa tässä toistaa. Muuta analyysiä numeerisista tuloksista. (10.4.20)

simpleCA1$rownames
```

```
## [1] "BE" "BG" "DE" "DK" "FI" "HU"
```

```
simpleCA1$rowdist
```

```
## [1] 0.1579735 0.6309909 0.1750128 0.6340627 0.3477331 0.5504040
```

```
simpleCA3$rowdist
```

```
## [1] 0.1474052 0.5902700 0.2059132 0.6885400 0.3835387 0.5078504
```

```
simpleCA1$colnames
```

```
## [1] "S" "s" "?" "e" "E"
```

```
simpleCA1$coldist
```

```
## [1] 0.5246525 0.3248840 0.3078230 0.2721699 0.6271108
```

```
simpleCA3$coldist
```

```
## [1] 0.5587368 0.3567818 0.3025459 0.2944703 0.7190317
```

Ensimmäinen tuloste on CA maapainoilla, ja toisessa maapainot on vakioitu. CA-funktiolle on siinä annettu dataksi riviprofiilit, rivisummat ovat ykkösiä. Rivimassat skaalautuvat niin, että niiden summa on 1.

Edellisessä jaksossa esimerkki siistimmästä taulukosta.

TODO (21.2.20)

(8) Tätä voisi käyttää esimerkkinä numeeristen tulosten vertailussa?

(9) Kokonaisinertia kasvaa (0,18 -> 0,26), koordinaatisto muuttuu mutta ei kovin radikaalisti.

(10) Kvaliteetti, kontribuutiot? Miten vertailla oleellisia asioita?

Belgian laatu putoavat merkittävästi,ja kontribuutiot pienenevät entisestään.

Saksan laatu paranee aika paljon, ja kontribuutiotkin jonkin verran. Aika outoa,että suurimman massan maiden (DE,BE lähes puolet datasta) kontribuutiot ovat niin pieniä (24.2.20)

(11) **Miksi sarakemassat muuttuvat?** Joku skaalaus, massojen "kaksoisrooli" painoina ja normalisoivina ("varianssin tapaan") muunnoksina. Hieman on hämärä, edelleen! Teknisiä tuloksia ei ehkä pidä käyttää mihinkään muuhun kuin kartan laadulliseen arviointiin ja kuvan tulkintojen varmistamiseen?(24.2.20)

zxy Taulukon käsitteiden läpikäynti ja pureskelu kuulunee seuraavaan lukuun?

MG & Blasius, "vihreä kirja": kontribuutiot inertiaan

---

4 Yksinkertaisen korrespondenssianalyysin laajennuksia 1

Korrespondenssianalyysi sallii rivien tai sarakkeiden yhdistelyn tai "jakamisen". Tämä onnistuu esimerkkiaineistossa lisäämällä rivejä eli jakamalla eri maiden vastausksia 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).

Nyt käytetään johdattelevan esimerkin dataa, johon muunnokset on jo alustavasti tehty.

Vanhaa koodia kolme koodilohkoa

4.1 Täydentävät muuttujat (supplementary points)

zxy Piste sinne piirretään, mutta muuttujassa on se tieto. "Täydentävät piste"" kuulostaa huonolta. Lisämuuttujat, havainnot, lisäpisteet?

Viite:CAip ss 89, HY2017_MCA.

Aineistossa on havaintoja (rivejä) tai muuttujia (sarakkeita), joista voi olla hyötyä tulosten tulkinnassa. Nämä lisäpisteet voidaan sijoittaa kartalle, jos niitä voidaan jotenkin järkevästi vertailla kartan luomisessa käytettyihin profiileihin (riveihin ja sarakkeisiin).

EDIT Lisätään Belgian ja Saksan aluejako täydentäviksi riveiksi. Sopii tarinaan, dimensioiden tulkinta ei ollut esimerkissä kovin kirkas. Viite CAip:n lukuun,

jossa vain todetaan että maita ei ole järkevää painottaa (massa) otoskoolla, vaan vakioidaan (jotenkin) sama (suhteellinen) massa kaikille. Samalla oikaistaan myös naisten yliedustus aineistossa.

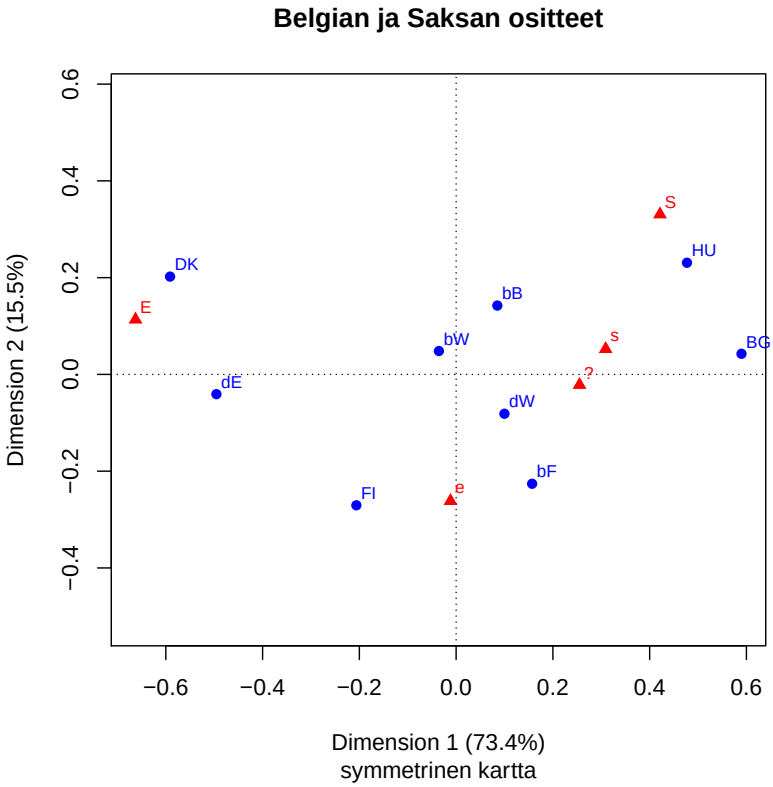
Käsitteitä: (@) Active point, aktiivinen piste (aktiivinen havainto tai muuttuja). (@) Supplementary pointtäydentävä piste (täydentävä havainto).

Täydentävien muuttujien kolme käyttötapaa:

1. Sisällöllisesti tutkimusongelman kannalta poikkeava tai erilainen rivi tai sarake
2. Outlayerit, poikkeava havainto jolla pieni massa (esimerkissä uusi sarake-muuttuja, jossa kovin vähän havaintoja)
3. osaryhmät **EDIT** capaper- jäsentelyssä ja bookdown-dokumentissa selitetty täydentävät/lisäpisteet tarkemmin (18.9.2018).

```
# Kömpelöä koodia, harjoitellaan taulukoiden yhdistelyä (CAtest1.Rmd)
# Belgian ja Sakasan jako lisäpisteinä 24.5.2018


# HUOM! Tässä ei vielä supp.points mukana!
suppointCA1 <- ca(~maa3 + Q1b,ISSP2012esim1.dat)
plot(suppointCA1, main = "Belgian ja Saksan ositteet",
                sub = "symmetrinen kartta")
```

```
# VANHAA 21.2.20
#
#Jos kuva kääntyy ympäri, kerrotaan koordinaattivektorit luvulla -1
#summary(suppointCA1)
#print(suppointCA1)
#str(suppointCA1)
#
# Käännetään kuva - EI TARVITA ENÄÄN, maa3-muuttujan järjestys sama kuin
# maa- muuttujan (C_ALPHAN - järjestys)

# suppointCA1b <- suppointCA1
# suppointCA1b$rowcoord <- suppointCA1b$rowcoord[,] * (-1)
# suppointCA1b$colcoord <- suppointCA1b$colcoord[,] * (-1)
# suppointCA1b$rowcoord
# suppointCA1b$colcoord
# plot(suppointCA1b, main = "Belgian ja Saksan ositteet - käännetty kartta")

# Miten lisärivit? (24.5.2018)
# Luetaan data tauluksi - ei toimi, char-table.Toimisiko nyt, ei chr? (4.2.20)
# yritetään uudestaan table-funktiolla

# data maa3-muuttujassa
```



Kuva 9: Belgian ja Saksan aluejako

```
# str(ISSP2012esim1.dat$maa3)
# attributes(ISSP2012esim1.dat$maa3)

suppoint1_df1 <- select(ISSP2012esim1.dat, maa3,Q1b)

#  tarkistuksiin jos koodi suoritetaan rivi kerrallaan
# str(suppoint1_tab1)

suppoint1_tab1 <- table(suppoint1_df1$maa3, suppoint1_df1$Q1b)
suppoint1_tab1
```

/	S	s	?	e	E
bF	51	241	262	312	146
bW	53	103	91	118	125
bB	87	107	85	122	110
BG	118	395	205	190	13
dW	133	313	138	375	208
dE	32	62	60	163	230
DK	70	238	152	232	696
FI	47	188	149	423	303
HU	219	288	225	190	75

```
#plot(ca(~maa2 + V6, suppoint1_df1)) #toimii
#
# Saksan ja Belgian summarivit
#
suppoint2_df <- filter(ISSP2012esim1.dat, (maa == "BE" | maa == "DE"))
suppoint2_df <- select(suppoint2_df, maa, Q1b)
#head(suppoint2_df)
#tail(suppoint2_df)
str(suppoint2_df)
```

