Metsanduslik statistika R ja MS Excel keskkonnas

Allan Sims, Andres Kiviste, Diana Laarmann

31.01.2025

Sisukord

Peatükk 1

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.

1.1 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, hiiruut-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.

1.1.1 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.

1.2 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.

Peatükk 2

Andmestik

Metsanduses kogutakse pidevalt mitmesuguseid andmeid, mis võivad erineda nii oma tüübilt kui ka struktuurilt. Andmekogumise protsessi üldnimetus on **takseerimine**, mis hõlmab mitmesuguseid meetodeid nagu mõõtmine, loendamine, määramine ja muud sarnased tegevused. Iga takseerimisviis määrab ühtlasi ka kogutavate andmete tüübi.

- **Mõõtmisel** saadakse enamasti **pidev tunnus**. See on tunnus, mille väärtused võivad olla mistahes arvud teatud vahemikus. See tähendab, et pideva tunnuse puhul ei ole väärtused piiratud kindlate, eraldiseisvate arvudega, vaid võivad muutuda sujuvalt ja pidevalt. Näiteks mõõdetud puu kõrgus või diameeter.
- **Loendamisel** saadakse **diskreetne tunnus**. See on tunnus, mille väärtused saavad olla ainult teatud kindlad, eraldiseisvad arvud, tavaliselt täisarvud. See tähendab, et diskreetsete tunnuste puhul ei saa väärtused olla sujuvad või pidevad, vaid peavad olema selgelt eristuvad. Näiteks puude ja aastarõngaste loendamine. Diskreetne on ka rinnaspindala mõõtmisel relaskoobiga loendamine, kus loendatakse puid, kuid teatud juhtudel kasutatakse ka konstanti 0,5.
- **Määramisel** saadakse üldjuhul **nominaalne tunnus**. See on selline tunnus, mis väljendab kvalitatiivseid ehk mittearvulisi andmeid, mida ei saa järjestada. Näiteks puuliik, kasvukohatüüp vms. Selle alla liigitub ka **ordinaalne tunnus** ehk järjestustunnus on tunnus, mille väärtustel on loogiline järjestus, kuid väärtuste vahed ei ole tingimata võrdsed. Näiteks on metsanduses selleks boniteediklass.

Statistikas tähendab **andmestik** vaatlusandmete kogumit, mis on kogutud mingi uuringu või eksperimendi käigus. See on toorandmete hulk, mida analüüsitakse, et avastada mustreid, seoseid ja teha järeldusi. Andmete tüüp mängib olulist rolli statistilises analüüsis, kuna see määrab, milliseid meetodeid ja visualiseerimisviise saab kasutada. Õige andmetüübi tuvastamine on seega esimene samm edukaks andmetöötluseks ja tulemuste tõlgendamiseks.

Andmestik on organiseeritud andmete kogum, mis on tavaliselt salvestatud tabelitesse, andmebaasidesse või muudesse struktuuridesse.

- **Tüübid:**

- **Vektor:** Sama tüüpi väärtuste jada (nt arvude vektor, tekstiväärtuste vektor).
- **Maatriks:** Sama tüüpi väärtuste kogum, kuid mis on mitmemõõtmeline.
- **Andmetabel:** Struktureeritud andmete kogum, kus iga rida esindab ühte objekti ja iga veerg esindab objekti atribuuti.

2.1 Andmestiku struktuur R keskkonnas

2.1.1 Vektor

Vektor on andmetüüp, mis võimaldab talletada mitut ühesugust andmetüüpi elementi ühes andmestruktuuris. Vektoris saab hoida näiteks numbreid, teksti või loogilisi väärtusi. Vektorit loetakse ühemõõtmeliseks andmestruktuuriks, kuna kõik selles sisalduvad elemendid asuvad ühes reas.

Vektorit saab luua kasutades funktsiooni `c()`, kuhu tuleb sisestada eraldatud komadega elemendid. Näiteks:

```
# Näide R keskkonnas

num_vektor <- c(1, 2, 3, 4, 5)
```

Vektori elementidele saab viidata indeksite abil, alustades indeksist 1. Näiteks:

```
# Näide R keskkonnas

# väljastab vektori num_vektor kolmanda elemendi väärtuse
num_vektor[3]
```

```
[1] 3
```

Vektori pikkus ehk elementide arv on võimalik saada funktsiooniga `length()`. Vektorit saab ühendada teise vektoriga funktsiooniga `c()`. Näiteks:

```
# Näide R keskkonnas

num_vektor_2 <- c(6, 7, 8)
kombineeritud_vektor <- c(num_vektor, num_vektor_2)
```

Võimalik on ka vektorit luua jada genereerimise abil, kasutades funktsiooni `seq()`. Näiteks:

```
# Näide R keskkonnas

# loob vektori, mis sisaldab arve vahemikus 1 kuni 10 sammuga 2
jarg_vector <- seq(1, 10, by = 2)
```

Vektoritega on võimalik teostada erinevaid arvutusi, teisendusi ning andmete filtreerimist. Vektorid on olulised R programmeerimises, kuna nende abil on mugav töödelda andmeid ja teostada statistilisi analüüse.

2.1.2 Maatriks

Andmetüüp maatriks on kahe- või mitmemõõtmeline andmete struktuur, kus andmed on paigutatud mitme rea ja veeruna, moodustades seeläbi mitmemõõtmelise andmestiku. Maatriksis olevad andmed peavad olema kõik samatüübilised, näiteks kõik arvud või kõik tekstid.

Maatriksi loomiseks kasutatakse funktsioone `matrix()` või `array()`, millest esimene võimaldab luua kahemõõtmelist ning teine juba mitmemõõtmelisi maatrikseid. Nendele funktsioonidele saab määrata erinevaid parameetreid, näiteks andmete sisu, ridade ja veergude arvu ning rea- ja veerunimed.

```
# Näide R keskkonnas
# funktsioon runif() loob juhuslike andmete vektori
# Loome kahemõõtmelise maatriksi juhuslikest arvudest
minu_maatriks <- matrix(runif(12), nrow=3, ncol=4)
print(minu_maatriks)
```

```
      [,1]  [,2]  [,3]  [,4]
[1,] 0,185 0,325 0,701 0,122
[2,] 0,141 0,476 0,770 0,459
[3,] 0,711 0,948 0,182 0,854
```

```
# Loome kolmemõõtmelise maatriksi juhuslikest arvudest
minu_array <- array(runif(12), dim = c(2,2,3))
```

Maatriksit saab kasutada erinevate arvutusoperatsioonide tegemiseks, nagu maatriksite korrutamine, liitmine, lahutamine jne. Samuti on võimalik maatriksit indekseerida ja välja printida konkreetseid ridu ja veerge.

2.1.3 Andmetabel

R-i andmetüüp andmetabel (data.frame) on üks olulisemaid andmestruktuure andmete haldamiseks ja analüüsimiseks R-keskkonnas. Tabelis on andmed korraldatud ridade (vaatluste) ja veergude (tunnuste) kaupa. Iga veerg andmetabelis vastab ühele muutujale, mis võib olla numbriline, tekstiline või faktor. Iga rida esindab ühte andmepunkti, mis sisaldab väärtusi vastavatele muutujatele.

