Bookdown-dokumetti - testi 1

Jussi Hirvonen

2018-07-03

Sisältö

Luku 1

Alkutoimia

1.1 Tärkeimmät ohjelmistot

```
system("pdflatex --version")
#getwd()

rmarkdown::pandoc_version()
```

```
## [1] '2.2.1'
```

Viimeinen rivi kertoo pandoc-version.

1.2 YAML-headerin säätöä

YAML-metadataa voi olla “kotisivulla”, jonka nimi oletuksena on index.Rmd. Lisäksi määrityksiä on tiedostoissa _output.yml ja _bookdown.yml. Näiden hierarkia on hieman epäselvä.

En säädä pdf-tulostusta, nyt toimii gitbook ja pdf_book tulostusformaatteina. Molemmat ovat html-paketteja, ja tarvitsevat ehkä r-datahakemistosta (omalta koneelta) libs-hakemiston jQuery- ja Gitbook-paketit (javaskriptiä ja css-tyylitiedostoja).

Bookdown-tulostuksessa voisi käyttää “one document” - optioita (funktioita). EI vileä kokeiltu.

Luku 2

Johdanto

Bookdown - formaatissa “juuritiedoston” indexBD.Rmd tekstit eivät tulostu jos siellä ei ole luvun (chapter) aloittavaa ensimmäisen tason otsikkoa.Siellä on YAML-headeri (metadata).

Lisää YAML-parametreja voi antaa tiedostoissa _bookdown.yml ja _output.yml. Nämä lienee välittyvät Pandocille?

Bookdown - demon esimerkkitiedostot ovat nämä:

ouput.yml (huomaa, että _ - merkki jätetty pois!) (tässä oli bookdown-demo-paketin yml-tiedostot, poistin 3.7.2018)

Nyt toimii gitbook ja pdf_book tulostusformaatteina. Molemmat ovat html-paketteja, ja tarvitsevat ehkä r-datahakemistosta (omalta koneelta) libs-hakemiston jQuery- ja Gitbook-paketit (javaskriptiä ja css-tyylitiedostoja).

Bookdown-tulostuksessa voisi käyttää “one document” - optioita (funktioita).

3.7.2018 PDF-tulostuksen säätöä, nyt saadaan jo virheilmoituksiakin! Piti tallentaa utf-8 - muodossa kertaalleen. Tulostus kaatuu valituksiin puuttuvista $-merkeistä kaavoissa (test_kaavat1.rmd). TeX-tiedoston voi kääntää PDF:ksi, mutta kaavat sekaisin ja paljon muutakin. Esim. sisällysluettelo.

Merkistöt olivat pielessä (ei utf-8!), ja samoin test1_preamble.tex -tiedosto. Laitoin kuntoon, nyt asiallinen virheilmoitus:

! LaTeX Error: Two documentclass or documentstyle commands.

Error: Failed to compile indexBD.tex. See indexBD.log for more info. Execution halted Poistetaan test1_preambe.tex eka rivi kokonaan: % documentclass[12pt,a4paper,leqno] (alusta puuttuu takakeno) % dispositiopaperista, poistettiin ekalta riviltä {article} Lisäksi usepackage[**utf8**]{inputenc}.

LaTeX Font Info: Redeclaring font encoding T1 on input line 48.))

! LaTeX Error: Option clash for package babel.

Palautetaan eka rivi, documeteclass {book}. Taas virheilmoitus ! LaTeX Error: Two documentclass or documentstyle commands. Poistetaan eka rivi. Ja sama virheilmoitus “option clash”.

Luku 3

Kaavat ja matemattiset merkinnät

Kaavat on esitettävä bookdown-paketin määrityksillä. Viittausnimien on oltava yksikäsitteisiä koko dokumentissa, jos käytetään “merge and knit” menetelmää. Jos taas jokainen lapsidokumentti on “itsenäinen” (“knit and merge”), tämä koskee vain kyseistä dokumenttia (kts. Bookdown - webkirja).

3.1 Kahden luokittelumuuttuja taulukko

Kahden luokittelumuuttujan riippuvuutta voidaan testata χ^2 - testillä. Testisuure saadaan laskemalla yhteen jokaisen solun havaittujen ja odotetettujen (riippumattomuushypoteesi) frekvenssien erotukset muodossa

$$\chi^2 = \frac{(havaittu - odotettu)^2}{odotettu} \tag{3.1}$$

Tämä voidaan esittää ca:han sopivammalla tavalla parilla muunnoksella, jolloin saamme riveittäin vastaavat termit rivisummalla painotettuna:

$$\text{rivisumma} \times \frac{(havaittu\ riviprofiili - odotettu\ riviprofiili)^2}{odotettu\ riviprofiili} \tag{3.2}$$

Kun jaamme nämä tekijät havaintojen kokonaismäärällä n, rivisumma muuntuu rivin massaksi, ja niiden summa muotoon $\frac{\chi^2}{n}$.

