Sissejuhatus

Allan Sims

31.01.2025

Sissejuhatus

Käesolev õpik on mõeldud metsandusüliõpilastele, kes soovivad omandada matemaatilise statistika põhitõdesid praktilises kontekstis. Materjali koostamisel on lähtutud eeldusest, et lugejatel ei pruugi olla sügavamat matemaatilist ettevalmistust. Terminoloogia on kooskõlas "Andmeanalüüsi ja statistika oskussõnastikuga" (https://sonaveeb.ee/ds/aso), kuid mõisted on lahtiseletatud metsandusvaldkonnast lähtuvalt.

Metsanduses toimub pidev andmete kogumine ja nende töötlemine. Kõiki soovitud andmeid ei ole võimalik alati mõõta ning seetõttu tuletatakse need erinevate mudelite abil. Sellised mudelid on saadud statistilise andmetöötluse ja modelleerimise abil. Seetõttu on andmete töötlemisel oluline tunda ka statistika põhitõdesid, et mõista paremini andmetest järelduste ja otsuste tegemisel. Õpiku näited keskenduvad metsanduslikele andmetele. Tarkvarana on kasutatud statistikatarkvara **R** ja **MS Excel**, kuid nende valikul on oluline märkida nende erinevaid võimekusi:

1. **R** on spetsialiseerunud statistiliseks modelleerimiseks ja analüüsiks, pakkudes laia valikut meetodeid (nt regressioon, testid). Kuid andmehalduse lihtsustamiseks võib olla vaja täiendavaid tööriistu.

2. **MS Excel** on kasutajasõbralik platvorm andmete korrastamiseks ja esmaseks visualiseerimiseks, kuid selle statistilised funktsioonid on piiratud (nt puuduvad keerukama modelleerimise võimalused).

Need tarkvarad täiendavad üksteist. MS Excel sobib hästi andmete esmaseks töötluseks, R aga detailsemaks analüüsiks. Näiteks võib MS Excelis ette valmistatud andmeid edasi analüüsida R-is. Samuti on võimalik integreerida R-skripte Exceli makrotesse või kasutada Pythoni skripte R-i keskkonnas, et laiendada funktsionaalsust, aga neid integratsioone käesolev õpik ei käsitle. Sellised moodulipõhised lahendused muudavad tarkvaravahendid paindlikumaks, kuid nõuavad aga pidevat oskuste arendamist.

Oluline on mõista, et üksik tarkvara ei suuda katta kõiki analüüsivajadusi. Seetõttu on õpikus rõhk praktilistel lahendustel, mis õpetavad ülesannete jaoks optimaalsete tööriistade valikut. Statistika valdkond areneb kiiresti – nii meetodid, tarkvarad kui rakendusvaldkonnad muutuvad. Käesolevat õpikut uuendatakse regulaarselt, et kajastada ajakohaseid standardeid ja tarkvarauuendusi.

Statistikaprogramm R

R on tarkvarakeskkond, mis on loodud spetsiaalselt statistiliseks arvutamiseks ja andmete graafiliseks esitamiseks. See on avatud lähtekoodiga projekt, mida toetab suur kogukond ning mida eelistavad andmeteadlased ja statistikud üle kogu maailma oma andmeanalüüsi vajadusteks. R pakub laia valikut tööriistu andmete importimiseks, töötlemiseks, modelleerimiseks, analüüsiks ja visualiseerimiseks. Järgnevalt on ülevaade R-i peamistest kasutusvõimalustest statistilise analüüsi tegemiseks.

Andmete importimine ja töötlemine

- **Andmete importimine.** R toetab andmete importimist paljudest allikatest, sealhulgas teksti-, CSV- ja MS Exceli failidest ning andmebaasidest ja isegi veebist.
- **Andmete puhastamine.** Andmete analüüsiks ettevalmistamisel pakub R võimalusi nende sortimiseks, filtreerimiseks, puuduvate väärtuste käsitlemiseks ja duplikaatide eemaldamiseks.

Andmete analüüs

- **Statistilised funktsioonid.** R sisaldab ulatuslikku valikut sisseehitatud funktsioone statistilisi tunnuste arvutamiseks, nagu keskmine, mediaan, standardhälve, kvantiilid ja palju muud, mis võimaldavad kiiresti arvutada olulisi statistilisi näitajaid.
- **Andmete manipuleerimine.** R pakub võimekaid pakette andmete manipuleerimiseks, mis võimaldavad efektiivselt andmete filtreerimist, valimist, rühmitamist ja kokkuvõtete tegemist.