Andmetabeli loomiseks saab kasutada funktsiooni `data.frame()`. Andmed saab andmetabelisse importida erinevatest andmeallikatest (nt CSV-failid, andmebaasid) või genereerida otse R-s.

```
# Näide R keskkonnas

# Loome data.frame andmestiku
minu_data_frame <- data.frame(x = runif(5), y = runif(5), z = 1)
```

Andmetabeli puhul on võimalik teha erinevaid andmehalduse ja analüüsi operatsioone nagu veergude ja ridade lisamine, kustutamine, filtreerimine, rühmitamine, sorteerimine ja statistiliste analüüside tegemine. Andmetabelil on oluline oskuslikult hallata ja kasutada, kuna see on väga levinud andmestruktuur R-keskkonnas ning võimaldab mugavalt töötada suurte andmekogumitega.

2.1.4 List

R-i andmetüüp list on objektide kogu, mis võimaldab ühte muutujasse salvestada erineva pikkuse ja tüübiga andmeid. Üks list võib koosneda erinevatest vektoritest, maatriksitest või tabelitest.

```
# Näide R keskkonnas

# Loome list andmestiku
minu_list <- list(
  x = runif(5),
```

```
  y = array(runif(16), dim = c(2,2,2)),
  z = c("MA","KU"))
```

2.2 Andmestike importimine R keskkonda

Andmete importimine CSV või Exceli failist R-i keskkonda on väga lihtne. Siin on näited mõlemast.

Andmete importimine CSV failist

Enne koodi käivitamist veendu, et CSV fail on salvestatud samas kaustas, kus Sa töötad R-is või et oled määranud õige tee failini.

```
# Näide R keskkonnas

andmed <- read.csv("failinimi.csv")
```

Kui CSV failil on teistsugune eraldusmärk (nt semikoolon), siis saab selle täpsustada lisades argumenti `sep=";"`.

```
# Näide R keskkonnas

andmed <- read.csv("failinimi.csv", sep=";")
```

Andmete importimine MS Exceli failist

Esmalt tuleb installida ja laadida sisse `openxlsx` pakett, mis võimaldab Exceli failidest andmete lugemist.

```
# Laadime peatükis kasutatavad paketid
library(dplyr)
library(readxl)
```

Seejärel kasutada funktsiooni `read_excel()`, et importida andmed Exceli failist.

```
# Näide R keskkonnas

andmed <- read_excel("failinimi.xlsx")
```

Arvutamine R-i keskkonnas

```
# Liitmine
1 + 1
```

```
[1] 2
```

```
# Lahutamine
5 - 3
```

```
[1] 2
```

```
# Korrutamine
2 * 4
```

```
[1] 8
```

```r
# Jagamine
10 / 2
```

```
[1] 5
```

```r
# Astendamine
2 ^ 3
```

```
[1] 8
```

```r
# Jäägi leidmine
10 %% 3
```

```
[1] 1
```

R keskkonnas on võimalik teha arvutusi tsüklina

```r
# Loome vektori 1st 10ni

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

Peatükk 3

Andmestike rühmitamine

Pideva tunnuse rühmitamine tähendab andmete jaotamist gruppidesse või kategooriatesse vastavalt nende väärtustele. See on oluline statistilise analüüsi meetod, kui soovime analüüsida suurt hulka andmeid ning leida seoseid või mustreid nende vahel.

Pideva tunnuse rühmitamine võib olla vajalik mitmel põhjusel:

- Andmete lihtsustamine. Pidevate tunnuste rühmitamine võib muuta andmed hõlpsamini tõlgendatavaks ja analüüsitavaks, eriti juhul, kui on suur hulk erinevaid väärtusi.
- Seoste või trendide avastamine. Rühmitades pidevaid tunnuseid, võime avastada seoseid või mustreid erinevate gruppide vahel. See võib aidata mõista näiteks, kuidas üks muutuja mõjutab teist või millised tegurid on omavahel seotud.
- Statistilise analüüsi täpsus. Pidevate tunnuste rühmitamine võib parandada statistiliste analüüside täpsust ja usaldusväärsust, kuna see võib aidata vähendada andmete varieeruvust ning korrigeerida võimalikke moonutusi.

Seega on pideva tunnuse rühmitamine oluline statistilise analüüsi tööriist, mis aitab andmeid paremini mõista ja interpreteerida ning seeläbi teha järeldusi ja otsuseid põhinevalt faktidel ja statistilistel seostel.

3.1 Jaotused

3.1.1 Empiiriline jaotus

Empiiriline jaotus on statistilise andmestiku jaotus, mis põhineb tegelikel vaatlustel või mõõtmistel. See erineb teoreetilisest jaotusest, mis on abstraktne ja ideaalne jaotus, mida kasutatakse statistilistes mudelites ja analüüsides.

Empiiriline jaotus saadakse andmete kogumisel ja nende analüüsimisel, et mõista nähtuse tegelikku jaotust ja omadusi. See võib olla esitatud graafikuna (nt histogrammina) või matemaatilise mudelina, mis kirjeldab

andmestiku jaotust.

Empiirilise jaotuse kasutamine on oluline, et saada parem arusaam uuritavast populatsioonist või nähtusest ning teha usaldusväärseid järeldusi statistiliste analüüside põhjal.

3.1.2 Teoreetiline jaotus

Teoreetiline jaotus on statistikas abstraktne kontseptsioon, mis kirjeldab tõenäosust, et mingi nähtuse väärtus võtab teatud vahemiku või konkreetse väärtuse. Teoreetiline jaotus arvutatakse tavaliselt matemaatiliste mudelite abil ning see aitab meil mõista andmete tõenäosuslikku käitumist mingi statistilise populatsiooni või nähtuse puhul. Teoreetilised jaotused on olulised statistiliste analüüside läbiviimisel ning nende põhjal saab teha järeldusi ja prognoose erinevate metsandustega seotud uuringute kohta.

3.2 Histogramm

Histogramm ehk sagedusjaotuse tulpdiagramm on graafiline esitusviis andmehulga sageduste jaotumise näitamiseks. Histogramm koosneb üksteise kõrvale paigutatud tulpadest, kus iga tulp esindab teatud vahemikku või klassi, ning selle kõrgus näitab antud klassi sagedust. Horisontaaltelg esitab klasside vahemikke või kategooriaid ning vertikaaltelg näitab vastavate klasside sagedusi või sageduste suhtarvu. Histogrammi abil saab hinnata andmehulga jaotust ning tuvastada erinevaid tendentse ja mustreid andmetes. Histogrammi loomise eesmärgiks on anda visuaalne ülevaade andmete koondumisest ja levikust ning seeläbi paremini mõista andmete struktuuri. Histogrammi loomisel on oluline valida sobiv klassilaius, mis võimaldab õigesti tõlgendada ja analüüsida andmeid. Histogramm on oluline tööriist statistilise info esitamiseks ja mõistmiseks nii akadeemilistes kui ka praktilistes seadetes, sealhulgas metsanduses.

3.2.1 R keskkonnas:

```
# Näide R keskkonnas
# Laadime peatükis kasutatavad paketid
library(dplyr)
library(ggplot2)
```

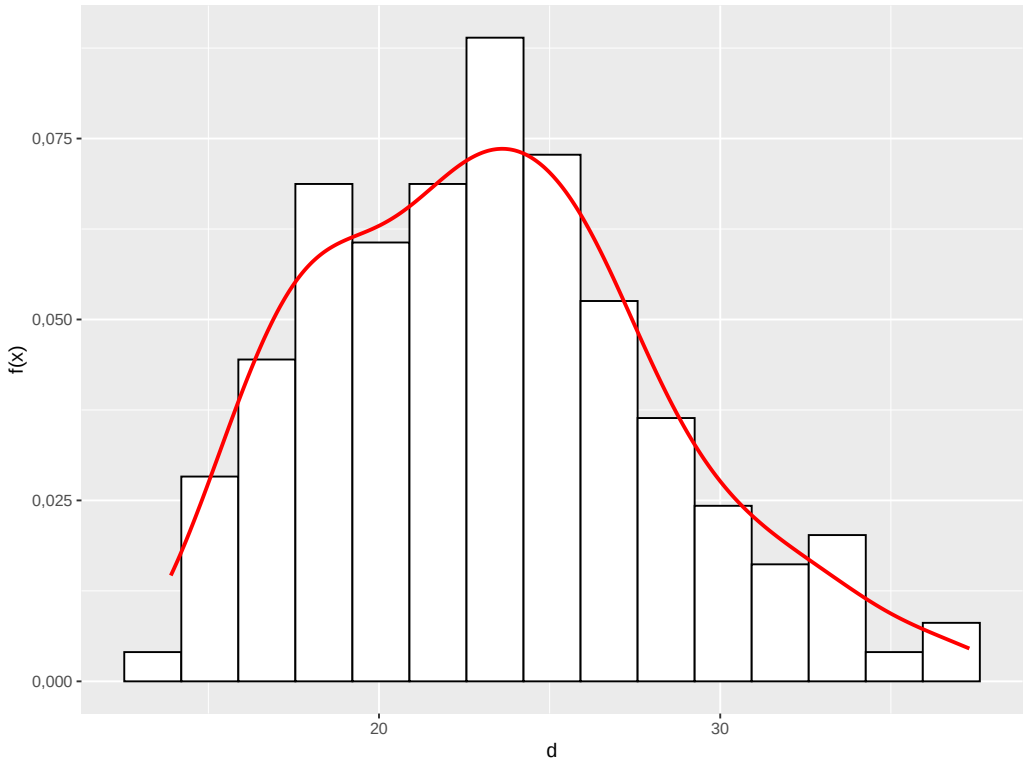
Esmalt loome pidevast tunnusest mõned juhuslikud andmed ning seejärel kuvame nende andmete põhjal histogrammi.

Histogramm

```
# Laadime proovitükkide andmed
df <- readxl::read_excel("data/naited.xlsx","prt_andmed")

# koostame diameetrite histogrammi
ggplot(df, aes(d)) +
  geom_histogram(
    aes(y = after_stat(density)),
    bins = 15,
    fill = "white",
    color = "black"
```

```
) +
geom_density(color = "red", lwd = 1) +
labs(y = "f(x)")
```



Joonis 3.1. Histogramm

See kood loob R-keeles `ggplot2` paketiga graafiku, mis visualiseerib andmete jaotust. Vaatame seda samm-sammult:

1. **`ggplot(df, aes(x))`**: See alustab graafiku loomist. `df` on andmetabel, mis sisaldab andmeid. `aes(x)` määrab, et x-teljele kuvatakse muutujat `x`. See on graafiku "põhi", millele järgnevad kihid lisatakse.

2. **`geom_histogram(...)`**: See lisab histogrammi.

   - `geom_histogram()` funktsioon loob histogrammi, mis näitab andmete sagedust erinevatesse gruppidesse jaotatuna.
   - `aes(y=after_stat(density))` on oluline osa. See määrab, et y-teljel kuvatakse *tihedust* (density), mitte lihtsalt sagedust (count). `after_stat()` funktsiooniga pääsetakse ligi statistilisele väärtusele, mis arvutatakse histogrammi joonistamisel. Tihedus on normaliseeritud sagedus, nii et histogrammi pindala on 1. See võimaldab histogrammi ja tihedusfunktsiooni kõrvuti kuvada.
   - `fill="white"` määrab histogrammi kastide sisemise värvi valgeks.
   - `color="black"` määrab histogrammi kastide piirjoone värvi mustaks.

3. `geom_density(color="red", lwd=1)`: See lisab tihedusfunktsiooni graafikule.
    - `geom_density()` funktsioon arvutab ja joonistab andmete tihedusfunktsiooni, mis on silutud kõver, mis näitab andmete jaotuse kuju.
    - `color="red"` määrab tihedusfunktsiooni joone värvi punaseks.
    - `lwd=1` määrab joone paksuse (line width) 1-ks.
4. `labs(y = "f(x)")`: See lisab y-teljele sildi "f(x)". See on hea tava, et telgi selgelt märgistada, eriti kui y-telg näitab tihedust, mitte sagedust. "f(x)" on levinud tähistus tõenäosustihedusfunktsioonile (probability density function).

Sageduste arvutamine

Tulpdiagrammile võib eelnevalt välja arvutada antud sagedused. Selleks saab kasutada funktsiooni `cut()`, mis vajab rühmade piire sisendiks ning seejärel saab juba funktsiooniga `table()` loendada kokku iga rühma liikmete arvu.

3.2.2 Exceli keskkonnas:

3.2.2.1 Histogramm

Histogrammi loomiseks Excelis peab andmed esmalt sisestama tabelisse ja seejärel kasutama selle jaoks sobivat tööriista.

1. Sisesta pideva tunnuse väärtused Exceli tabelisse.
2. Vali need lahtrid, kuhu soovid luua histogrammi.
3. Mine menüüsse "Lisa" ja vali "Diagramm".
4. Vali "Histogramm" ja klikka "OK".
5. Seejärel on võimalik vormindada telje suvandeid, millega määratakse rühmade (MS Exceli keskkonnas nimetusega "salv") parameetrid.

3.2.2.2 Sageduste arvutamine

MS Exceli funktsioon `FREQUENCY()` võimaldab kasutajatel määrata, kui sageli väärtused esinevad teatud väärtusvahemikes. See funktsioon sobib hästi suurte andmekogumite analüüsimiseks, et mõista andmete jaotust ilma iga üksiku väärtuse manuaalse üle vaatamiseta.

Enne `FREQUENCY()` funktsiooni kasutamist peate määrama rühmade vahemike piirid, millesse soovite oma andmed jaotada. Need piirid tuleks sisestada eraldi veergu Exceli töölehel. Näiteks, kui soovite analüüsida testitulemusi vahemikus 0-100, võite määrata piirid 0, 20, 40, 60, 80, 100.

1. **Andmete ja piiride sisestamine:**

- Sisestage oma andmekogum ühte veergu (nt A2:A101).

- Sisestage vahemike piirid teise veergu (nt B2:B7, eeldades, et esitasite näiteks eelmises punktis toodud piirid).

2. **Funktsiooni rakendamine:**

- Valige tühi ala, kuhu soovite tulemused väljastada. Sellel peaks olema sama palju lahtrid kui määratletud vahemike piire. Kui teil on 6 piiri, valige 6 lahtrit vertikaalselt.
- Sisestage `FREQUENCY()` funktsioon. Kuna `FREQUENCY()` on massiivifunktsioon, tuleb see sisestada massiivivalemiga. Algusesse minev andmevahemik on teie andmekogum ja teine vahemik on teie vahemike piirid. Näiteks: `=FREQUENCY(A2:A101, B2:B7)`
- Pärast funktsiooni sisestamist lõpetage sisestus, vajutades **Ctrl+Shift+Enter**. Excel käitab nüüd `FREQUENCY()` funktsiooni massiivina ja täidab valitud lahtrid andmete sagedustega, mis vastavad määratud vahemikele.

Peatükk 4

Karakteristikud

Statistilised karakteristikud on kvantitatiivsed või kvalitatiivsed omadused, mis kirjeldavad ja iseloomustavad andmekogumeid või andmete jaotusi. Neid kasutatakse andmete kokkuvõtlikuks esitamiseks, et saada aru andmete üldisest käitumisest, trendidest ja omadustest. Statistilised karakteristikud võimaldavad teha järeldusi kogu populatsiooni kohta, analüüsides selleks valimit. Peamised statistilised karakteristikud jagunevad kaheks suureks rühmaks: keskmise ja hajuvuse karakteristikud.

Juhuslik suurus on täielikult kirjeldatud üldkogumi jaotus- või tihedusfunktsiooniga (jaotusega). Juhusliku suuruse iseloomustamiseks teatud aspektist kasutatakse erinevaid arvkarakteristikuid. Neist tähtsaim on **keskväärtus** ehk matemaatiline ootus, mida tähistatakse EX ja mis on defineeritud järgnevalt.

Pideva juhusliku suuruse korral:

$$EX = \sum_{-\infty}^{\infty} x \cdot f(x)dx$$

Diskreetse juhusliku suuruse korral:

$$EX = \sum_{i=1}^{n} x_i \cdot p_i$$

4.1 Keskmised

4.1.1 Aritmeetiline keskmine

Juhusliku suuruse jaotusseadus on uurijale enamasti teadmata. Seetõttu kasutatakse üldkogumi karakteristikute hindamiseks valimi vastavaid karakteristikuid. Üldkogumi keskväärtusele vastab valimi aritmeetiline keskmine, mida tähistatakse ülakriipsuga vastava muutuja kohal $\bar{x}$ ja arvutatakse järgnevalt.

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{k} n_i \cdot x_i$$

Kus x_i on iga andmehulga liige ning n on andmehulga suurus.

R keskkonnas:

```
# Laadime proovitükkide andmed
df <- readxl::read_excel("data/naited.xlsx","prt_andmed")
# Tunnusest arvutatakse keskmine:
mean(df$d)
```

```
[1] 23,3
```

Exceli keskkonnas:

```
=MEAN(andme_vektor)
```

4.1.2 Ruutkeskmine

Ruutkeskmine on üks keskmise leidmise meetoditest, mis annab ülevaate andmehulgast ning aitab mõõta andmete hajuvust. Ruutkeskmise leidmiseks tuleb iga andmehulga liige ruutu võtta, seejärel leida nende aritmeetiline keskmine ning võtta sellest ruutjuur. Matemaatiliselt väljendatuna on ruutkeskmine järgmine:

$$\bar{x}_{ruut} = \sqrt{\frac{1}{N}\sum_{i=1}^{k} n_i \cdot x_i^2}$$

,

kus x_i on iga andmehulga liige ning n on andmehulga suurus.

Metsanduses on ruutkeskmine kasutusel keskmise diameetri arvutamisel.

R keskkonnas:

```
# Tunnusest arvutatakse ruut ning keskmisest võetakse ruutjuur:
sqrt(mean(df$d^2))
```

```
[1] 23,8
```

Exceli keskkonnas:

```
=SQRT(MEAN(andmeplokk^2))
```

4.1.3 Geomeetriline keskmine

Geomeetriline keskmine on keskmise arvutamise meetod, mille puhul kõik antud arvud korrutatakse omavahel ning seejärel võetakse saadud korrutisest n-nda astme juur, kus n tähistab arvude kogust. Geomeetriline keskmine $\bar{x}_{geom}$ arvutatakse järgneva valemiga:

$$\bar{x}_{geom} = \exp\left(\frac{1}{N}\sum_{i=n}^{k} n_i \cdot \ln(x_i)\right)$$

kus:

- n on arvude kogus
- $x_1, x_2, \ldots, x_n$ on positiivsed arvud

R keskkonnas:

```
# Tunnusest arvutatakse logaritm ning keskmisest võetakse exponent:
exp(mean(log(df$d)))
```

```
[1] 22,7
```

Exceli keskkonnas:

```
=GEOMEAN(andmeplokk)
```

4.1.4 Harmooniline keskmine

Harmooniline keskmine on statistiline mõõde, mis arvutatakse jagades arvude hulk nende vastandväärtuste summa järgi. Harmooniline keskmine on erinev aritmeetilisest ja geomeetrilisest keskmisest ning seda kasutatakse eelkõige siis, kui on vaja leida keskmine, mis arvestab vastandväärtusi või suhtarve.

Harmoonilist keskmist kasutatakse sageli olukordades, kus on vaja arvutada keskmine mingis suhtarvulises kontekstis, näiteks kiiruste, ajade või muude vastandväärtustega seotud suuruste puhul. Näiteks võib harmooniline keskmine kiirus olla kasulik, kui on vaja arvutada keskmine kiirus teepikkuse ja sõiduaegadega seotud olukorras.

Harmoonilist keskmist arvutatakse valemiga:

$$\bar{x}_{harm} = \frac{N}{\sum_{i=1}^{k} n_i/x_i},$$

kus $x_1, x_2, \ldots, x_n$ on arvude hulk, mille harmoonilist keskmist soovitakse leida ning n on arvude hulga suurus.

R keskkonnas:

```
# Tunnuse pöördväärtuse keskmisest võetakse pöördväärtus:
1/mean(1/df$d)
```

```
[1] 22,2
```

Exceli keskkonnas:

```
=HARMEAN(andmeplokk)
```

4.2 Kvantiilid

Juhusliku suuruse p-kvantiiliks (0 < p < 1) nimetatakse sellist juhusliku suuruse väärtust xp, millest väiksemate väärtuste esinemise tõenäosus on p:

$$P(X < x_p) = p$$

Teatud kvantiilide jaoks kasutatakse erinimetusi:

- $x_{0,50}$ – mediaan, tähistatakse ka Me;
- $x_{0,25}$ – alumine kvartiil;
- $x_{0,75}$ – ülemine kvartiil;
- $x_{0,10}$ – alumine detsiil;
- $x_{0,9}$ – ülemine detsiil;
- $x_{0,01}$ – alumine protsentiil;
- $x_{0,99}$ – ülemine protsentiil.

Enamkasutatavam kvantiil on **mediaan**, mida võib pidada (nagu keskväärtustki) juhusliku suuruse tsentriks. Mediaan on juhusliku suuruse selline väärtus, millest nii väiksemate kui ka suuremate väärtuste esinemise tõenäosus on 0,5. Mediaan iseloomustab juhusliku suuruse asendi poolest keskmist väärtust. Kui on teada juhusliku suuruse jaotusfunktsioon F(x), siis saab juhusliku suuruse kvantiili xp arvutada kui jaotusfunktsiooni pöördfunktsiooni, sest kvantiili definitsioonist tuleneb F(xp) = p. Juhul kui on tegemist vaatlusandmestikuga, saab kvantiile ligikaudu hinnata empiiriliselt jaotusfunktsiooni graafikult.

4.3 Mediaan

Mediaan väärtus, mis jagab andmed kaheks võrdselt suureks pooleks. See ei sõltu ekstreemsetest väärtustest ega hajutatud andmetest.

Mediaani arvutamine:

```
median(df$d)
```

```
[1] 23,2
```

Exceli keskkonnas:

```
=MEDIAN(andmeplokk)
```

Peatükk 5

Jaotusfunktsioonid

Statistikas nimetatakse **jaotusfunktsiooniks** funktsiooni, mis kirjeldab juhusliku suuruse X tõenäosusjaotust. See määrab iga reaalarvu x korral tõenäosuse, et juhusliku suuruse X väärtus on väiksem või võrdne x-ga. Formaalselt väljendub jaotusfunktsioon kui $F(x) = P(X \leq x)$, seostes iga x-i väärtuse vastavusse tõenäosuse $P(X \leq x)$.

Jaotusfunktsiooni saab kasutada nii pidevate kui ka diskreetsete juhuslike suuruste puhul. Pideva juhusliku suuruse korral on jaotusfunktsioon integraal tema tihedusfunktsioonist. Diskreetse juhusliku suuruse korral on jaotusfunktsioon kumulatiivne tõenäosuste summa.

Jaotusfunktsioonil on järgmised omadused:

- See on *monotoonselt kasvav*, st $F(x) \leq F(y)$ kui $x \leq y$.
- See on *paremalt pidev*, st $\lim_{x \to y^+} F(x) = F(y)$ iga y korral.
- See läheneb 0-le kui x läheneb $-\infty$-le, ja 1-le kui x läheneb $+\infty$-le.

Jaotusfunktsiooni saab kasutada juhusliku suuruse tõenäosusjaotuse visualiseerimiseks. Selle abil saab arvutada ka tõenäosusi, et juhuslik suurus langeb teatud vahemikku. Näiteks tõenäosus, et X langeb vahemikku [a, b], on F(b) - F(a).

Jaotusfunktsioon on statistikas väga oluline tööriist. See aitab mõista juhuslike suuruste käitumist ja teha järeldusi andmete kohta.

Lisaks ülaltoodule on jaotusfunktsioonil veel teisi olulisi omadusi ja kasutusvõimalusi. Näiteks:

- Jaotusfunktsiooni abil saab leida juhusliku suuruse kvantiile. Kvantiil on väärtus, millest väiksemaks või võrdseks jääb juhusliku suuruse väärtus teatud tõenäosusega.
- Jaotusfunktsiooni abil saab arvutada juhusliku suuruse keskväärtust, dispersiooni ja teisi momente.
- Jaotusfunktsiooni abil saab genereerida juhuslikke arve, mis järgivad teatud jaotust.

Jaotusfunktsioon on statistikas fundamentaalne mõiste ja sellel on lai valik rakendusi.

Teoreetilisi jaotusi kasutatakse hüpoteeside testimisel statistilistes analüüsides selleks, et teha kindlaks, kas vaatlusandmete hajuvus on juhuslik või on olemas mingi muster või seos. Oluline on märkida, et enamik statistilisi teste eeldab teatud jaotuse tundmist, mille põhjal tehakse järeldusi populatsiooni kohta.

Kõige tuntum teoreetiline jaotus on normaaljaotus, mis on sümmeetriline kõver ning annab tihti paljudele looduslikele nähtustele lähedase jaotuse. Normaaljaotus on oluline paljudes statistilistes testimistes, näiteks t-testis või ANOVA analüüsis. Kui andmed järgivad normaaljaotust, saab kasutada erinevaid statistilisi meetodeid.

Lisaks normaaljaotusele on olulised ka teised jaotused, nagu näiteks binoomjaotus (kasutatakse binomiaalsete katsete puhul), eksponentsiaaljaotus (kasutatakse elupikkuste analüüsimisel) ja Poissoni jaotus (kasutatakse haruldaste sündmuste analüüsimisel).

Hüpoteeside testimisel võrreldakse empiirilisi andmeid teoreetiliste jaotustega, et teha järeldusi populatsiooni kohta. Näiteks võib uurida, kas kahe rühma keskmiste vahel on oluline erinevus või kas mingi seos on juhuslik või mitte. Sellisel juhul kasutatakse statistilisi teste, et hinnata, kas tulemus võib olla juhuslik või on tõenäoliselt tingitud tegelikust mõjust või seosest.

5.1 Hüpotees

Hüpotees on statistikas väide või eeldus, millele toetudes teostatakse uuringuid, analüüsi või katseid. Hüpotees võib olla kas uurija poolt esitatud eeldus, mida on vaja tõestada või ümber lükata või juba olemasolev teooria, mida soovitakse kontrollida.

- **Nullhüpotees (Ho)**: See on väide, mida püütakse ümber lükata. Nullhüpotees eeldab, et uuritavas populatsioonis puudub mingi mõju, seos või erinevus. Näiteks võib nullhüpotees väita, et kahe rühma keskmised väärtused on võrdsed või et tunnus ei mõjuta teise tunnuse jaotust. Nullhüpotees on alati konkreetne väide, mille kohta on võimalik arvutada tõenäosust.
- **Sisukas ehk alternatiivne hüpotees (H1)**: See on väide, mida uurija soovib tõestada. See hüpotees esitab oletuse mingi mõju, seose või erinevuse olemasolu kohta populatsioonis. Näiteks võib alternatiivne hüpotees väita, et kahe rühma keskmised väärtused on erinevad või et tunnus mõjutab teise tunnuse jaotust. Sisukaid hüpoteese võib olla ka rohkem kui üks.
- Hüpoteesipaar püstitatakse üldkogumi jaoks.
- Hüpoteeside püstitus sõltub uurimisküsimusest ja sellest, kas oodatakse kindlasuunalist kõrvalekallet või mitte.
  - **Kahepoolne hüpotees:** Kontrollitakse, kas parameeter (nt keskväärtus) erineb mingist väärtusest. Näiteks Ho: $\mu = \mu_0$ ja H1: $\mu \neq \mu_0$.

  - **Ühepoolne hüpotees:** Kontrollitakse, kas parameeter on suurem või väiksem mingist väärtusest. Näiteks H0: $\mu \leq \mu_0$ ja H1: $\mu > \mu_0$ (parempoolne) või H0: $\mu \geq \mu_0$ ja H1: $\mu < \mu_0$ (vasakpoolne).

- **Olulisuse nivoo (α)**: See on tõenäosus, millega ollakse nõus tegema I liiki viga ehk lükkama tagasi kehtiva nullhüpoteesi. Tavaliselt kasutatakse olulisuse nivood 0,05 (5%) või 0,01 (1%), kuid see sõltub uurimisvaldkonnast. Mida väiksem on olulisuse nivoo, seda väiksem on tõenäosus teha I liiki viga. Olulisuse nivoo on maksimaalne lubatud p-väärtus väärtus.

- **Kriitilised väärtused:** Need on teststatistiku väärtused, mis määratlevad kriitilise piirkonna. Kriitilised väärtused leitakse teststatistiku jaotusest vastavalt valitud olulisuse nivool.

  - Kahepoolse hüpoteesi korral on kaks kriitilist väärtust, mis asuvad jaotuse mõlemal pool.
  - Ühepoolse hüpoteesi korral on üks kriitiline väärtus, mis asub kas jaotuse paremal või vasakul pool.
  - Kriitilised väärtused on seotud ka usaldusvahemiku laiusega.

- **Teststatistiku väärtuse võrdlemine kriitilise väärtusega:** Võrreldakse, kas empiiriline teststatistiku väärtus langeb kriitilisse piirkonda.

  - Kui empiiriline väärtus ületab kriitilise väärtuse (langeb kriitilisse piirkonda), siis lükatakse nullhüpotees tagasi ja võetakse vastu sisukas hüpotees.
  - Kui empiiriline väärtus ei lange kriitilisse piirkonda, siis ei ole alust nullhüpoteesi tagasi lükata. Jäädakse nullhüpoteesi juurde, aga see ei tähenda, et nullhüpotees on tõestatud.

- **Olulisuse tõenäosus (p-väärtus):** See on tõenäosus saada antud või veelgi äärmuslikum valim, eeldusel, et nullhüpotees on tõene. Mida väiksem on p-väärtus, seda väiksem on tõenäosus, et nullhüpotees kehtib.

  - Kui p-väärtus on väiksem kui olulisuse nivoo ($p < \alpha$), siis lükatakse nullhüpotees tagasi.
  - Kui p-väärtus on suurem või võrdne olulisuse nivooga ($p \geq \alpha$), siis ei lükata nullhüpoteesi tagasi.
  - Olulisuse tõenäosust saab kasutada nullhüpoteesi sobivuse hindamiseks antud valimiga.

- Teststatistiku empiirilise väärtuse võrdlemine kriitilise väärtusega ja olulisuse tõenäosuse võrdlemine olulisuse nivooga on ekvivalentsed meetodid.

Tuleb arvestada, et nullhüpoteesi ei saa "tõestada", kuna statistiliste testidega ei saa nullhüpoteesi tõestada samamoodi, nagu matemaatikas võidakse tõestada väiteid. Selle asemel saab statistiliste testide abil hinnata, kas on piisavalt tõendeid nullhüpoteesi tagasilükkamiseks. Kui testi tulemus näitab, et andmed on nullhüpoteesiga vastuolus, siis võidakse nullhüpotees tagasi lükata. Kui aga testi tulemus ei näita olulist vastuolu nullhüpoteesiga, siis ei lükata nullhüpoteesi tagasi. Siiski, see ei tähenda, et nullhüpotees oleks "tõestatud"; pigem öeldakse, et ei leitud piisavalt tõendeid nullhüpoteesi ümberlükkamiseks. Oluline on mõista, et nullhüpoteesi mitte-tagasilükkamine ei tõesta selle paikapidavust. See viitab lihtsalt sellele, et andmed ei andnud piisavalt alust nullhüpoteesi ümber lükata. Samuti on oluline meeles pidada, et statistiliste testide tulemused

annavad ainult tõenäosusliku hinnangu ning need tuleb alati panna laiemasse konteksti koos teiste tõendite ja teadmistega valdkonnast.

5.2 Jaotusfunktsioon

Jaotusfunktsioon on viis näidata, kui tõenäoline on, et mingi arv (mida me nimetame muutujaks) on väiksem või võrdne kindla arvuga x. See aitab meil mõista, kuidas arvud (või muutujad) ühes grupis paiknevad või jaotuvad, näiteks kui me tahame teada, kui tõenäoline on, et metsa pindala on teatud suurusega või väiksem.

Jaotusfunktsiooni olulised omadused on:

1. Vahemik: jaotusfunktsioon võtab väärtuseid 0 ja 1 vahemikus, mis tähendab, et tõenäosus on alati selles vahemikus.
2. Mittelangus: jaotusfunktsioon ei kasva kunagi tagurpidi; see liigub alati kas võrdse kõrguse juurde või kõrgemale.
3. Ühekordsus: Ükskõik millised on väärtused, jaotusfunktsiooni koguulatus on alati 1.

Jaotusfunktsiooni abil saame lahendada paljusid statistilisi ülesandeid. See aitab meil leida näiteks keskmisi väärtusi, hajuvust (ehk kui laiali arvud on jaotunud), tõenäosusi (kui tõenäoline on mingi sündmus), ja isegi määrata arvude jaotuse erinevaid tasemeid. See tähendab, et jaotusfunktsioon on tööriist, mis aitab meil mõista, kuidas andmed käituvad, ja teha selle põhjal täpseid otsuseid või järeldusi.