```
## tibble [3,727 x 2] (S3: tbl_df/tbl/data.frame)
##  $ maa: Factor w/ 6 levels "BE","BG","DE",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Q1b: Factor w/ 5 levels "S","s","?","e",..: 4 5 1 4 2 2 2 2 1 1 ...
```

```
# attributes(suppoint2_df) # korvaa attr(x, which) tms. liian pitkä tulostus
# attr(suppoint2_df, which = "class")
# attr(suppoint2_df, which = "name")
# summary(suppoint2_df)
suppoint2_df %>% table1() # miksi ei tulosta mitään (4.2.20)
```

```
##
## -------------------------
##        Mean/Count (SD/%)
```

```
##         n = 3727
##  maa
##     BE 2013 (54%)
##     BG 0 (0%)
##     DE 1714 (46%)
##     DK 0 (0%)
##     FI 0 (0%)
##     HU 0 (0%)
##  Q1b
##     S  356 (9.6%)
##     s  826 (22.2%)
##     ?  636 (17.1%)
##     e  1090 (29.2%)
##     E  819 (22%)
## -------------------------
```

```
suppoint2_tab1 <- table(suppoint2_df$maa, suppoint2_df$Q1b)
suppoint2_tab1 # tarkistus
```

/	S	s	?	e	E
BE	191	451	438	552	381
BG	0	0	0	0	0
DE	165	375	198	538	438
DK	0	0	0	0	0
FI	0	0	0	0	0
HU	0	0	0	0	0

```
suppoint2_tab1 <- suppoint2_tab1[-2,]
# kömpelösti kolme kertaa
suppoint2_tab1 <- suppoint2_tab1[-3,]
suppoint2_tab1 <- suppoint2_tab1[-3,]
suppoint2_tab1 <- suppoint2_tab1[-3,]

suppoint2_tab1 # Belgian ja Saksan summat yli ositteiden
```

/	S	s	?	e	E
BE	191	451	438	552	381
DE	165	375	198	538	438

```
#lisätään rivit maa3-muuttujan taulukkoon

suppoint1_tab1 <- rbind(suppoint1_tab1, suppoint2_tab1)
suppoint1_tab1
```

	S	s	?	e	E
bF	51	241	262	312	146
bW	53	103	91	118	125
bB	87	107	85	122	110
BG	118	395	205	190	13
dW	133	313	138	375	208
dE	32	62	60	163	230
DK	70	238	152	232	696
FI	47	188	149	423	303
HU	219	288	225	190	75
BE	191	451	438	552	381
DE	165	375	198	538	438

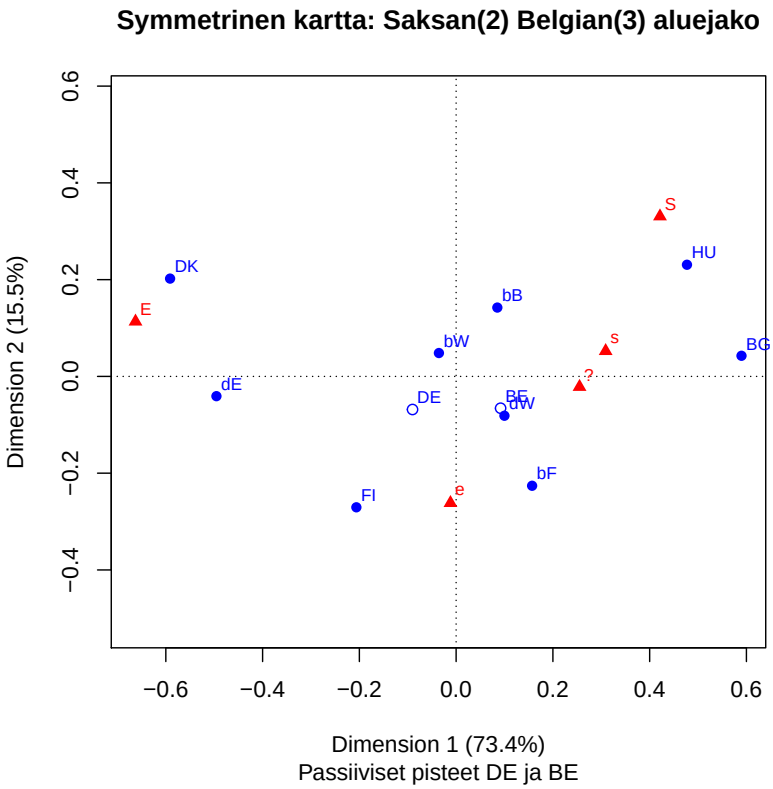
```
suppointCA2 <- ca(suppoint1_tab1[,1:5], suprow = 10:11)
plot(suppointCA2, main = "Symmetrinen kartta: Saksan(2) Belgian(3) aluejako",
     sub = "Passiiviset pisteet DE ja BE" )
```

```
# käännetään kuva VANHAA (21.2.20)
# suppointCA2b <- suppointCA2
# suppointCA2b$rowcoord <- suppointCA2b$rowcoord[,] * (-1)
# suppointCA2b$colcoord <- suppointCA2b$colcoord[,] * (-1)

# plot(suppointCA2b,main = "Saksan ja Belgian ositteet",
#     sub = "Symmetrinen kartta, passiiviset pisteet DE ja BE" )
# ca- output
#names(suppointCA2b)
#str(suppointCA2b)
#str(suppointCA2b$rowcoord)
#suppointCA2b
#suppointCA2b$rowcoord
#apply(suppointCA2b$rowcoord, 2, sum)
#suppointCA2b$rowdist
#suppointCA2b$coldist
# summary(suppointCA2)
```

Kääntöä ei tarvita, kun maiden järjestys on sama myös muuttujassa maa3 (mukana maiden jaot)

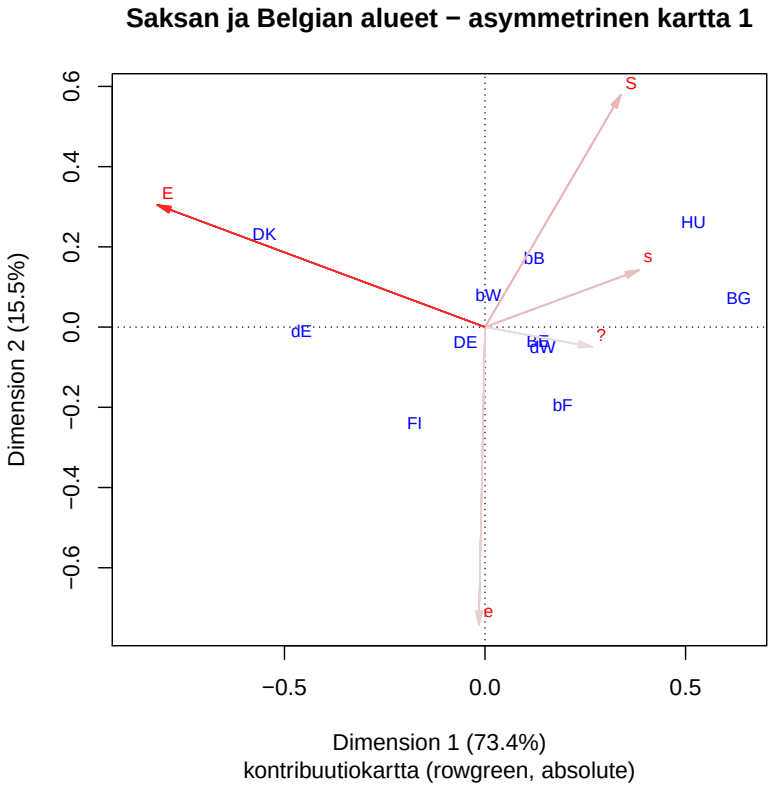
Saksan ja Belgian summarivit ovat ositteiden painotettuja keskiarvoja (sentroideja), läntisen ja itäisen Saksan rivipisteiden välisellä janalla on koko maan summapiste DE.

Piirretään vertailun vuoksi vielä asymmettrinen kartta ("kontribuutio-kartta, kontribuutio-kaksoiskuva"). **edit 3.5.20** Minne katoavat pisteet?



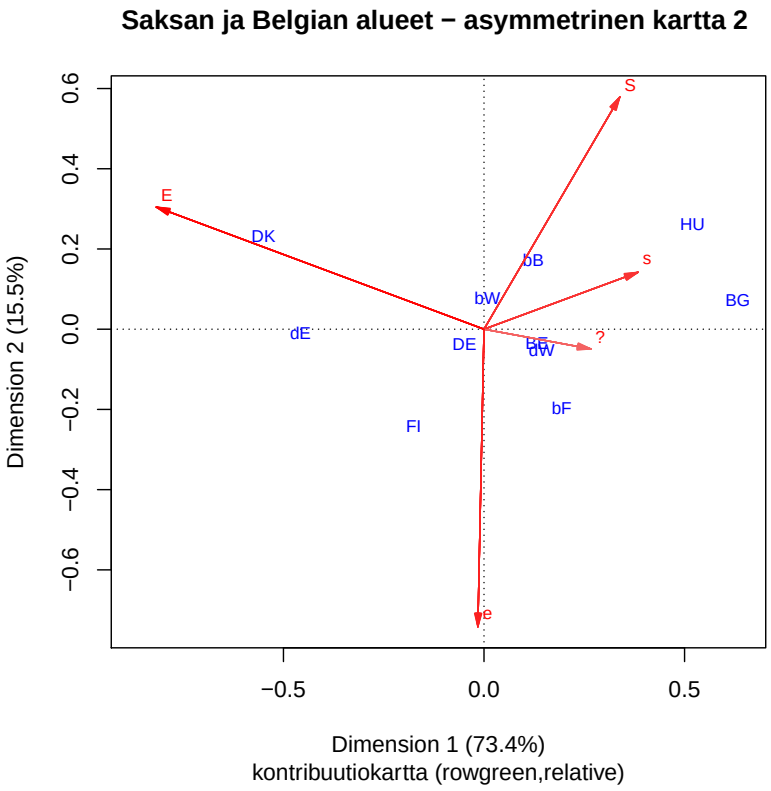
Kuva 10: Belgian ja Saksan aluejako