$\chi^2 \frac{}{n = \phi^2}$

(3.3)

Tunnusluku ϕ^2 on korrespondenssianalyysissä kokonaisinertia (total inertia). Se kuvaa, kuinka paljon varianssia taulukossa on ja on riippumaton havaintojen lukumäärästä. Tilastotieteessä tunnusluvulla on useita vaihtoehtoisia nimiä (esim. mean square contingency coefficient), ja sen neliöjuurta kutsutaan ϕ - kertoimeksi.

Tässä siirrytään kahden luokittelumuuttujan taulukosta suhteellisten frekvenssien taulukkoon, ja pieni pohdinta taulukoista yleensä olisi paikallaan. Yhtälöihin voi viitata (**??**) . Kokeillaan vielä, toimivatko kirjallisuusviittet, kuten tärkeä lähde(**?**).

Luku 4

Taulukot ja kuvat

Tähän taulukoita ja kuvia, esimerkkiaineistoilla.

Kirjallisuutta on myös (**?**), ja (**?**) esittelee geometrisen tulkinnan peruskäsitteet yksinkertaisen kahden luokittelumuuttujan korrespondenssianalyysin avaulla.

Viitteet saa tulostusasetuksilla yhdelle sivulle, oletuksena on viitteiden esittäminen jokaisen sivun alareunassa.

4.1 Taulukoita

```
#käytetään CA-paketin smoke - dataa
data("smoke")
```

Taulukot tulostetaan funktiolla knitr::kable(). Taulukko numeroidaan ja se saa automaattisesti labelin etutunnisteella 'tab', ja siihen liitetään chunk-label (esim alla tab:smoketable1).

Tämä koodipätkä ei antaa yhden kappaleen esikatselussa virheilmoituksen, "smoke"-dataa ei löydy.

```
knitr::kable(smoke[,1:4], booktabs = TRUE,
  caption = 'CA-paketin smoke-data (keinotekoinen)'
)
```

```
# Taulukkoon viittaaminen tekstissä \@ref(label)
```

Taulukossa **??** on kahden luokittelumuuttujan keinotekoinen esimerkkiaineisto tupakonnin määrästä henkilöstöryhmittäin (SM = senior managemet, JM = junior management, SE ja JE vastaavasti ryhmälle employee, SC = secretary).

Taulukko 4.1: CA-paketin smoke-data (keinotekoinen)

	none	light	medium	heavy
SM	4	2	3	2
JM	4	3	7	4
SE	25	10	12	4
JE	18	24	33	13
SC	10	6	7	2

Taulukko 4.2: Riviprofiilit ja keskiarvoprofiili

	none	light	medium	heavy
SM	0.364	0.182	0.273	0.182
JM	0.222	0.167	0.389	0.222
SE	0.490	0.196	0.235	0.078
JE	0.205	0.273	0.375	0.148
SC	0.400	0.240	0.280	0.080

none	light	medium	heavy
0.316	0.233	0.321	0.13

Useampi taulukko saadaan taulukkoympäristöön (table environment) yhdistämällä data-objektit listaksi.

```
# riviprofiilit
smoke.rpro <- smoke / rowSums (smoke)
# keskiarvoprofiili
smoke.avrpro <- colSums(smoke) / sum(smoke)

knitr::kable(
  list(smoke.rpro, t(smoke.avrpro)   ), digits = 3,
  caption = 'Riviprofiilit ja keskiarvoprofiili', booktabs = TRUE
)
```

Taulukossa **??** on laskettu jokaisen rivin riviprofiilit. Ne saadan kun rivin luvut jaetaan rivin summalla. Yhden rivin taulukossa on esitetty riviprofiilien keskikarvo, sarakesummat jaettuna koko taulukon havaintojen lukumäärällä. Sen prosenttiluvut kertovat tupakoititapojen jakauman koko henkilöstössä.

Jos PDF-tulostuksessa ei haluta ns. kelluvaa taulukkoa (float), voi kable-funktiossa käyttää LaTeXin pakettia longtable. Silloin on myös muistettava ottaa paketti käyttöön (usepackage{}) LaTeX - pohjatiedostossa (preamble).