```
# Näide R keskkonnas
# Laadime peatükis kasutatavad paketid
library(dplyr)
library(ggplot2)
library(ggpmisc)
```

- **Andmete jaotus.** ECDF graafik annab intuitiivse ülevaate andmete jaotusest, näidates, kui suur osa andmestikust on alla või võrdne iga võimaliku väärtusega. See on eriti kasulik andmete jaotuse mõistmiseks ilma eeldamata mingit konkreetset jaotustüüpi.

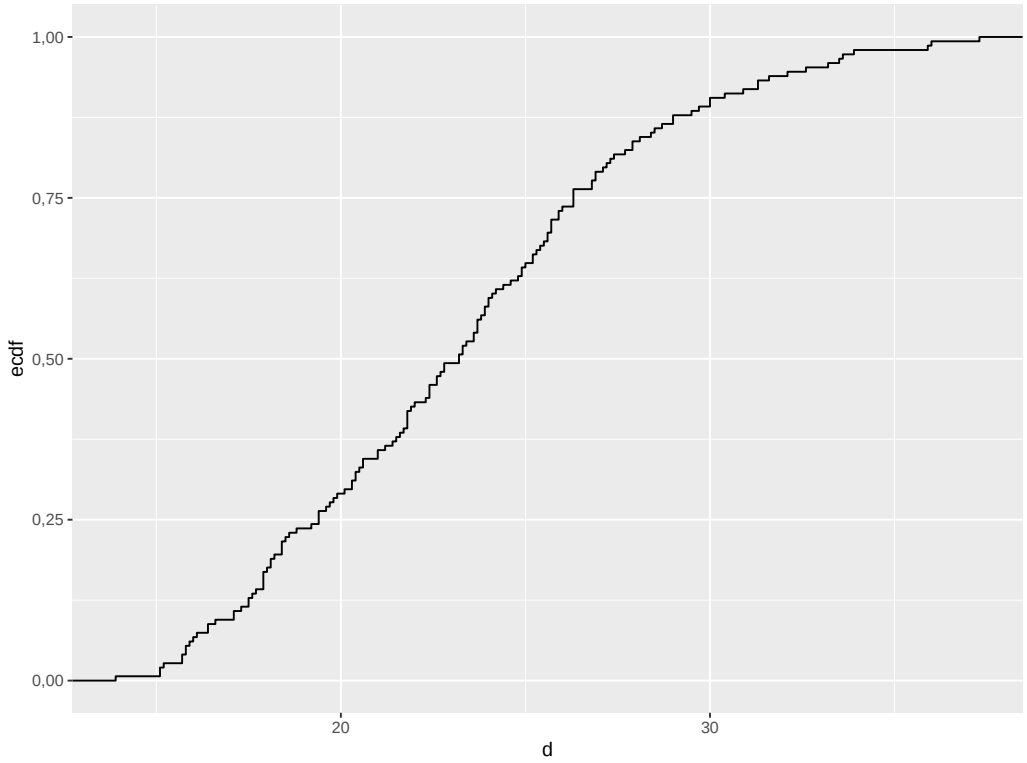
ECDF graafik on väärtuslik tööriist andmete uurimisel, võimaldades kiiresti hinnata andmete jaotuse omadusi ja tuvastada andmete levikuga seotud probleeme.