```
plot(suppointCA2, map = "rowgreen",
     contrib = c("absolute", "absolute"),
     mass = c(TRUE,TRUE),
     arrows = c(FALSE, TRUE),
     main = "Saksan ja Belgian alueet - asymmetrinen kartta 1",
     sub = "kontribuutiokartta (rowgreen, absolute)")
```



Kuva 11: Belgian ja Saksan aluejako

```
# Sama kuva, maasummat lisäpisteinä (4.2.20)

plot(suppointCA2, map = "rowgreen",
     contrib = c("relative", "relative"),
     mass = c(TRUE,TRUE),
     arrows = c(FALSE, TRUE),
     main = "Saksan ja Belgian alueet - asymmetrinen kartta 2 ",
     sub = "kontribuutiokartta (rowgreen,relative)")
```



Kuva 12: Belgian ja Saksan aluejako

Kaksi konrtibuutio-karttaa (MG:n keksintö) osoittavat, että tulkinnan hankaluuksista huolimatta symmetrinen kartta on usein selkeämpi. Molemmissa ideaalipisteet sijatsevat kaukana (vaikka ne on skaalattu hieman lähemmäs origoa), ja maapisteiden hajontaa on aika vaikeaa nähdä. Belgian täydentävä maapiste (BE) peittyy läntisen Saksan (dW) alle.

Tulostetaan numeeriset taulukot.

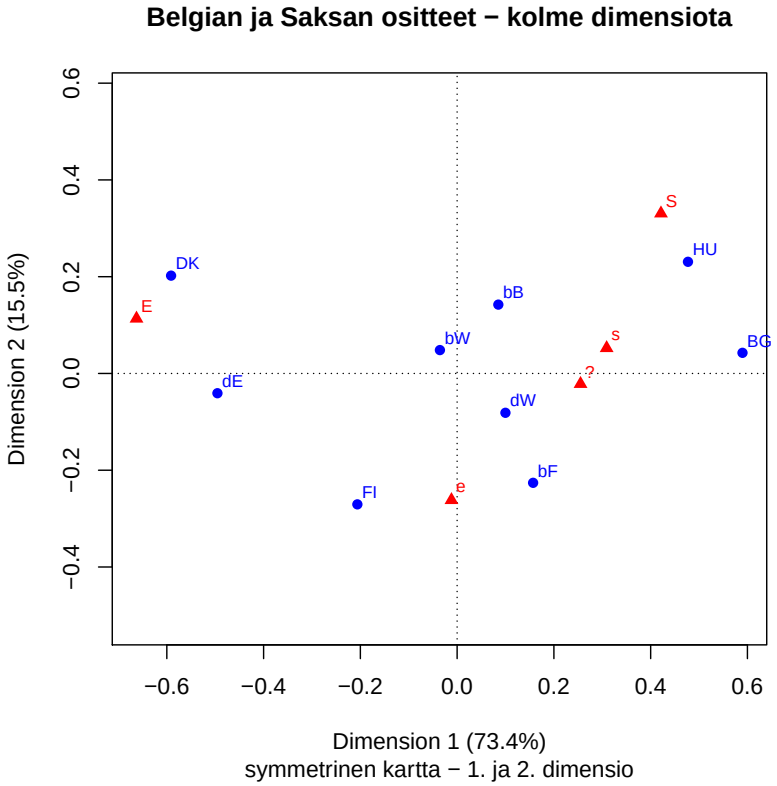
```
# CA - numeeriset tulokset

summary(suppointCA2)
```

```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.154101  73.4  73.4  ******************
##  2      0.032489  15.5  88.9  ****
##  3      0.014294   6.8  95.7  **
##  4      0.008944   4.3 100.0  *
##         -------- -----
##  Total: 0.209828 100.0
##
##
## Rows:
##       name   mass  qlt  inr    k=1 cor  ctr    k=2 cor  ctr
## 1  |    bF |  124  650   69 |  157 212   20 | -226 438  195 |
## 2  |    bW |   60  388    3 |  -36 137    0 |   48 252    4 |
## 3  |    bB |   63  481   17 |   85 127    3 |  142 354   39 |
## 4  |    BG |  113  878  215 |  590 874  255 |   43   5    6 |
## 5  |    dW |  143  345   33 |  100 208    9 |  -81 138   29 |
## 6  |    dE |   67  966   82 | -495 960  107 |  -41   7    3 |
## 7  |    DK |  170  971  327 | -591 869  387 |  202 102  214 |
## 8  |    FI |  136  957   79 | -206 352   38 | -271 605  307 |
## 9  |    HU |  122  927  177 |  477 751  181 |  231 176  201 |
## 10 | (*)BE | <NA>  512 <NA> |   92 338 <NA> |  -66 173 <NA> |
## 11 | (*)DE | <NA>  418 <NA> |  -90 265 <NA> |  -68 153 <NA> |
##
## Columns:
##     name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 |    S |   99  816  167 |  421 505 115 |  331 311 335 |
## 2 |    s |  238  781  143 |  309 759 147 |   52  22  20 |
## 3 |      |  168  594   88 |  255 589  71 |  -22   4   2 |
## 4 |    e |  261  871   98 |  -12   2   0 | -262 870 550 |
## 5 |    E |  234  999  505 | -663 971 667 |  113  28  93 |
```

Kolmiulotteisesta kuvasta voi tulostaa molempien akseleiden ja uuden kolmannen

akselin kartat. R-ohjelmistossa voi tulostaa näytölle kolmiulotteisen kuvan, siitä voisi ehkä ottaa kuvakaappauksena esimerkin raporttiin? **edit** Kannattaako (30.3.20)

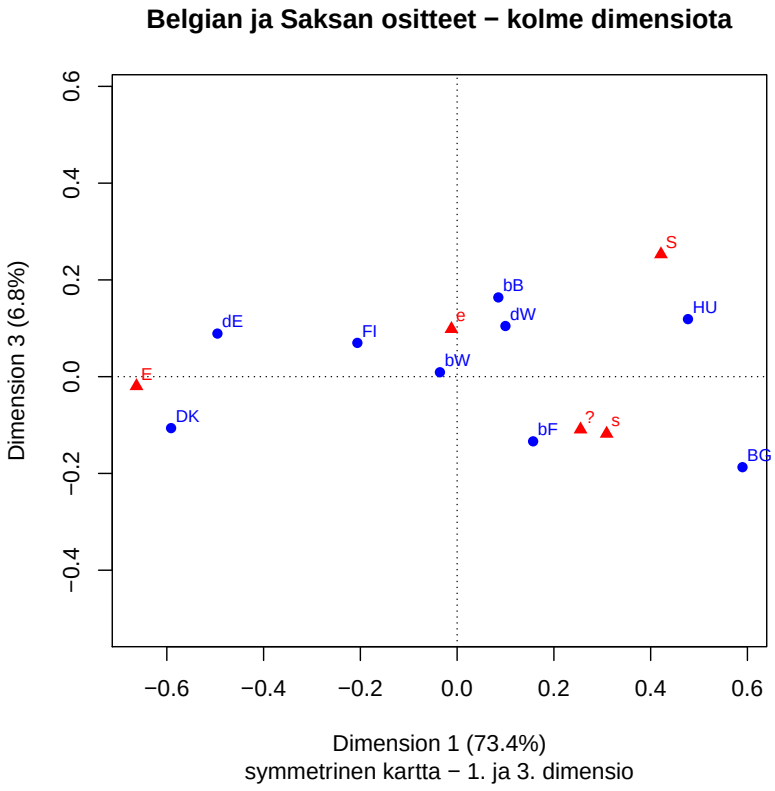
```
# Näkyisikö Belgian aluejako kolmannessa dimensiossa? (19.2.20)
# Toimii, mutta siistittävä, samoin koko maajakoskripti!

suppointCA3 <- ca(~maa3 + Q1b,ISSP2012esim1.dat, nd = 3)
# (24.2.20)
# Tulostetaan kolme karttaa - ensimmäinen ja toinen akseli uuden kolmannen kera

plot(suppointCA3, dim = c(1,2),
                main = "Belgian ja Saksan ositteet - kolme dimensiota",
                sub = "symmetrinen kartta - 1. ja 2. dimensio")
```



Kuva 13: Belgian ja Saksan aluejako - 3D

```
plot(suppointCA3, dim = c(1,3),
                main = "Belgian ja Saksan ositteet - kolme dimensiota",
```