Pandoc tukee monia Markdownin taulukkotyyppejä. Viittaaminen vaaati labeloidun otsikon, ja sen on oltava otsikkotestin alussa määrämuotoisena (esim. ab:hienotaulu). Tämä vaatii tarkkuutta, jos taulukon pitää toimia html- ja LaTex-outputissa. kable-funktiota kannattaa käyttää!

4.2 Korrespondenssianalyysin numeeriset tulokset taulukoina

Korrespondenssianalyysin idea on vähentää aineiston dimensioita, ja esittää taulukon rivien ja sarakkeiden riippuvuudet yleensä kaksiulotteisena karttana.

```
smokeCA <- ca(smoke)
temp1 <- smokeCA
numres1CA1 <- summary(smokeCA)
#str(smokeCA)
#knitr::kable( smokeCA,
#  digits = 3,
#  caption = 'Riviprofiilit ja keskiarvoprofiili', booktabs = TRUE
#)
#str(temp1)
#stargazer(temp2$rows, type = "text", title = "CA-tuloksia")
# LateX-tulostuksessa float vaatii jotain tällaista:Table: (\#tab:cataul1)
#str(temp2)
#str(temp2$scree)
#temp2$scree
```

Taulukko 4.3: Korrespondenssianalyysin diagnostiikkaa - rivit

name	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
SM	57	893	31	-66	92	3	-194	800	214
JM	93	991	139	259	526	84	-243	465	551
SE	264	1000	450	-381	999	512	-11	1	3
JE	456	1000	308	233	942	331	58	58	152
SC	130	999	71	-201	865	70	79	133	81

Taulukko 4.4: Korrespondenssianalyysin diagnostiikkaa - sarakkeet

name	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
none	316	1000	577	-393	994	654	-30	6	29
lght	233	984	83	99	327	31	141	657	463
medm	321	983	148	196	982	166	7	1	2
hevy	130	995	192	294	684	150	-198	310	506

Taulukot ovatkin aika vaikeita, tulostiedoista! Stargazer toki tekee monenlaista, mutta kun kyse on hyvin yksinkertaisista tulostaulukoista kablen pitäisi toimia.

Kokeillaan summary(smokeCA) - listan dataframe-olioden tulostusta kablella. Voisi harkita funktiota, joka poimii CA:n tuloslistasta sopivat objektit kable-funktiolle? Stargazer taas vaatisi (luultavasti) jonkun ehdollisen tulostuksen (PDF ja html)?

```
knitr::kable( numres1CA1$rows,
    digits = 3,
    caption = 'Korrespondenssianalyysin diagnostiikkaa - rivit', booktabs = TRUE
)
```

Rivien ja sarakkeiden diagnotiikkataulukot eivät mahdu rinnakkain, siksi ne tulostetaan erikseen.

```
knitr::kable( numres1CA1$columns,
    digits = 3,
    caption = 'Korrespondenssianalyysin diagnostiikkaa - sarakkeet', booktabs = TRUE
)
```

Taulukoiden **??** ja **??** luvut on kerrottu tuhannella (“per milles”).

Dimensioiden ominaisarvot (eli niiden osuus kokonaisinertiasta) saadaan ca-funktion tuloslistasta taulukoksi. Se esitetään joskus myös ns. scree - kuvana, jos dimensoita on paljon ja joudutaan pohtimaan kuinka monta valitaan (vaikea kysymys!).

```
knitr::kable( numres1CA1$scree,
    digits = 3,
    caption = 'Korrespondenssianalyysin diagnostiikkaa - ominaisarvot', booktabs = TRUE
)
```

Taulukko **??** vaatii selityksen, mutta kuvaa ei tässä tapauksessa tarvita.

Taulukko 4.5: Korrespondenssianalyysin diagnostiikkaa - ominaisarvot

	values	values2	values3
1	0.075	87.756	87.756
2	0.010	11.759	99.515
3	0.000	0.485	100.000

4.3 Kuvat

chunk-optiot

fig.cap: R plot - kuvat figure-ympäristöön, automaattiset labelit (fig: + koodipätkän label) ja niihin voi viitata.

