

Rakendustarkvara: R. Kevad 2018, 3. praktikum*

1 Toimingud andmestikuga

Impordime uuesti andmestiku, mida korra kasutasime ka eelmises praktikumis (Massachusettsi andmestik), ja vaatame andmestikust ülevaadet

```
andmed <- read.table("http://kodu.ut.ee/~annes/Rkursus/mass.txt", header = T, sep = "\t")
str(andmed)
```

1.1 Veergude ja ridade eraldamine andmestikust: indeksi ja nime järgi

Kui tahame andmestikust ainult üht veergu uurida, siis kõige mugavam on kasutada dollari-sümbolit:

```
vanused <- andmed$AGEP
median(andmed$AGEP)
median(vanused)
```

Üldiselt on `data.frame` kahemõõtmeline tabel, mis tähendab, et iga elemendi asukoht selles tabelis on ära määratud rea ja veeru numbriga. Rea- ja veerunumbrite abil andmestikust infot eraldades tuleb kasutada kantsulgusid:

```
andmed[3, 2] # kolmas rida, teine veerg
andmed[, 2]  # kogu teine veerg
andmed[3, ]  # kogu kolmas rida
```

Korruga on võimalik eraldada ka mitut rida või veergu, kasutades selleks käsku `c()`:

```
andmed[, c(2, 4)] # teine ja neljas veerg
valik <- c(2, 4)  # tekitame objekti, milles on kirjas huvipakkuvate veergude numbrid
andmed[, valik]   # kasutame seda objekti andmestikust veergude eraldamiseks
andmed[c(5, 3, 9), ] # viies, kolmas ja üheksas rida
```

Tihti on veeruindeksite asemel mugavam kasutada veergude nimesid (peavad olema jutumärkides):

```
andmed[, c("AGEP", "WAGP")] # eraldame veerud "AGEP" ja "WAGP"
```

Kui andmestikus on ka ridadel nimed, siis saab neid sama moodi kasutada ridade eraldamisel. Selle läbiproovimiseks lisame oma andmestikule praegu ise reanimed:

```
rownames(andmed) <- paste("rida", rownames(andmed), sep = "-") # paneme ise nimed kujul 'rida-jrk'
head(andmed[, 5:9]) # vaatame milline on tulemus
andmed[c("rida-23", "rida-62"), 5:9] # eraldame read 23 ja 62
```

1.1.1 Uue tunnuse lisamine

Uue tunnuse lisamiseks andmestikku tuleb valida tunnuse nimi, mida andmestikus veel ei esine, ja omistada uuele andmeveerule valitud väärtused. Lisame oma andmestikku näiteks keskmise kuusissetuleku veeru, mis on arvutatud olemasoleva 12 kuu sissetuleku tunnuse `WAGP` abil. Uus tunnus lisatakse viimaseks veeruks andmestikus:

*Praktikumijuhendid põhinevad aine MTMS.01.092 Rakendustarkvara: R (2 EAP) materjalidel, mille autorid on Mait Raag ning Raivo Kolde

```
andmed$kuusissetulek <- andmed$WAGP/12
# või
andmed[, "kuusissetulek"] <- andmed$WAGP/12
str(andmed)
```

1.2 Veergude ja ridade eraldamine andmestikust: tõeväärtusvektori abil

Tõeväärtusvektoreid on väga mugav kasutada nn filtritena. Nimelt on `data.frame` puhul võimalik ridu (ja ka veerge!) eraldada mitte ainult indeksi või -nime järgi, vaid ka tõeväärtusvektori abil. Kui eraldada ridu, siis vastav tõeväärtusvektor peab olema sama pikk kui on andmestikus ridu ning väärtused selles vektoris näitavad, kas vastavat rida kasutada (`TRUE`) või mitte(`FALSE`). Tõeväärtusvektorite kombineerimisel saab andmestikust väga spetsiifilisi alamhulki eraldada.

```
# eraldame kõik read, kus SEX == "Male" ning salvestame selle uueks objektiks
mehed <- andmed[andmed$SEX == "Male", ]

# moodustame kaks filtritunnust ja kombineerime need alagrupi valikuks
filter_kod <- andmed$CIT == "Not a citizen of the U.S." # mittekodanikud
filter_vanus <- andmed$AGEP >= 80                      # vähemalt 80 aastased
alamandmestik <- andmed[filter_kod & filter_vanus, ]    # Ära unusta: [read, veerud]
```