```
                sub = "symmetrinen kartta - 1. ja 3. dimensio")
```

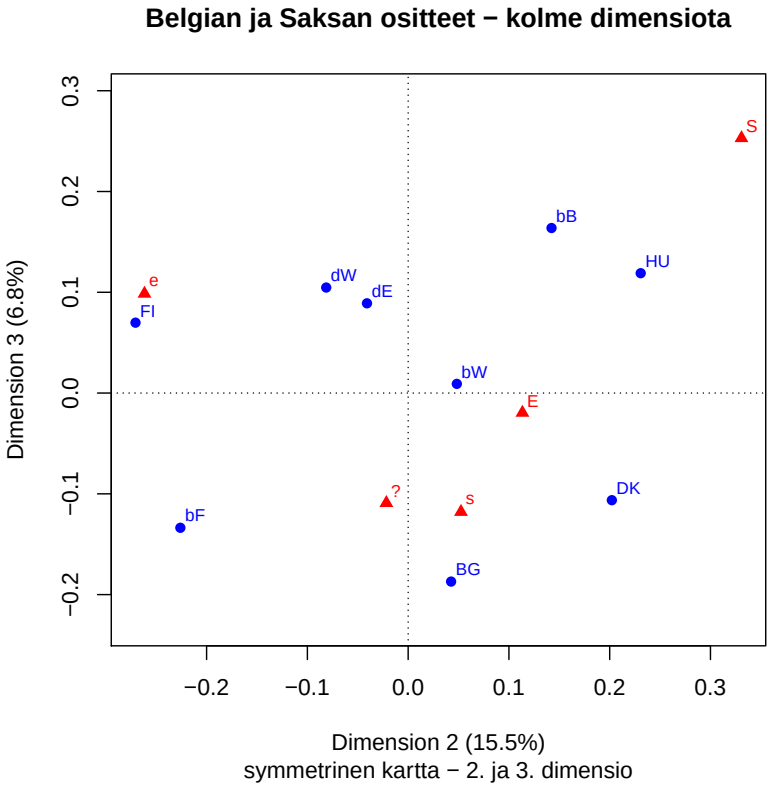


Kuva 14: Belgian ja Saksan aluejako - 3D

```
plot(suppointCA3, dim = c(2,3),
                main = "Belgian ja Saksan ositteet - kolme dimensiota",
                sub = "symmetrinen kartta - 2. ja 3. dimensio")
```

```
# summary(suppointCA3) - virheilmoitus (21.2.20)

# Virheilmoitus "Error in rsc %*% diag(sv) : non-conformable arguments" ?!
# onko vika täydentävissä pisteissä? Ei ole, eivät ole mukana
# ISSP2012esim1.dat %>% tableX(maa3, Q1b)

suppointCA3
```

```
##
##  Principal inertias (eigenvalues):
##            1        2        3        4
```



Kuva 15: Belgian ja Saksan aluejako - 3D

```
## Value      0.154101 0.032489 0.014294 0.008944
## Percentage 73.44%   15.48%   6.81%    4.26%
##
##
##  Rows:
##                bF        bW       bB        BG        dW        dE        DK
## Mass     0.124279  0.060174 0.062753  0.113103  0.143313  0.067174  0.170453
## ChiDist  0.341469  0.096258 0.239034  0.630991  0.219094  0.505720  0.634063
## Inertia  0.014491  0.000558 0.003586  0.045032  0.006879  0.017180  0.068528
## Dim. 1   0.400065 -0.090631 0.216912  1.502458  0.254323 -1.262007 -1.506022
## Dim. 2  -1.254042  0.267998 0.789358  0.236498 -0.451124 -0.226595  1.121468
## Dim. 3  -1.118212  0.076188 1.369786 -1.564654  0.875735  0.744856 -0.889187
##                FI       HU
## Mass     0.136313 0.122436
## ChiDist  0.347733 0.550404
## Inertia  0.016483 0.037091
## Dim. 1  -0.525222 1.215462
## Dim. 2  -1.500986 1.280342
## Dim. 3   0.584116 0.994772
##
##
##  Columns:
##                S         s         ?         e         E
## Mass    0.099472  0.237627  0.167874  0.260960  0.234066
## ChiDist 0.592824  0.354761  0.332288  0.280549  0.672594
## Inertia 0.034959  0.029907  0.018536  0.020540  0.105887
## Dim. 1  1.073310  0.787257  0.649789 -0.029859 -1.688108
## Dim. 2  1.835133  0.290929 -0.119934 -1.451548  0.629110
## Dim. 3  2.116048 -0.986156 -0.912379  0.824777 -0.163282
```

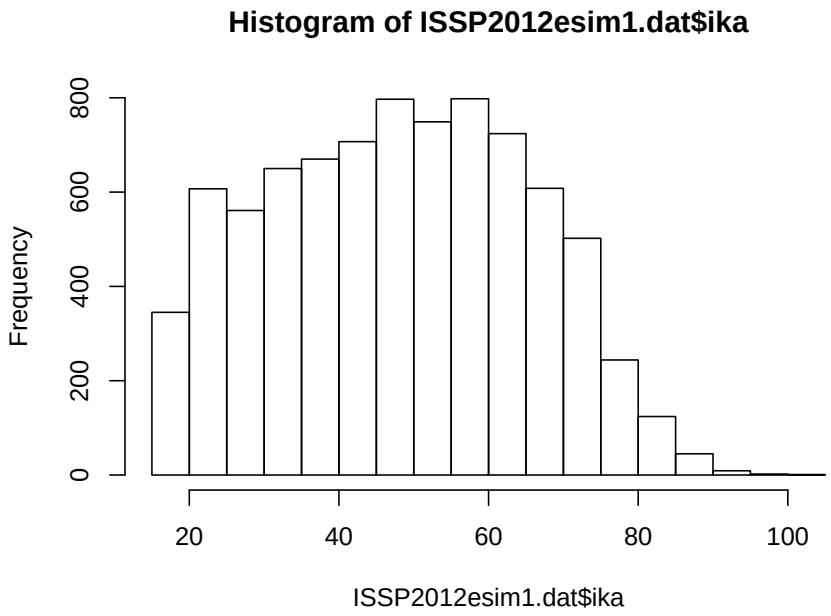
```
# Virheilmoituksen selvittelyä (24.2.20)
# str(suppointCA3)
# diag(suppointCA3$sv)

# Kolmiulottein kuva grafiikkaikkunaan

plot3d(suppointCA3, c(1,2,3))

# Hyödyllinen, mutta aika vaikea
```

4.2 Lisämuuttujat: ikäluokka ja sukupuoli

zxy Otsikkoa pitää harkita, CAip - kirjassa tämä on ensimmäinen esimerkki yksinkertaisen CA:n laajennuksesta. Otsikkona on “multiway tables”, ja tästä yhteisvaikutusmuuttujan (interactive coding) luominen on ensimmäinen esimerkki. Menetelmää taivutetaan sen jälkeen moneen suuntaan.

Luodaan luokiteltu ikämuuttua age_cat, ja sen avulla iän ja sukupuolen interaktiomuuttuja ga. Maiden välillä on hieman eroja siinä, kuinka nuoria vastaajia on otettu tutkimuksen kohteeksi. Suomessa alaikäraja on 15 vuotta, monessa maassa se on hieman korkeampi. Ikäluokat ovat (1=15-25, 2 =26-35, 3=36-45, 4=46-55, 5=56-65, 6= 66 tai vanhempi). Vuorovaikutusmuuttuja ga koodataan f1,. . . , f6 ja m1,. . . ,m6. Muuttujien nimet kannattaa pitää mahdollisimman lyhyinä.

```
# Iän ja sukupuolen vuorovaikutusmuuttujia 1
#
# Uusi R-data: ISSP2012esim2.dat - MIKSI, TARVITAANKO? *esim1.dat kelpaisi?(4.2.20)
#
#age_cat
#AGE 1=15-25, 2 =26-35, 3=36-45, 4=46-55, 5=56-65, 6= 66 and older
#
#summary(ISSP2012esim1.dat$AGE)
hist(ISSP2012esim1.dat$ika)
```




```
ISSP2012esim2.dat <- mutate(ISSP2012esim1.dat, age_cat = ifelse(ika %in% 15:25, "1",
                        ifelse(ika %in% 26:35, "2",
                        ifelse(ika %in% 36:45, "3",
                        ifelse(ika %in% 46:55, "4",
                        ifelse(ika %in% 56:65, "5", "6"))))))
```

```r
ISSP2012esim2.dat <- ISSP2012esim2.dat %>%   # uusi (4.2.20)
        mutate(age_cat = as_factor(age_cat)) # järjestys omituinen! (4.2.20)
# Tarkistuksia

# str(ISSP2012esim2.dat$age_cat)
# levels(ISSP2012esim2.dat$age_cat)
# ISSP2012esim2.dat$age_cat %>% summary()

# Järjestetään ikäluokat uudelleen

ISSP2012esim2.dat <- ISSP2012esim2.dat %>%
        mutate(age_cat =
                fct_relevel(age_cat,
                            "1",
                            "2",
                            "3",
                            "4",
                            "5",
                            "6")
               )

