R: Nano-opas

Pekka Pere
matematiikan ja systeemianalyysin laitos
Aalto-yliopisto

24.8.2023

1. R

- R on ohjelmointikieli, joka sopii erityisen hyvin tilastotieteellisten analyysien ja kuvioiden tekemiseen.
- Ross Ihaka ja Robert Gentleman loivat R:n kaupallisen S-kielen vastineeksi Aucklandin yliopistossa Uudessa-Seelannissa. Nimi R kunnioittaa S:ää ja viittaa ohjelman luojien etukirjaimiin. S-kielen tärkeimpiä kehittäjiä oli John Chambers. Edellä mainittujen lisäksi R:ää ovat kehittäneet ja kehittävät lukuisat tutkijat ympäri maailmaa. Kehitystyötä johdetaan R-säätiöstä (*R Foundation*), joka on rekisteröity Wieniin Itävaltaan.
- Versio 1.0 julkaistiin helmikuussa 2000.
- R on ilmainen! Sitä voi käyttää yliopisto-opintojen jälkeenkin.
- R on niin laajasti käytetty, että sitä voi kutsua tilastotieteen *lingua francaksi*.

Lisätietoja on The Comprehensive R Archive Network (CRAN) -sivulla `https://www.r-project.org/`. CRAN on valtaisa tietovarantoverkosto.

2. R:n asennus

Asennus Windows-käyttöjärjestelmään:

- Mene sivulle `https://ftp.acc.umu.se/mirror/CRAN/`.[1]
- Klikkaa linkkiketju *Download R for Windows* → *install R for the first time* → *Download R-“versionumero” for Windows*. R:n asennusohjelma imuroituu nyt tietokoneellesi (vie vähän aikaa).
- Avaa imuroitu tiedosto R-“versionumero”-win.exe (löytyy selaimesi oikean yläkulman valikoista tai tiedostonhallinnasta kohdasta Ladatut tiedostot (*Downloads*). Hyväksy klikkaamalla asennus ja seuraavat esiintulevat oletusarvoiset ehdotukset ja lopussa ruksaa kuvakkeen työpöydälle ja pikakäynnistysnäppäimen luomiset. R asentuu.
- Tietokoneesi aloitusnäkymässä on nyt R-logo.

[1] Myös Mac- ja Linux-käyttöjärjestelmille löytyvät linkit täältä.

3. R:n käyttö

Merkintöjä:

- komentokehote >
- sijoitusoperaatio <-
- R:n palautteen *n*. alkio [n]
- R:n jatkamiskehote (esimerkki alempana) +
- kommentti (esimerkki alempana) #
- R:n palaute (oppaan merkintätapa) ##
- R-palautteesta poistettua tekstiä (oppaan merkintätapa) - - .

R käynnistyy kaksoisklikkaamalla  -logoa. R:ää komennetaan kirjoittamalla tekstiä komentokehotemerkillä > alkavalle komentoriville ja painamalla syöttönäppäintä ↲.

Yksinkertaisimmillaan R on (symbolinen) laskin:

```
> 1+2
## [1] 3
> a <- 1
> b <- 2
> a
## [1] 1
> b
## [1] 2
> a+b
## [1] 3
```

Ensin laskettiin $1 + 2 = 3$. R:n palaute oli `[1] 3` . Siinä `[1]` osoittaa palautteen 1. alkion — tässä 3:n. Kaksi risuaitaa ## ovat oppaan käytäntö osoittaa R:n palaute. R ei tuota tällaista merkintää.

Seuraavaksi ohjattiin lukuarvot 1 ja 2 muuttujien a ja b arvoiksi sijoitusoperaatiolla `<-`. Komennoilla a ja b katsottiin, millaisia a ja b ovat. Käskyllä a+b laskettiin summa ja saatiin vastaukseksi 3.

Palaute voi koostua useammasta alkiosta, jolloin R osoittaa hakasuluissa, kuinka mones palautteen alkio on rivillä ensimmäisenä. Komento `seq(1,30,1)` (*sequence*) tuottaa luvut 1:stä 30:een yhden välein (- - osoittaa poistettua R-palautetta):

```
> seq(1,30,1)
## [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 - -
## [26] 26 27 28 29 30
```

Palaute koostuu luvuista $1, \ldots, 30$. Toisen rivin ensimmäinen alkio on palautteen 26. alkio. Siksi R on tulostanut sen edelle `[26]`. Yleensä alkioiden järjestysnumero ei ole yhtä helposti hahmotettavissa. Silloin R:n käytäntö on avuksi.

Huom1! Kokeile komentoja edellä! Tietokoneohjelmia oppii vain käyttämällä niitä. Huom2! Komentoja kutsutaan R:ssä funktioiksi. Ne eivät ole funktioita matemaattisessa mielessä. Selvyyden vuoksi tässä puhutaan komennoista. Huom3! Komentokehotetta > ei osoiteta komentojen edellä jatkossa.