Statistilised mudelid ja analüüsimeetodid

- **Hüpoteeside testimine.** R võimaldab kasutajatel teostada erinevaid statistilisi teste, nagu t-testid, hii-ruut-testid, ANOVA ja palju muud, et uurida andmete omavahelisi seoseid ja testida hüpoteese.
- **Regressioon ja korrelatsioon.** R toetab mitmesuguseid regressioonianalüüse võimaldades seeläbi uurida muutujate vahelisi seoseid.

Andmete visualiseerimine

- **Graafikud ja diagrammid.** R pakub mitmeid pakette andmete visualiseerimiseks võimaldades luua keerukaid ja esteetiliselt meeldivaid graafikuid.
- **Interaktiivsed visualisatsioonid.** Lisaks staatilistele graafikutele võimaldab R luua interaktiivseid visualisatsioone.

Laiendatavus ja kogukond

- **Pakettide ökosüsteem.** R toetab laialdast pakettide ökosüsteemi, mis katab spetsiifilisi statistilisi meetodeid, andmete visualiseerimise tehnikaid, masinõpet ja palju muud, võimaldades kasutajatel laiendada R-i funktsionaalsust vastavalt vajadusele.
- **Tugev kogukond.** R-i kasutajate ja arendajate kogukond on väga aktiivne, pakkudes tuge, juhendeid ja ressursse, mis aitavad lahendada keerulisi analüütilisi väljakutseid.
- **Õpikud ja juhendid.** R-i kasutamiseks on koostatud erinevaid õpikud, mis on mõeldud nii algajatele kui ka edasijõudnutele. Veebilehel https://bookdown.org/ on mitmed sellised õpikud.

R-i paketid

Statistikaprogramm põhineb pakettidel. Iga pakett on meetodite ja funktsioonide kogumik ning need võimaldavad teha kindlaid protseduure. Kasutajal on võimalik installeerida teiste poolt koostatud pakette või neid vajadusel ise koostada. R-i paketid on tööriistade, funktsioonide ja andmete kogumikud, mis laiendavad R-i põhifunktsionaalsust ning võimaldavad lahendada erinevaid andmetöötluse, statistilise analüüsi ja visualiseerimise ülesandeid. Enamik pakette on loodud R-i kogukonna liikmete poolt. CRAN (Comprehensive R Archive Network) on ametlik hoidla, kuhu arendajad saavad pakette üles laadida ning kasutajad neid alla laadida.

Pakettide installeerimiseks kasutatakse funktsiooni `install.packages()`. Näiteks:

```
install.packages("ggplot2")
```

Kui pakett on installitud, tuleb see sessioonis laadida, kasutades funktsiooni `library()`:

```
library(ggplot2)
# või
library("ggplot2")
```

Pakettide puhul tuleb jälgida, et installimise funktsioonis peab paketi nimetus olema jutumärkide vahel, aga paketi sisse laadimisel võib see olla ka jutumärkideta.

Lisaks on ka funktsioon `require()`, mis võimaldab testida paketi olemasolu. Selline testimise vajadus tekib, kui skripti kasutatakse korduvalt erinevates arvutites ning ühes arvutis skripti loomisel võib olla vajalikud paketid juba installitud ning teises arvutis ei pruugi veel vajalikke pakette olla. Näiteks, kontrollime, kas antud pakett on olemas ning kui see puudub, siis installime:

```
if(!require(ggplot2)){
    install.packages("ggplot2")
}
```

Funktsioon `require()` loeb ka paketi sisse, kui see on installitud, aga seda on soovituslik kasutada pigem selleks, et kontrollida paketi olemasolu. Kui koostatud on arvutusskript, milles antud funktsiooniga laaditakse pakett, mis puudub, siis skripti töö jätkub, aga see ebaõnnestub, sest vajalikku paketti ei ole.

Igal paketil on põhjalik dokumentatsioon, millele pääseb ligi, kasutades `help()` või `?` süntaksit. Paljud paketid pakuvad ka sissejuhatavaid õpetusi ja veebipõhiseid juhendeid.

R-i keskkonnas arvutamine

R-is saab teha mitmesuguseid arvutusi, alates lihtsatest aritmeetilistest operatsioonidest kuni keerukate statistiliste analüüsideni.