# Tarkistuksia

# Iso taulukko, voi tarkistaa että muunnos ok.
# test6 %>% tableX(AGE, age_cat, type = "count")
# taulu42 <- ISSP2012esim2.dat %>% tableX(maa,age_cat,type = "count")
# kable(taulu42,digits = 2, caption = "Ikäluokka age_cat")
#

# UUdet taulukot (4.2.20)

ISSP2012esim2.dat %>%
    tableX(maa,age_cat,type = "count") %>%
    kable(digits = 2, caption = "Ikäluokka age_cat")
```

Taulukko 61: Ikäluokka age_cat

	1	2	3	4	5	6	Total
BE	208	333	336	375	368	393	2013
BG	77	115	159	148	198	224	921
DE	205	223	274	358	288	366	1714
DK	207	213	245	271	234	218	1388
FI	152	166	165	223	238	166	1110
HU	103	161	198	171	196	168	997

	1	2	3	4	5	6	Total
Total	952	1211	1377	1546	1522	1535	8143

```
ISSP2012esim2.dat %>%
    tableX(maa,age_cat,type = "row_perc") %>%
    kable(digits = 2, caption = "age_cat: suhteelliset frekvenssit")
```

Taulukko 62: age_cat: suhteelliset frekvenssit

	1	2	3	4	5	6	Total
BE	10.33	16.54	16.69	18.63	18.28	19.52	100.00
BG	8.36	12.49	17.26	16.07	21.50	24.32	100.00
DE	11.96	13.01	15.99	20.89	16.80	21.35	100.00
DK	14.91	15.35	17.65	19.52	16.86	15.71	100.00
FI	13.69	14.95	14.86	20.09	21.44	14.95	100.00
HU	10.33	16.15	19.86	17.15	19.66	16.85	100.00
All	11.69	14.87	16.91	18.99	18.69	18.85	100.00

Ikäjäkauma painottuu kaikissa maissa jonkinverran vanhempiin ikäluokkiin. Nuorempien ikäluokkien osuus on (alle 26-vuotiaan ja alle 26-35 - vuotiaat) varsinkin Bulgariassa (BG) ja Unkarissa (HU) pieni.

zxy Siistimmät versioit muuttujien luonnista (case_when - rakenne) (19.9.2018).

```
# ga - ikäluokka ja sukupuoli

# case_when: ikä ja sukupuoli
ISSP2012esim2.dat <- mutate(ISSP2012esim2.dat, ga = case_when((age_cat == "1")&(sp == "m") ~
                                        (age_cat == "2")&(sp == "m") ~ "m2",
                                        (age_cat == "3")&(sp == "m") ~ "m3",
                                        (age_cat == "4")&(sp == "m") ~ "m4",
                                        (age_cat == "5")&(sp == "m") ~ "m5",
                                        (age_cat == "6")&(sp == "m") ~ "m6",
                                        (age_cat == "1")&(sp == "f") ~ "f1",
                                        (age_cat == "2")&(sp == "f") ~ "f2",
                                        (age_cat == "3")&(sp == "f") ~ "f3",
                                        (age_cat == "4")&(sp == "f") ~ "f4",
                                        (age_cat == "4")&(sp == "f") ~ "f4",
                                        (age_cat == "5")&(sp == "f") ~ "f5",
                                        (age_cat == "6")&(sp == "f") ~ "f6",
                                        TRUE ~ "missing"
                                    ))

#ISSP2012esim1.dat %>% tableX(ga,ga2) # tarkistus uudelle muuttujan luontikoodille
```

```
# muuttujien tarkistuksia 19.9.2018
str(ISSP2012esim2.dat$ga) # chr-muuttuja, mutta toimii (4.2.20)
```

```
##  chr [1:8143] "f5" "f3" "m5" "f2" "f4" "f4" "m4" "m3" "f5" "m5" "m3" "f5" ...
```

```
# str(ISSP2012esim2.dat)
# str(ISSP2012esim1.dat$ga2)
# ga on merkkijono, samoin ga2, pitäisikö muuttaa faktoriksi?
# str(ISSP2012esim1.dat)

#Tulostetaan taulukkoina ga2 - muuttuja.

ISSP2012esim2.dat %>% tableX(maa,ga,type = "count") %>%
kable(digits = 2, caption = "Ikäluokka ja sukupuoli ga")
```

Taulukko 63: Ikäluokka ja sukupuoli ga

	f1	f2	f3	f4	f5	f6	m1	m2	m3	m4	m5	m6	Total
BE	116	198	174	199	186	185	92	135	162	176	182	208	2013
BG	40	64	94	85	114	149	37	51	65	63	84	75	921
DE	102	120	152	186	135	185	103	103	122	172	153	181	1714
DK	83	110	136	146	128	99	124	103	109	125	106	119	1388
FI	94	95	94	118	142	91	58	71	71	105	96	75	1110
HU	54	86	95	91	94	104	49	75	103	80	102	64	997
Total	489	673	745	825	799	813	463	538	632	721	723	722	8143

```
ISSP2012esim2.dat %>% tableX(maa,ga,type = "row_perc") %>%
kable(digits = 2, caption = "ga: suhteelliset frekvenssit")
```

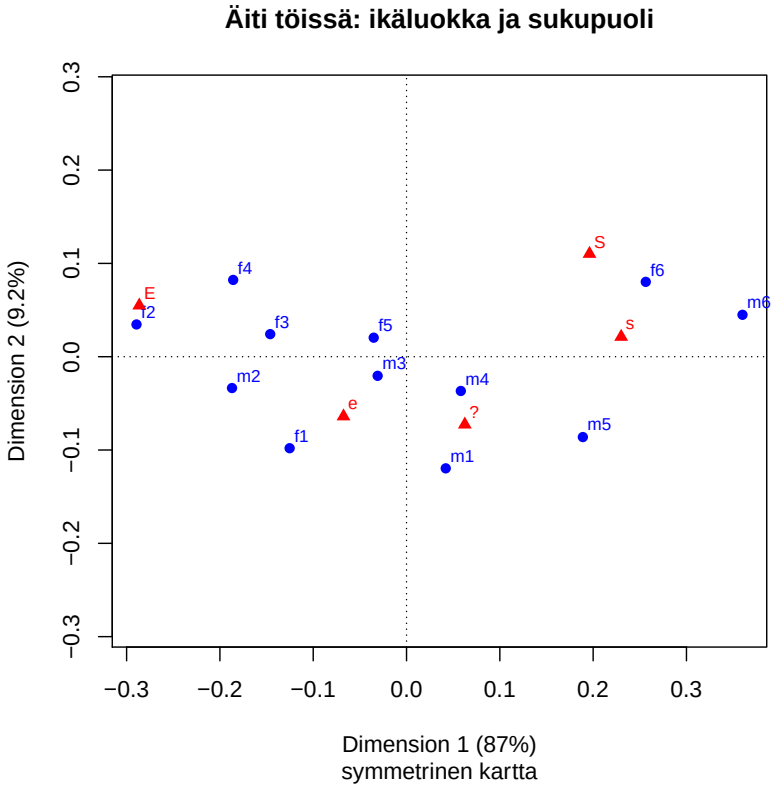
Taulukko 64: ga: suhteelliset frekvenssit

	f1	f2	f3	f4	f5	f6	m1	m2	m3	m4	m5	m6	Total
BE	5.76	9.84	8.64	9.89	9.24	9.19	4.57	6.71	8.05	8.74	9.04	10.33	100.00
BG	4.34	6.95	10.21	9.23	12.38	16.18	4.02	5.54	7.06	6.84	9.12	8.14	100.00
DE	5.95	7.00	8.87	10.85	7.88	10.79	6.01	6.01	7.12	10.04	8.93	10.56	100.00
DK	5.98	7.93	9.80	10.52	9.22	7.13	8.93	7.42	7.85	9.01	7.64	8.57	100.00
FI	8.47	8.56	8.47	10.63	12.79	8.20	5.23	6.40	6.40	9.46	8.65	6.76	100.00
HU	5.42	8.63	9.53	9.13	9.43	10.43	4.91	7.52	10.33	8.02	10.23	6.42	100.00
All	6.01	8.26	9.15	10.13	9.81	9.98	5.69	6.61	7.76	8.85	8.88	8.87	100.00

edit Vain tarkistuksiin, toisen voi poistaa (19.9.2018)!

CAiP, ch16, täällä myös maa- ja sukupuoli- uudelleenpainotus.