Huom2! Komentokehotetta > ei osoiteta komentojen edellä jatkossa.

Havaintoja voi tuoda monella tavalla R:ään. Yhden muuttujan havainnot saa R:ään kätevästi ketjutuskomennolla `c(·)` (*concatenate*):

```
x <- c(49,35,32,39,45)
x
## [1] 49 35 32 39 45
y <- c(45,37,30,
## +
32,40)
y
## [1] 45 37 30 32 40
```

Havainnot luettiin ensin muuttujaan x. Sen jälkeen katsottiin, että x todella koostuu niistä. Toiset havainnot luettiin seuraavaksi kahdessa erässä muuttujaan y. Yllä + on jatkamiskehote, joka ilmoittaa, että R odottaa komennon loppuosaa. Seuraavalla rivillä komento täydennettiin loppuun.

Huom! Jatkamiskehote + ei suorita yhteenlaskua!

Suuret aineistot kannattaa tuoda R:ään esimerkiksi `read.table`-komennolla. Sitä ei havainnollisteta tässä.

R erottelee isot ja pienet kirjaimet:

```
X
## Error: object 'X' not found
x
## [1] 49 35 32 39 45
```

R ei tunnista X:ää, koska sitä ei määritelty edellä. Vain pieni x määriteltiin.

Huom1! Kokeile komentoja edellä! Tietokoneohjelmia oppii vain käyttämällä niitä.

Huom2! Komentoja kutsutaan R:ssä funktioiksi. Ne eivät ole funktioita matemaattisessa mielessä. Selvyyden vuoksi tässä puhutaan komennoista.

R:ssä on lukuisia komentoja, joilla analysoida aineistoja. Muuttujan x otoskeskiarvo, otoshajonta ja muita tunnuslukuja saadaan `mean(x)`-, `sd(x)`- ja `summary(x)`-komennoilla:

```
mean(x)
## [1] 40
sd(x)
## [1] 7
summary(x)
## Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   32      35      39      40      45      49
```

Otoskeskiarvo ja -hajonta ovat 40 ja 7. Aineiston pienin ja suurin havainto ovat 32 ja 49. Mediaani on 39, ja 1. ja 3. kvartiili ovat 35 ja 45.

Muuttujien x ja y otoskorrelaatio lasketaan `cor(x,y)`-komennolla:

```
cor(x,y)
## [1] 0.8725105
```

Tarkempia tietoja laskuista edellä saa komennolla `help`: `help(mean)`, `help(sd)`, `help(summary)` tai `help(cor)`.
R operoi sujuvasti lukuisilla jakaumilla kuten normaali-, χ^2-, t- ja F-jakaumilla. Komennot `pnorm(x)` ja `qnorm(x)` laskevat standardinormaalijakauman kertymäfunktion arvon pisteessä x ja vastaavan kvantiilin pisteessä x ($p \leftrightarrow$ *probability*; $q \leftrightarrow$ *quantile*). Vastaavat komennot normaalijakaumalle odotusarvolla `m` ja keskihajonnalla `s` ovat `pnorm(x,m,s)` ja `qnorm(x,m,s)`.

Esimerkkejä:

```
pnorm(-1.964)
## [1] 0.02476505
qnorm(0.02476505)
## [1] -1.964
pnorm(-1.964,-2,2)
## [1] 0.5071806
qnorm(0.5071806,-2,2)
## [1] -1.964
```

Huom! Normaalijakauma määritellään oppikirjoissa järjestään odotusarvon (μ) ja varianssin (σ^2) — ei keskihajonnan (σ) kuten edellä — avulla: $N(\mu, \sigma^2)$.

Vastaavat komennot χ^2-, t- ja F-jakaumille ovat `pchisq(x,df)`, `pt(x,df)` ja `pf(x,df1, df2)` sekä `qchisq(x,df)`, `qt(x,df)` ja `qf(x,df1, df2)`. Niissä `df`, `df1` ja `df2` viittaavat jakauman vapausasteisiin (*degrees of freedom*).

Samalla rivillä voidaan antaa useampia komentoja ;-merkillä erotettuna:

```
pt(-1.964,20); qt(0.03179091,20)
## [1] 0.03179091
## [1] -1.964
```

4. Hyviä käytäntöjä

Usein komennot kannattaa ajaa tiedostosta käsin ja lopuksi tallettaa tiedosto: Valitse R:ssä *File* (näytön vasen yläkulma) ja *New script*. R:n tekstieditori avautuu. Kirjoita tarvitsemasi komennot peräkkäisille omille riveille, ja aja ne painamalla näppäinyhdistelmää Ctrl-R kunkin rivin kohdalla. Talleta tiedosto valitsemalla *File* ja *Save as*.