Üks näide ka tõeväärtusvektori abil veergude eraldamise kohta: valime kõik need veerud, mille päis algab sõnega *MAR* (tunnused, mis seotud abieluga)

```
onMAR <- startsWith(names(andmed), "MAR")
# või
#onMAR <- substr(names(andmed), 1, 3) == "MAR"
abielu <- andmed[, onMAR]
str(abielu)
```

1.2.1 Ülesanded (Massachusettsi andmestik)

1. Selekteeri andmestikust iga 5. rida ja salvesta see alamandmestik uue nimega `valik5`. Mitu vaatlust on selles andmestikus?
2. Moodusta alamandmestik, kuhu kuuluvad uuritavad, kelle kohta pole teada, kas nad on viimase aasta jooksul elukohta vahetanud. Kes sellesse andmestikku kuuluvad?

1.3 Lihtsam kirjeldav statistika

Allpool on loetletud mõned käsud, mille abil mõnd konkreetset tunnust (veergu andmestikust) kirjeldada.

- `min(tunnus)`, `max(tunnus)`, `median(tunnus)`, `mean(tunnus)`, `sd(tunnus)` – arvulise tunnuse karakteristikud
- `quantile(tunnus, kvantiil)` – saab leida kvantiile ehk protsentiile arvulisele tunnusele
- `length(tunnus)` – mitu elementi on antud veerus
- `table(tunnus)` – saab koostada sagedustabelit (kasulik `Factor`-tüüpi tunnuse kirjeldamisel)
- `table(tunnus1, tunnus2)` – koostab kahemõõtmelise sagedustabeli
- `t(table)` – vahetab tabeli read ja veerud (transponeerib)
- `fable(tunnus1 + tunnus2 ~ tunnus3 + tunnus4, data = andmed)` – teeb mitmemõõtmelise sagedustabeli, mis on inimesele lihtsasti loetav

Kui on soov korrata mitme arvulise tunnuse keskmisi arvutada, sobib selleks käsk `colMeans(andmetabel)`, sarnane käsk on `rowMeans()`. Ridade või veergude summasid saab leida käskudega `rowSums()` ja `colSums()`. Seda saab näiteks kasutada sagedustabeli põhjal protsentide arvutamiseks:

```
sagedustabel <- table(andmed$SEX, andmed$LANX)
sagedustabel / rowSums(sagedustabel) #proovi, mis juhtub, kui kasutada /colSums()
```

```
##
##           No, speaks only English Yes, speaks another language
##   Female           0.8031949           0.1968051
##   Male             0.8147019           0.1852981
```

Sagedustabeli põhjal protsentide arvutamiseks (jaotustabeli arvutamiseks) on eelnevalt näidatud konstruktsioonist mugavam kasutada käsku `prop.table()`:

```
prop.table(sagedustabel) # ühisjaotus
prop.table(sagedustabel, margin = 1) # iga rida kokku 1 (ehk 100%)
prop.table(sagedustabel, margin = 2) # iga veerg kokku 1 (ehk 100%)
```

1.3.1 Ülesanded (Massachusettsi andmestik)

1. Mitu protsenti vastajatest on mehed?
2. Milline on palk, millest väiksemat palka saab 80% inimestest?
3. Kas lahutatute osakaal on suurem meeste või naiste hulgas?
4. Mitu üle 74 aasta vanust doktorikraadiga naist on andmestikus?
5. Mitmel inimesel on bakalaureuse-, magistri- või doktorikraad?
6. Milline on keskmine aastapalk meestel, milline naistel?
7. Kas Massachusettsi andmete alamandmestikus `valik5`(tekitatud eelmises ülesanneteplokis) on meeste ja naiste keskmised aastapalgad samasugused kui kogu andmestikus?