```
gaTestCA1 <- ca(~ga + Q1b,ISSP2012esim2.dat)
plot(gaTestCA1, main = "Äiti töissä: ikäluokka ja sukupuoli",
     sub = "symmetrinen kartta")
```



Kuva 16: Iän ja sukupuolen yhdistetty muuttuja

```
summary(gaTestCA1)
```

```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.037448  87.0  87.0  **********************
##  2      0.003977   9.2  96.2  **
##  3      0.001041   2.4  98.6  *
##  4      0.000590   1.4 100.0
##         -------- -----
##  Total: 0.043055 100.0
```

```
##
##
## Rows:
##      name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1  |   f1 |   60  990   36 | -125 614  25 |  -98 376 145 |
## 2  |   f2 |   83  997  163 | -289 983 185 |   35  14  25 |
## 3  |   f3 |   91  984   47 | -146 958  52 |   24  26  13 |
## 4  |   f4 |  101 1000   97 | -186 836  93 |   82 164 172 |
## 5  |   f5 |   98  879    4 |  -35 658   3 |   20 221  10 |
## 6  |   f6 |  100  951  176 |  256 866 175 |   80  85 162 |
## 7  |   m1 |   57  659   32 |   42  72   3 | -120 587 205 |
## 8  |   m2 |   66  977   57 | -187 946  62 |  -34  30  19 |
## 9  |   m3 |   78  457    5 |  -31 318   2 |  -20 139   8 |
## 10 |   m4 |   89  674   14 |   58 482   8 |  -37 192  30 |
## 11 |   m5 |   89  988   90 |  189 818  85 |  -86 170 166 |
## 12 |   m6 |   89  978  277 |  360 963 307 |   45  15  45 |
##
## Columns:
##     name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 |    S |   99  915  128 |  196 695 102 |  110 220 304 |
## 2 |    s |  238  969  304 |  230 961 336 |   21   8  27 |
## 3 |      |  168  777   46 |   62 330  17 |  -73 447 223 |
## 4 |    e |  261  897   58 |  -68 473  32 |  -64 424 268 |
## 5 |    E |  234  997  464 | -286 962 513 |   55  35 177 |
```

zxy Ei kovin kiinnostava, mutta voi verrata sekä edellisiin maa-vertailuihin että maan, ikäluokan ja sukupuolen yhteisvaikutusmuuttujan tuloksiin. MG tutkailee eri kysymyksellä tätä samaa asiaa, ja havaitsee että (a) maiden erot suuria ja sukupuolten pieniä (b) naiset liberaalimpia kuin miehet.

zxy miten pitäisi tulkita "oikealle kaatunut U - muoto" miehillä ja naisilla? Järjestys ei toimi, jotain muuta pelissä?

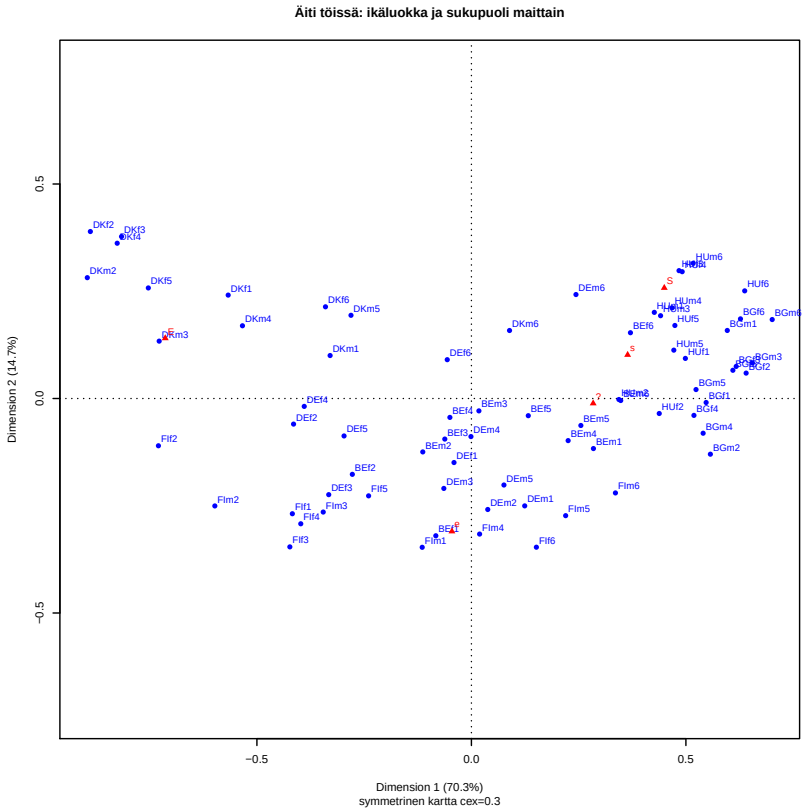
zxy On kiinnostava, mutta aika yksiuloitteinen (87 prosenttia ensimmäisellä dimensiolla!). **pisteet voisi yhdistää? (29.9.18)**

```
# Luodaan aineistoon kolmen muuttujan yhdysvaikutusmuuttuja maaga, maa, ikäluokka ja sukupuo
# Yleensä ei yhdysvaikuksissa mennä yli kolmen luokittelumuuttujan, ja tässäkin vain maiden
# tekee tarkastelun aika helpoksi.

ISSP2012esim2.dat <- mutate(ISSP2012esim2.dat, maaga = paste(maa, ga, sep = ""))

# tarkistus, muunnos ok
# ISSP2012esim2.dat %>% tableX(maa, maaga)

# head(ISSP2012esim2.dat)
# str(ISSP2012esim2.dat)
```

Maa - ikäluokka - sukupuoli - interaktiomuuttuja maaga

Tehty jo 26.9.2018!

```
maagaCA1 <- ca(~maaga + Q1b,ISSP2012esim2.dat)

par(cex = 0.5)
plot(maagaCA1, main = "Äiti töissä: ikäluokka ja sukupuoli maittain", "cex" = 0.3,
                sub = "symmetrinen kartta cex=0.3")
```



Kuva 17: Ikä, sukupuoli ja maa

```
#Kuvatiedoston koko säädettävä tarkemmin (30.3.20)

# Miten kuvatiedosto talteen?
```

Ratkaisun numeerisia tuloksia voi katsoa, löytyykö profiileja joilla on pieni massa mutta suuri vaikutus akseleihin.

```
# (24.2.20) Miten voisi kätevästi tarkistaa, että mikään pienen massa piste ei
# vaikuta (kontribuutiot) liikaa karttaan?
#str(maagaTestCA1)

ISSP2012esim2.dat %>% tableX(maaga, Q1b) # aika pieniä frekvenssejä soluissa!
```

maaga/Q1b	S	s	?	e	E	Total
BEf1	5	15	28	43	25	116
BEf2	10	26	34	66	62	198
BEf3	19	27	33	53	42	174
BEf4	21	34	40	55	49	199
BEf5	21	38	46	48	33	186
BEf6	25	58	50	30	22	185
BEm1	9	19	30	24	10	92
BEm2	10	19	31	40	35	135
BEm3	18	33	31	44	36	162
BEm4	19	46	37	51	23	176
BEm5	15	61	34	49	23	182
BEm6	19	75	44	49	21	208
BGf1	2	21	7	9	1	40
BGf2	7	28	17	12	0	64
BGf3	10	44	21	18	1	94
BGf4	14	30	15	24	2	85
BGf5	16	51	21	25	1	114
BGf6	27	66	26	27	3	149
BGm1	8	12	9	7	1	37
BGm2	4	21	12	14	0	51
BGm3	5	33	16	11	0	65
BGm4	7	19	21	15	1	63
BGm5	12	29	21	19	3	84
BGm6	6	41	19	9	0	75
DEf1	5	28	13	33	23	102
DEf2	9	14	14	37	46	120
DEf3	10	22	12	59	49	152
DEf4	11	31	20	53	71	186
DEf5	8	27	12	43	45	135
DEf6	31	40	15	50	49	185
DEm1	6	26	20	36	15	103
DEm2	7	26	13	39	18	103
DEm3	11	24	15	45	27	122
DEm4	22	39	17	57	37	172
DEm5	11	43	19	54	26	153
DEm6	34	55	28	32	32	181
DKf1	7	11	9	15	41	83
DKf2	4	15	7	13	71	110

maaga/Q1b	S	s	?	e	E	Total
DKf3	3	20	15	14	84	136
DKf4	5	24	8	19	90	146
DKf5	6	16	11	22	73	128
DKf6	5	26	11	17	40	99
DKm1	10	21	18	28	47	124
DKm2	2	11	9	16	65	103
DKm3	2	13	12	23	59	109
DKm4	4	24	14	24	59	125
DKm5	11	14	23	18	40	106
DKm6	11	43	15	23	27	119
FIf1	3	9	13	36	33	94
FIf2	5	6	3	34	47	95
FIf3	2	8	13	39	32	94
FIf4	3	15	13	47	40	118
FIf5	6	26	17	52	41	142
FIf6	3	22	21	34	11	91
FIm1	1	9	13	22	13	58
FIm2	2	5	6	28	30	71
FIm3	2	10	9	27	23	71
FIm4	8	23	13	43	18	105
FIm5	5	31	15	35	10	96
FIm6	7	24	13	26	5	75
HUf1	11	13	16	11	3	54
HUf2	15	19	25	22	5	86
HUf3	22	26	26	12	9	95
HUf4	24	25	20	14	8	91
HUf5	21	28	19	19	7	94
HUf6	33	30	18	21	2	104
HUm1	9	15	12	8	5	49
HUm2	18	13	15	22	7	75
HUm3	15	38	24	16	10	103
HUm4	14	29	17	13	7	80
HUm5	19	31	24	21	7	102
HUm6	18	21	9	11	5	64
Total	810	1935	1367	2125	1906	8143

```
maagaCA1num <- summary(maagaCA1)
# str(maagaCA1num) numeeriset tulokset tibbleksi - rivit
maagaCAnum2 <- as_tibble(maagaCA1num$rows, .name_repair = c("unique"))
```