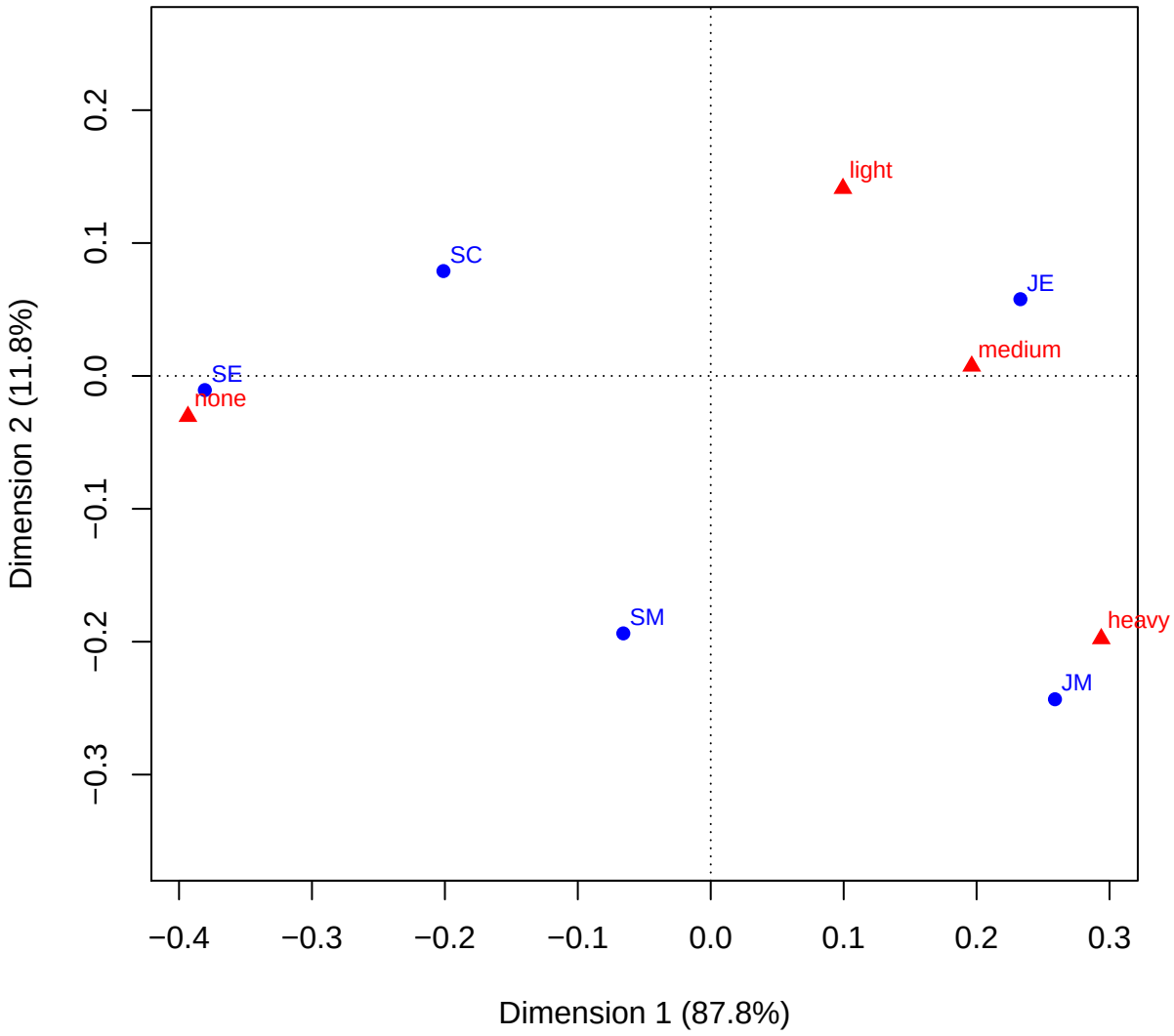
fig.asp oikeaan arvoon 1.

```
plot(smokeCA)
```

Kuviin (kuten **??**) ja taulukoihin voi viitata tekstissä. Kuvan otsikko tulostuu kuvan alapuolelle, ehkä vähän huono idea?

Näköjään stargazer-kokeilu tulostusoptiolla “html” loi R-projektihakemistoon kansion ja sinne png-kuvan. finnish.ldf tiedoston muokkaus MikTeX:ssä tehty, mutta se ei vaikuta html-viiteotsikkoon. Korjattu “ehdollisessa viitesivussa” viitteet.Rmd jossa html-viiteluettelon otsikko annetaan.

Saisiko numeeristen tulosten scree-kuvan samalla tavalla kuvaksi?



Kuva 4.1: CA-kartta

Lïhteet