1.4 Faktortunnus

Faktortunnus pole tegelikult nn elementaartüüp (nagu näiteks `integer`), vaid keerulisem konstruktsioon. Nimelt faktortunnus on sildistatud koodide tunnus. Kui andmestik R-i sisse loetakse, siis vaikinisi pannakse tähelisi väärtuseid sisaldavate tunnuste tüübiks `factor` (seda võib ka `read.table()` argumentiga `stringsAsFactors` keelata) ning iga erinev väärtus kodeeritakse mingi täisarvuga, aga lisaks tehakse kodeerimistabel, kus on kirjas iga täisarvu (kodeeringu) tekstiline väärtus (ehk silt).

Et teada saada, mitu erinevat väärtust antud faktortunnusel võib üldse olla, kasutatakse käsku `levels()`. Sealjuures ei pruugi kõiki faktori väärtustasemeid antud andmetes üldse esineda:

```
levels(mehed$SEX)
```

```
## [1] "Female" "Male"
```

```
table(mehed$SEX)
```

```
##
## Female   Male
##      0    3125
```

Faktortunnuse tekitamiseks saab kasutada käsku `factor()`. Vaikinisi pannakse faktortunnuse tekitamisel faktori tasemed tähestiku järjekorda. Seda saab aga muuta käsu `factor()` argumenti `levels` kasutades:

```
table(andmed$MARHT) # Mitu korda abielus olnud?
andmed$MARHT <- factor(andmed$MARHT, levels = c("One time", "Two times", "Three or more times"))
table(andmed$MARHT)
```

Mõnikord on imporditud andmestikus arvulised tunnused faktorkujul sellepärast, et sisseloetavas failis oli viga ning ühes lahtris oli arvu asemel mingi tekst. Kui nüüd proovida `as.numeric(.)` käsuga see tunnus arvuliseks teisendada, tekib segadus: **Factor**-tüüpi tunnus on juba täisarvude tunnus (kuigi neil on sildid juures) ning seetõttu antakse tulemuseks need täisarvud ehk kodeeringud. Segadust ei teki, kui faktortunnus kõigepealt sõneks teisendada (kodeeringud kaotatakse, jäävad ainult sildid) ning alles seejärel arvudeks.

```
(x <- factor(c("1", "8", "ei vastanud", "12"))) # välimised sulud tingivad väljatrüki
```

```
## [1] 1          8          ei vastanud 12
## Levels: 1 12 8 ei vastanud
```

```
as.numeric(x)
```

```
## [1] 1 3 4 2
```

```
as.numeric(as.character(x))
```

```
## Warning: NAs introduced by coercion
```

```
## [1] 1 8 NA 12
```

Mõnikord soovime arvulist tunnust muuta nn ordinaaltunnuseks (st selliseks, kus on mõned üksikud kategooriad, mis on omavahel järjestatud). Näiteks palkade statistika esitamisel on soov teada infot palgavahemike kaupa. Arvulist tunnust aitab lõikudeks tükeldada käsk `cut(.)`. Selle käsu tulemusel tekib faktortunnus, mille silte saab `cut(.)` käsu argumendiga `labels` ette anda:

```
palgad <- cut(andmed$WAGP, breaks = c(0, 999, 4999, Inf), include.lowest = T,
             labels = c("0-999", "1000-4999", ">= 5000"))
```

```
table(palgad)
```

```
## palgad
```

```
##      0-999 1000-4999  >= 5000
```

```
##      1903      296      3114
```

1.4.1 Ülesanded

1. Tekita tunnus, kus oleks kirjas, millisesse vanusgruppi inimene kuulub: 0–17, 18–49, 50–64, 65 või vanem.
2. Kas USA kodakonsust mitteomavate naiste ja meeste hulgas on vanusgrupid erinevalt jaotunud?