Tiedoston voi hakea uudelleen käytettäväksi valinnoilla *File* ja *Open script*. Näin voi helposti toistaa aiemmat analyysinsa tai muokata niitä samantapaisiin uusiin tehtäviin. Jos antaa tiedoston jollekulle, hän voi toistaa ja tarkistaa analyysit.

R sivuuttaa risuaidalla # merkityt kommenttirivit. Ne helpottavat hahmottamista, mitä on tehty:

```
x <- c(49,35,32,39,45)
# y on laskettu Albert Neron artikkelissa ‘‘Elämän salaisuus’’:
y <- c(45,37,30,32,40)
```

Kommentoi koodiasi!

5. Korjaaminen ja peruminen

Virheellisen komennon voi korjata painamalla nuoli ylös -näppäintä ↑:

```
x <- c(49 35 32 39 45)
## Error: unexpected numeric constant in "x <- c(49 35"
x <- c(49,35,32,39,45)
x
## [1] 49 35 32 39 45
```

Nuolinäppäin tuo virheellisen komennon `x <- c(49 35 32 39 45)` komentoriville, josta sen voi korjata helposti lisäämällä pilkut lukujen väliin. (Tehty 3. rivillä yllä.)

R:n tekstieditorissa korjaaminen on erityisen helppoa, koska komento jää esille sen ajamisen jälkeen.

Jos R pyytää täydentämään komentoa, muttet halua tai osaa, peru toiminto poistumisnäppäimellä Esc:

```
y <- c(45,37,33,
## +
```

Jos nyt huomaat antaneesi kolmannen lukuarvon väärin (33 eikä 30), keskeytä komento painamalla Esc.

6. Ohjelmoinnista ja paketeista

R:n yksinkertainen käyttö ei vaadi ohjelmointitaitoja. Hyvin paljon saa tehtyä hyvin simppeleillä komennoilla tai lyhyillä komentoketjuilla.

Monia vaativampia tehtäviä varten on lisäosia eli paketteja (*package*). Niitä on jo yli 18 500 (`https://cran.r-project.org/web/packages/`). Esimerkiksi psykologeille erityisen sopivia tilastollisia menetelmiä on koottu psych-pakettiin. Se haetaan Internetistä omalle koneelle ja otetaan käyttöön komennoilla `install.packages("psych")` ja `library(psych)`. Ensimmäistä pakettia asennettaessa R kysyy, miltä palvelimelta paketti haetaan. Sopiva valinta on esimerkiksi Ruotsissa (*Sweden*) sijaitseva palvelin, jolle CRANin tietovaranto on kopioitu.

Monimutkaisiin tehtäviin löytyy usein myös valmiita komentojonoja R-oppaista tai Internetistä. Koodit voi kopioida, ja säätää omiin tarpeisiin. Sitä tehdään paljon. Koodin tekijään tulee viitata, kun näin tekee. Esimerkiksi erikoisen kuvion piirtäminen voi olla tällainen tehtävä. Tällöinkin vaihtoehto saattaa olla ladata sopiva paketti (esim. ggplot2).

7. Lopettaminen ja viittaaminen

Lopeta R-istunto `quit()`- tai lyhyemmin `q()`-komennolla:

```
q()
```

Vastaa R:n esittämiin seuraaviin kysymyksiin kieltävästi. Istuntosi on sen jälkeen päättynyt.

Jos käytät opinnäytteessäsi tai muussa teoksessasi R:ää, kerro se. Ohjeen viittaamiseen saat `citation()`-komennolla:

```
citation()
## To cite R in publications use:
##
##   R Core Team (2023). _R: A language and environment for statistical
##   computing_. R Foundation for Statistical Computing, Vienna, Austria.
##   <https://www.R-project.org/>.
 - -
## We have invested a lot of time and effort in creating R, please cite it
## when using it for data analysis. See also ’citation("pkgname")’ for
## citing R packages.
```

8. Oppaita ja neuvoja

R-opintoja voi syventää seuraavien opusten avulla:

- W. J. Braun ja D. J. Murdoch (2021): A First Course in Statistical Programming with R, 3. laitos. CUP.
- M. J. Crawley (2013): The R Book, 2. laitos. Wiley.
- R. I. Kabacoff (2022): R in Action, 3. laitos. Manning.
- P. Väkeväinen (2018): Ohjeita tilastollisen tutkimuksen toteuttamiseksi R-ohjelmiston avulla. Informaatiotieteiden yksikön raportteja 64. Tampereen yliopisto.

Huom1! Oppaita ei ole tarkoitettu luettaviksi kannesta kanteen! Oppaasta kannattaa tyypillisesti lukea vain alkuluvut, ja sen jälkeen konsultoida sitä tarpeen tullen.

Huom2! Oppaita kannattaa opiskella ei vain lukien vaan samalla itse koodeja kokeillen.

Internetistä löytyy usein nopeasti apu mitä erilaisimpiin kysymyksiin R:n käytöstä.