```
## New names:
## * cor -> cor...6
## * ctr -> ctr...7
```

```
## * cor -> cor...9
## * ctr -> ctr...10
```

```
# maagaCAnum2
# str(maagaCAnum2)
summary(maagaCAnum2)
```

name	mass	qlt	inr	k=1	cor...6	ctr..
BEf1 : 1	Min. : 5.00	Min. :108.0	Min. : 1.00	Min. :-895.00	Min. : 0.0	Min. :
BEf2 : 1	1st Qu.:10.00	1st Qu.:704.8	1st Qu.: 6.75	1st Qu.:-330.00	1st Qu.:351.2	1st Qu.:
BEf3 : 1	Median :13.00	Median :838.0	Median :11.00	Median : 82.50	Median :667.5	Median :
BEf4 : 1	Mean :13.97	Mean :772.6	Mean :13.88	Mean : 46.49	Mean :573.3	Mean :1
BEf5 : 1	3rd Qu.:17.00	3rd Qu.:953.2	3rd Qu.:15.00	3rd Qu.: 472.50	3rd Qu.:830.2	3rd Qu.:
BEf6 : 1	Max. :26.00	Max. :999.0	Max. :57.00	Max. : 701.00	Max. :982.0	Max. :6
(Other):66	NA	NA	NA	NA	NA	NA

```
arrange(maagaCAnum2 ,mass)
```

name	mass	qlt	inr	k=1	cor...6	ctr...7	k=2	cor...9	ctr...10
BGf1	5	531	11	547	531	8	-9	0	0
BGm1	5	940	7	596	878	9	159	62	3
BGm2	6	830	9	557	788	11	-130	43	3
HUm1	6	935	5	426	766	6	201	170	6
FIm1	7	787	5	-115	78	1	-347	710	22
HUf1	7	723	9	499	698	9	93	25	1
BGf2	8	860	14	640	853	17	59	7	1
BGm3	8	709	19	655	698	19	83	11	1
BGm4	8	771	11	540	754	12	-81	17	1
HUm6	8	726	15	517	529	11	315	197	20
BGm6	9	692	27	701	647	24	184	45	8
FIm2	9	977	14	-598	832	17	-250	146	14
FIm3	9	998	6	-345	629	6	-265	369	16
FIm6	9	911	6	336	637	6	-220	274	12
HUm2	9	381	11	344	381	6	-2	0	0
BGf4	10	932	12	519	927	15	-39	5	0
BGm5	10	979	11	524	977	15	21	2	0
DKf1	10	991	15	-567	839	18	241	152	15
HUm4	10	999	10	468	830	12	211	169	11
BEm1	11	429	9	284	367	5	-117	62	4
FIf6	11	835	7	151	134	1	-347	701	35
HUf2	11	689	11	438	685	11	-35	4	0
HUf4	11	768	18	491	564	15	296	204	25
BGf3	12	815	21	617	804	24	75	12	2
DKf6	12	808	9	-340	579	8	214	229	14

name	mass	qlt	inr	k=1	cor. . . 6	ctr. . . 7	k=2	cor. . . 9	ctr. . . 10
FIf1	12	980	11	-417	693	11	-269	287	21
FIf2	12	927	26	-730	907	34	-110	21	4
FIf3	12	984	13	-423	590	11	-346	394	36
FIm5	12	734	7	220	289	3	-273	446	23
HUf3	12	808	18	484	586	15	298	222	27
HUf5	12	850	13	474	753	14	170	97	9
DEf1	13	425	3	-41	29	0	-149	395	7
DEm1	13	912	4	124	180	1	-250	732	20
DEm2	13	766	4	38	16	0	-259	749	22
DKm2	13	989	43	-895	900	55	282	89	26
DKm3	13	982	28	-728	950	38	134	32	6
DKm5	13	643	9	-281	435	6	194	208	13
FIm4	13	837	6	19	3	0	-316	834	33
HUf6	13	671	34	637	581	28	251	90	21
HUm3	13	957	12	441	803	13	193	154	12
HUm5	13	942	12	472	891	15	113	51	4
BEf1	14	678	9	-83	43	1	-320	635	38
BGf5	14	880	23	609	870	28	66	10	2
DKf2	14	991	49	-888	831	58	389	160	53
FIf4	14	991	14	-398	644	12	-292	347	32
DEf2	15	938	10	-415	919	14	-60	19	1
DEm3	15	737	4	-64	63	0	-210	674	17
DKm1	15	981	7	-329	898	9	100	83	4
DKm4	15	941	19	-534	855	24	170	86	11
DKm6	15	355	5	89	85	1	158	270	9
DKf5	16	998	38	-753	894	48	258	105	27
BEm2	17	372	5	-113	169	1	-125	203	7
DEf5	17	839	7	-297	772	8	-87	67	3
DKf3	17	963	53	-816	793	60	377	170	61
FIf5	17	952	8	-240	502	5	-227	450	23
BGf6	18	921	32	627	846	39	186	74	16
DKf4	18	977	57	-826	820	66	362	157	61
DEf3	19	846	13	-333	582	11	-224	264	24
DEm5	19	603	5	76	75	1	-202	529	20
BEm3	20	108	1	17	29	0	-29	79	0
BEf3	21	320	3	-62	96	0	-95	224	5
DEm4	21	137	5	-1	0	0	-89	137	4
BEm4	22	966	5	225	812	6	-98	154	5
BEm5	22	728	8	255	686	8	-63	42	2
DEm6	22	849	12	244	427	7	242	422	34
BEf5	23	332	5	133	304	2	-40	28	1
BEf6	23	832	17	371	710	17	153	121	14
DEf4	23	985	13	-390	982	19	-18	2	0
DEf6	23	116	8	-56	32	0	90	84	5

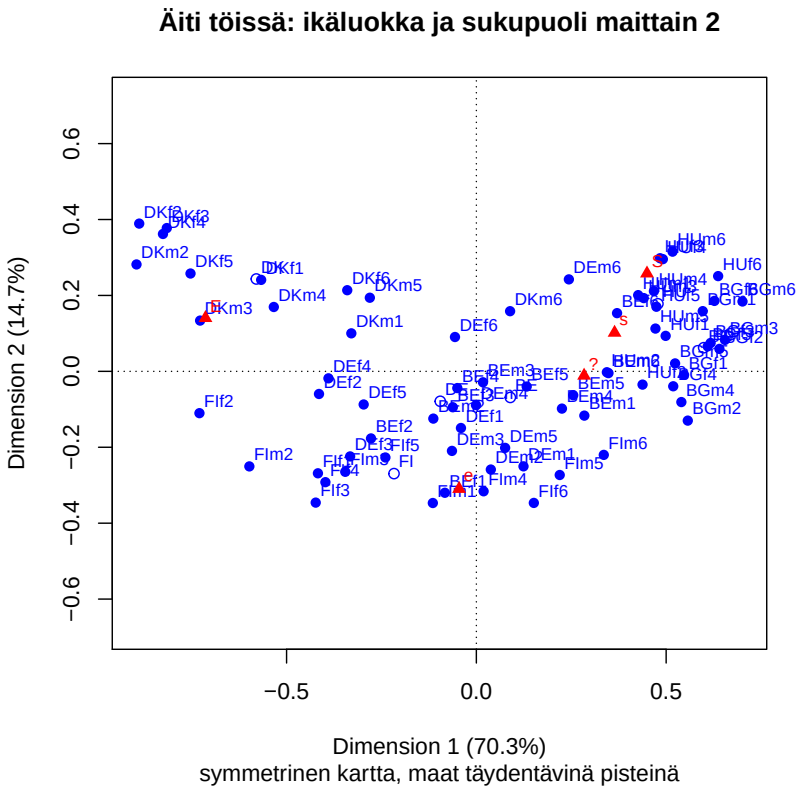
name	mass	qlt	inr	k=1	cor...6	ctr...7	k=2	cor...9	ctr...10
BEf2	24	914	11	-278	650	10	-177	264	20
BEf4	24	164	3	-50	92	0	-44	71	1
BEm6	26	788	15	348	788	17	-5	0	0

```
# plot(maagaCAnum2, x = c("mass"), y = c("ctr...7"), xlim = c(0,30), ylim = c(0, 1000))


# Hieman hankalaa kätevästi järjestää numeerisia tuloksia massan mukaan

#str(maagaCA1num)
#maagaCA1num$rows
#maagaRows.df <- maagaCA1num$rows
# sarakenimet eivät yksikäsitteisiä
#maagaRows.df
#str(maagaRows.df)
#names(maagaRows.df)
#str(maagaRows.df$mass)
# ei toimi AscmaagaRows.df <- maagaRows.df[order(mass),]
```

Maapisteet täydentäviksi pisteiksi - tarkistuksia.

```
# Miten maa-rivit täydentäviksi riveiksi - alla siisti ratkaisu
# Miten labelit hieman lähemmäkis pistettä? offset-jotenkin toimii...