Aritmeetilised operatsioonid. R toetab põhilisi aritmeetilisi operatsioone nagu liitmine (+), lahutamine (-), korrutamine (*), jagamine (/) ja astendamine (^).

```
1 + 1
1 - 1
1 * 1
1 / 1
1 ^ 1
```

Muutujad: R-is saab väärtusi salvestada muutujatesse, kasutades omistamisoperaatorit (`<-` või `=`).

```
x = 1 + 1
x <- 1 - 1
```

Andmetüübid. R toetab erinevaid andmetüüpe, sealhulgas numbrid (nii täis- kui ka reaalarvud), tekst (stringid), loogilised väärtused (TRUE/FALSE) ja faktorid (kategoorilised andmed).

```
x <- 1
y <- TRUE
z <- "MA"
```

Vektorid ja maatriksid. R on eriti tugev vektor- ja maatriksarvutustes.

```
#Vektoreid saab luua funktsiooniga `c()`,
x <- c(1, 2, 3, 4, 5)
#Maatriksid saab luua funktsioonidega
x <- matrix(1:9, nrow = 3, ncol=3)
dim(x)
```

```
## [1] 3 3
```

```
x <- array(1:16, dim = c(2,2,2))
dim(x)
```

```
## [1] 2 2 2
```

Arvutused tsüklitena. R võimaldab teha korduvate tsüklitena.

```
for(x in 1:10){
  y <- x + 1
}
```

MS Excel

MS Excel on laialdaselt kasutatav tabelarvutusprogramm, mis pakub mitmekülgseid tööriistu andmete analüüsimiseks, visualiseerimiseks ja esitlemiseks. MS Exceli kasutusvõimalused statistilise analüüsi tegemiseks hõlmavad mitmesuguseid funktsioone ja lisavahendeid, mis võimaldavad kasutajatel teostada põhjalikke analüüse erinevate andmekogumite põhjal. Järgnevalt on toodud ülevaade peamistest MS Exceli kasutusvõimalustest statistiliseks analüüsiks.

Andmete importimine ja korraldamine

- **Andmete importimine.** Excel võimaldab importida andmeid erinevatest allikatest, sealhulgas tekstifailidest, CSV-failidest, andmebaasidest ja veebiteenustest.
- **Andmete puhastamine ja ettevalmistamine.** Kasutajad saavad kasutada tööriistu, nagu sortimine, filtreerimine ja eemaldamine, et korrastada andmeid analüüsi jaoks. Samuti on võimalik kasutada tingimusvormingut, et esile tõsta olulisi andmepunkte või mustreid.

Andmete analüüsimine

- **Põhilised statistilised funktsioonid.** Excel sisaldab mitmesuguseid statistilisi funktsioone, nagu keskmine (AVERAGE), mediaan (MEDIAN), standardhälve (STDEV), minimaalne (MIN) ja maksimaalne (MAX) väärtus.
- **Liigendtabelid.** Liigendtabelid võimaldavad kiiresti kokku võtta, analüüsida, uurida ja esitada andmeid. Kasutajad saavad hõlpsasti grupeerida ja segmenteerida andmeid, et tuvastada mustreid või suundumusi.

Statistilised testid ja analüüsimeetodid

- **Hüpoteeside testimine.** Excelis saab teostada t-Testi ja ANOVA (Analysis of Variance) teste, et võrrelda andmekogumeid või analüüsida erinevusi rühmade vahel.
- **Regressioonianalüüs.** Kasutajad saavad kasutada regressioonianalüüsi, et uurida muutujate vahelisi seoseid ja ennustada tulemusi.

Andmete visualiseerimine - **Diagrammid ja graafikud.** Excel pakub erinevaid viise andmete visualiseerimiseks, sealhulgas tulp-, joon-, piirkonna-, ring-, hajuvus- ja mullidiagramme. - **Tingimuslik vormindamine:** See võimaldab andmeid visuaalselt esile tõsta, kasutades värve, ikoone ja andmeribasid, et lihtsustada andmestiku mustrite ja trendide tuvastamist.

Makrod ja VBA skriptimine - **Automatiseerimine ja kohandamine.** Kasutajad saavad luua makrosid ja kasutada Visual Basic for Applications (VBA) skriptimist, et automatiseerida korduvaid ülesandeid ja kohandada analüütilisi protseduure, mis ületavad Exceli standardfunktsioonide piire.

Lisandmoodulid ja tööriistad - **Analysis ToolPak.** See on MS Exceli lisandmoodul, mis pakub täiendavaid statistilisi funktsioone, sealhulgas keerulisemaid analüütilisi tööriistu ja statistilisi meetodeid.

MS Exceli paindlikkus ja lai funktsionaalsus muudavad selle heaks tööriistaks statistiliseks analüüsiks. Siiski on oluline märkida, et keerukamate statistiliste analüüside ja suurte andmekogumite korral pakuvad spetsialiseeritud statistikatarkvarad rohkem võimalusi.