# rakennetaan taulukko, jossa alimpina riveinä "maa-rivit"
# otetaan karttaan mukaan täydentävinä pisteinä
# karttaa on helpompi tulkita, kun nähdään miten ikä-sukupuoli-ryhmät sijatsevat keskiarvons

#ikäluokka - sukupuoli ja maa - maaga-muuttuja
maagaTab1 <- table(ISSP2012esim2.dat$maaga, ISSP2012esim2.dat$Q1b)
#dim(testTab1) #72 riviä,  5 saraketta


# maa-rivit
maagaTab_sr <- table(ISSP2012esim2.dat$maa, ISSP2012esim2.dat$Q1b)
#maagaTab_sr

maagaTab1 <- rbind(maagaTab1,maagaTab_sr)
# str(maagaTab1)
# maagaTab1
# dim(maagaTab1) #78 riviä, 5 saraketta, 1-72 data ja 73-78 täydentävät rivit

spCAmaaga1 <- ca(maagaTab1[,1:5], suprow = 73:78)
#X11()
```

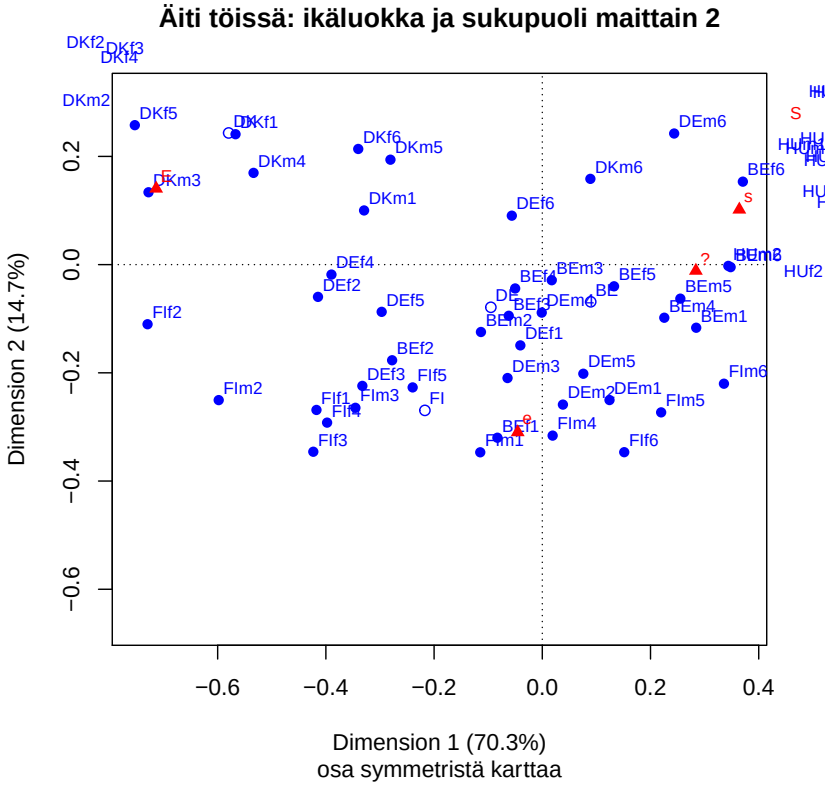
```
# Plot toimii (4.2.20), mutta par() ei, sama virheilmoitus (varoitus)
# kuin edellisessä koodilohkossa (24.2.20)
# par("cex"= 0.75, "asp" = 1, "offset" = 0.5)
plot(spCAmaaga1, main = "Äiti töissä: ikäluokka ja sukupuoli maittain 2",
                 sub = "symmetrinen kartta, maat täydentävinä pisteinä"
                     )
```



Kuva 18: Ikä-sukupuoli-maa

```
#par()

#asymmetrinen kartta
#X11()
#par("cex"= 0.75, "asp" = 1, "offset" = 0.5)
#plot(spCAmaaga1, main = "Äiti töissä: ikäluokka ja sukupuoli maittain 3 (kontribuutiot) -
#                  map = "rowgreen",
#                  contrib= c("absolute", "absolute"),
#                  mass = c(TRUE,TRUE),
```

```
#                   arrows = c(FALSE,TRUE)
#                   )

# Zoomaus - esimerkki (24.2.20) xlim=c(-0.5,0.5), ylim=c(-0.6,0.4)
# EI TOIMI

plot(spCAmaaga1, xlim = c(-0.75,0.37), ylim = c(-0.4,0.05),
                 main = "Äiti töissä: ikäluokka ja sukupuoli maittain 2",
                 sub = "osa symmetristä karttaa"
                     )
```



Kuva 19: Ikä-sukupuoli-maa

```
# ei toimi ihan toivotulla tavalla - tarkoitettu komentoriviltä
# grafiikkaikkunaan tulostukseen ?
```

Kuvissa on aika ahdasta. Kuvan voisi rajata johonkin alueeseen erityisesti oikea yläosa on täynnä pisteitä. Maiden täydentävät pisteet ovat ikäluokka-sukupuoli -luokkien keskiarvopisteitä. Maiden väliset erot dominoivat, mutta maiden välillä

on isoja eroja.

Kartan herkkyyttä joillekin pienen massan rivipisteille pitää tutkia tarkemmin.

Vertailu voi tehdä

1.Maiden sisällä, ikä-sukupuoli - luokkien välillä. Ovatko naiset kaikissa ikäluokissa mies-ikäluokkien oikealla vai vasemmalla puolella?

2.Maiden välillä

a. miten ikä-sukupuoliluokat sijaitsevat suhteessa maiden keskiarvopisteisiin

b. mikä on niiden järjestys

---

5 Yksinkertaisen korrespondenssianalyysin laajennuksia 2

TODO Vielä kuuden maan aineistolla ilman puuttuvia havaintoja? Helpompi havainnollistaa taulukoiden pinoamista / liittämistä (concate, stack). Ja voisi jatkaa ehkä pari pointtia pienellä aineistolla?

```
# str(ISSP2012jh1d.dat)
# Yksinkertaisuuden vuoksi muuttujat tähän

isodatVars1 <- ISSP2012jh1d.dat %>% names()
isodatVars1 <- isodatVars1[24:73]
demogrVars1 <- c("maa","maa3","sp","ika")
isodatVars1 <- isodatVars1[21:50]
isodatVars1 <- c(demogrVars1, isodatVars1)
isodatVars1
```

```
##  [1] "maa"     "maa3"    "sp"      "ika"     "Q1a"     "Q1b"     "Q1c"
##  [8] "Q1d"     "Q1e"     "Q2a"     "Q2b"     "Q3a"     "Q3b"     "edu"
## [15] "msta"    "sosta"   "nchild"  "lifsta"  "urbru"   "Q1am"    "Q1bm"
## [22] "Q1cm"    "Q1dm"    "Q1em"    "Q2am"    "Q2bm"    "Q3am"    "Q3bm"
## [29] "edum"    "mstam"   "sostam"  "nchildm" "lifstam" "urbrum"
```

```
ISSP2012jh1d.dat %>% select(all_of(isodatVars1)) %>%
    summary()
```

maa	maa3	sp	ika	Q1a	Q1b	Q1c
FR : 2409	FR-France : 2409	m:14789	Min. : 15.00	S :11116	S :2747	S :2838
BE : 2192	CZ-Czech Republic: 1804	f:18034	1st Qu.: 36.00	s :12352	s :8389	s :8263
CZ : 1804	AU-Australia : 1557	NA	Median : 50.00	? : 3382	? :5949	? :6000
DE : 1761	RU-Russia : 1525	NA	Mean : 49.52	e : 4074	e :9003	e :8706
AU : 1557	NO-Norway : 1444	NA	3rd Qu.: 63.00	E : 1051	E :5547	E :5960

maa	maa3	sp	ika	Q1a	Q1b	Q1c
RU : 1525	DK-Denmark : 1403	NA	Max. :102.00	NA's: 848	NA's:1188	NA's:105
(Other):21575	(Other) :22681	NA	NA	NA	NA	NA

Data on valmiina, edellisen luvun ikäluokka, ikä-sukupuoli- muuttuja ja ikä-sukupuoli- maa muuttujien luontia voi harkita.

edit Tässä keskityttävä data-analyysin **tutkimusongelmiin**, johdantoa MCA-lukuun.

5.1 Päällekkäiset matriisit (stacked matices)

Ref:CAip, CA_Week2.pdf (kalvot MCA-kurssilta 2017)

Concatenated tables (yhdistetyt taulut tai matriisit): (a) kaksi luokittelumuuttujaa (b) useita muuttujia stacked ("pinotaan").

MCA 2017 laskareissa ja kalvoissa esitetään, miten nämä saadaan kätevästi CA-paketin MJCA-funktion BURT-optiolla.

5.2 Matched matrices

Ref:CAip ss. 177, HY2017_MCA, Greenacre JAS 2013 (sovellus ISSP 1989, 4 kysymystä 'pitäisikö äidin olla kotona', 8 maata), tässä artikkelissa "SVD-based methods", joista yksi CA (muut biplots, PCA, compositional data/log ratios).

Edellisen menetelmän variantti, jossa ryhmien väliset ja sisäiset erot saadaan esiin. Inertian jakaminen. Samanlaisten rivien ja sarakkeiden kaksi samankokoista taulua, esimerkiksi sukupuolivaikutusten arviointi. Alkuperäinen taulukko jaetaan kahdeksi tauluksi sukupuolen mukaan. Matriisien yhdistäminen (concatenation) riveittäin tai sarakkeittain ei näytä optimaalisesti mm - matriisien eroja.

Ryhmien välisen ja ryhmien sisäinen inertian erottaminen, **ABBA** on yksi ratkaisu (ABBA matrix, teknisesti block circulanMat matrix).

Luokittelu voi olla myös kahden indikaattorimuuttujan avulla jako neljään taulukkoon (esim. miehet vs. naiset länsieuroopassa verratuna samaan asetelmaan itä-Euroopassa). Samaa ideaa laajennetaan.

Esimerkkinä "Attitudes to women working in 2012".