```
# Laadime proovitükkide andmed
df <- readxl::read_excel("data/naited.xlsx","prt_andmed")

ggplot(df, aes(d)) +
  stat_ecdf()
```

Selgitus:

- `ggplot(df, aes(d))`: See algatab graafiku loomise, kasutades `ggplot2`.
  - `ggplot(df, ...)` määrab, et graafiku andmed pärinevad `df` andmeraamist.



Joonis 5.1. Empiiriline jaotusfunktsioon.

  - `aes(x)` seostab andmeraami veeru `d` graafiku x-teljega. See ütleb ggplot2-le, et veeru `x` väärtused määravad punktide asukoha horisontaalteljel.

- `+ stat_ecdf()`: See lisab graafikule kihi, nimelt ECDF-i.
  - `stat_ecdf()` arvutab ECDF-i. ECDF antud väärtusel x tähistab andmepunktide osakaalu, mis on väiksemad või võrdsed x-ga.

5.3 Tihedusfunktsioon

Tihedusfunktsioon on statistikas kasutatav funktsioon, mis kirjeldab tõenäosust, et juhuslik suurus langeb mingisse kindlasse vahemikku. Tihedusfunktsiooni saab kasutada tõenäosusjaotuse (nt normaaljaotus, Bernoulli jaotus jne) karakteriseerimiseks. See näitab, kui tõenäoline on konkreetse väärtuse saamine selle jaotuse järgi.

Tihedusfunktsioon on graphiline joon, mis esitab tõenäosuse jaotuse ühtlase tervikuna. Üksiku väärtuse tõenäosus on antud tihedusfunktsiooni abil vahemikuna, mida funktsioon katab selles punktis. Tavaliselt on tihedusfunktsioon positiivne ja integreerub üle vahemiku 1, mis tähendab, et kõikidele võimalikele väärtustele vastavate tõenäosuste summa on alati 1.

Metsanduses võib tihedusfunktsiooni kasutada näiteks puistu tiheduse jaotuse kirjeldamisel, kus tiheduse-

funktsioon näitab puude arvu tõenäosust kindlas pindalas. Tihedusfunktsioonide abil saab hinnata erinevaid statistilisi parameetreid nagu keskmine, mediaan, variatsioon jt, ning teha ennustusi ja otsuseid metsaressursi haldamisel ja planeerimisel.

Tihedusfunktsioon (inglise keeles probability density function, lühidalt PDF) on statistikas oluline mõiste, mida kasutatakse tõenäosusjaotuse kirjeldamiseks. Tihedusfunktsioon annab meile informatsiooni selle kohta, kuidas on tõenäosus mingi juhusliku muutuja väärtuse jaotustel. Allpool on toodud mõned olulised statistilised omadused tihedusfunktsiooni kohta:

Integreeruvus: tihedusfunktsioon peab olema integreeruv kogu kandva ala suhtes, st selle integraal alates miinimum- kuni maksimumväärtuseni peab olema võrdne ühega. Teisisõnu, kogu võimaliku vaadeldava mahu (nt kogu metsaala, kus uuritakse puude tihedust) peab olema 100%.

Positiivsus: Tihedusfunktsiooni väärtused peavad olema alati positiivsed. See tähendab, et tõenäosus väärtuse sattumiseks konkreetsesse intervalli ei saa olla negatiivne.

Normaliseerumine: Tihedusfunktsioon peab olema normaliseeritud, mis tähendab, et alati, kui me integreerime tihedusfunktsiooni kogu liinil või tasandil, siis tulemuseks peab olema üks. See tagab, et tihedusfunktsioon annab meile täpse tõenäosuse, et vaadeldav suurus langeb mingisse kindlasse vahemikku.

Mitte-negatiivsus: Tihedusfunktsioon ei või mitte kunagi võtta negatiivseid väärtusi. See tähendab, et tõenäosus mingi suuruse väärtuse jaoks on alati null või suurem.

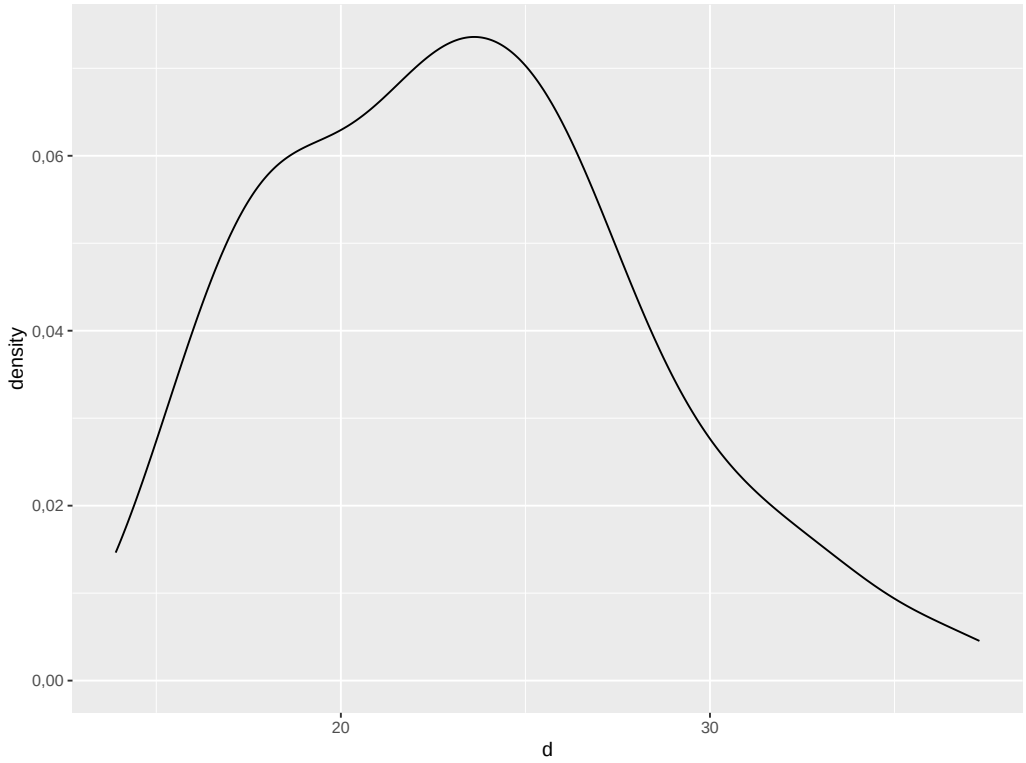
Tipukoha lokaliseerimine: Tihedusfunktsiooni tippkoht näitab moodi ehk suurimat tõenäosust mingi väärtuse saamiseks. Mood on statistikas näitaja, mis näitab, milline väärtus esineb andmehulgas kõige sagedamini ehk mis väärtusega esineb andmehulkade hulgas enim kordi. Mood on üks keskmiste näitajate tüüpe, mis aitab anda ülevaadet andmete esinemissagedusest konkreetse väärtuse juures.

Need omadused on olulised tihedusfunktsiooni mõistmisel ja kasutamisel statistilistes analüüsides, sealhulgas metsanduses, kus tihedusfunktsiooni abil saab kirjeldada erinevaid muutujaid nagu puu tihedus, metsa elujõulisus jne.

Näide R-i koodist, mis genereerib normaaljaotuse tihedusfunktsiooni graafiku:

```
# Loome andmestiku
ggplot(df, aes(d)) +
  geom_density()
```

See kood genereerib normaaljaotuse tihedusfunktsiooni graafiku, kus x-teljel on andmete väärtused ja y-teljel tihedusfunktsiooni väärtused.



Joonis 5.2. Empiiriline tihedusfunktsioon.

5.4 Normaaljaotus

Normaaljaotus on statistikas ja tõenäosusteoorias üks tähtsamaid pidevaid jaotusi. Selle tunnusmärk on kellakujuline sümmeetriline graafik, mistõttu seda nimetatakse sageli ka Gaussi jaotuseks, kuna seda uuris põhjalikult Carl Friedrich Gauss. Normaaljaotuse keskmes paikneb enamik väärtusi, harvemini esinevad väärtused jäävad aga sellest kaugemale.

Normaaljaotuse omadused

- **Sümmeetria:** Normaaljaotus on sümmeetriline keskväärtuse suhtes, mis tähendab, et jaotuse graafiku mõlemad pooled on üksteise peegelpildid.
- **Keskmine, mediaan ja mood:** Normaaljaotuse korral on aritmeetiline keskmine, mediaan ja mood võrdsed.
- **Hajuvus:** Normaaljaotust iseloomustab hajuvus, mida mõõdetakse standardhälbe abil. Standardhälve näitab, kui palju väärtused keskmisest erinevad.
- **Ekstsess (kurtosis):** Normaaljaotuse püstakus on null. Püstakus näitab, kui terav on jaotuse tipp.
- **Asümmeetria (skewness):** Normaaljaotus on asümmeetriline, mis tähendab, et jaotus ei kaldu ei vasakule ega paremale.
- **Väärtuste jaotus:** Normaaljaotuse puhul jääb umbes 68,3% väärtustest ühe standardhälbe kaugusele

keskmisest, umbes 95,4% väärtustest kahe standardhälbe kaugusele keskmisest ja umbes 99,7% väärtustest kolme standardhälbe kaugusele keskmisest.

- **Teoreetiline ulatus:** Teoreetiliselt võib normaaljaotusega muutuja omada väärtusi vahemikus −inf kuni inf.

Normaaljaotuse kasutusalad

Normaaljaotust kasutatakse laialdaselt statistikas ja erinevates teadusvaldkondades.

- **Andmete kirjeldamine ja analüüs:**
  - Normaaljaotust kasutatakse üldkogumi keskväärtuse usaldusvahemike arvutamiseks.
  - Normaaljaotust kasutatakse sageli hüpoteeside testimisel.
  - Normaaljaotust kasutatakse andmete standardiseerimiseks.
  - Normaaljaotust saab kasutada erindite tuvastamiseks.
- **Tsentraalne piirteoreem:**
  - Tsentraalse piirteoreemi kohaselt alluvad piisavalt suure valimi keskmised ligikaudu normaaljaotusele.
- **Statistilised mudelid:**
  - Normaaljaotust kasutatakse regressioonimudelite vigade jaotuse eelduseks.
- **Tõenäosusteooria:**
  - Normaaljaotust kasutatakse tõenäosusteooria arvutustes.
- **Rakendusalad:**
  - Normaaljaotust kasutatakse paljudes valdkondades, nagu majandus, meditsiin, inseneriteadused, psühholoogia ja keskkonnateadused.

Standardiseeritud normaaljaotus

Standardiseeritud normaaljaotus on normaaljaotuse erijuhtum, mille keskmine on 0 ja standardhälve on 1.

Normaaljaotuse piirangud

- Normaaljaotus eeldab andmete sümmeetriat ja teatud hajuvust.
- Normaaljaotus ei ole sobiv ekstremaalsete väärtuste analüüsimiseks.
- Normaaljaotuse omadused kehtivad suure valimi korral.

R-is on järgnevad normaaljaotuse funktsioonid:

- `pnorm()` arvutab tõenäosuse, et juhuslik muutuja on väiksem või võrdne kui antud väärtus.
- `dnorm()` arvutab tõenäosustihedusfunktsiooni väärtuse antud punktis.
- `qnorm()` arvutab väärtuse, millest väiksem või võrdne on antud protsent andmetest.

```
pnorm(1.96, 0, 1)
```

```
[1] 0,975
```

```
dnorm(1.96, 0, 1)
```

```
[1] 0,0584
```

```
qnorm(0.975, 0, 1)
```

```
[1] 1,96
```

MS Exceli keskkonnas normaaljaotusega seotud arvutuste tegemiseks saab kasutada järgmisi funktsioone:

- Näide: NORM.DIST(1,96; 0; 1; TRUE) annab tõenäosuse, et normaaljaotuse väärtus on väiksem või võrdne 1.96.
- Näide: NORM.S.DIST(1,96; TRUE) annab tõenäosuse, et standardnormaaljaotuse väärtus on väiksem või võrdne 1.96.
- Näide: NORM.INV(0,975; 0; 1) annab normaaljaotuse kvantiili, mis vastab 0.975 tõenäosusele.
- Näide: `NORM.S.INV(0,975)` annab standardnormaaljaotuse kvantiili, mis vastab 0.975 tõenäosusele.

5.5 T-jaotus

T-jaotus, tuntud ka kui Studenti t-jaotus, on pidev tõenäosusjaotus, mida kasutatakse statistilises järeldamisel, eriti kui valimi suurus on väike ja populatsiooni standardhälve on teadmata. T-jaotus sarnaneb kujult normaaljaotusega, kuid on lamedama tipu ja raskemate sabadega. See tähendab, et t-jaotuses on suurem tõenäosus saada äärmuslikke väärtusi võrreldes normaaljaotusega. T-jaotus on sümmeetriline nulli suhtes.

T-jaotuse omadused

1. **Sümmeetria:** T-jaotus on sümmeetriline nulli suhtes, mis tähendab, et jaotuse graafiku mõlemad pooled on üksteise peegelpildid.
2. **Kuju:** T-jaotus on kellakujuline, sarnaselt normaaljaotusega, aga selle kuju sõltub vabadusastmetest. Mida väiksem on vabadusastmete arv, seda lamedam on jaotus ja raskemad sabad. Suuremate vabadusastmete arvu korral läheneb t-jaotus normaaljaotusele.
3. **Vabadusastmed:** T-jaotust iseloomustab parameeter, mida nimetatakse vabadusastmeteks (ν). Vabadusastmed on seotud valimi suurusega ja neid arvutatakse erinevalt olenevalt rakendusest. Näiteks ühe valimi t-testi korral on vabadusastmete arv tavaliselt n-1, kus n on valimi suurus.
4. **Keskmine, mediaan ja mood:** T-jaotuse keskväärtus, mediaan ja mood on kõik null, kui jaotuse keskmine on null.
5. **Hajuvus:** T-jaotuse hajuvus on suurem kui standardiseeritud normaaljaotuse hajuvus, eriti väikeste vabadusastmete korral. See peegeldab ebakindlust, mis tekib, kui populatsiooni standardhälve on teadmata ja seda tuleb hinnata valimi standardhälbe abil.

T-jaotuse kasutusalad

T-jaotust kasutatakse peamiselt olukordades, kus populatsiooni standardhälve on tundmatu ja valimi suurus

on väike. T-jaotus on eriti oluline järgmistes statistikameetodites:

1. **Ühe valimi t-test:**
   - Kasutatakse hüpoteeside testimiseks ühe valimi keskmise kohta, kui populatsiooni standardhälve ei ole teada.
   - Näiteks saab t-testi abil kontrollida, kas valimi keskmine erineb oluliselt etteantud väärtusest.
2. **Kahe sõltumatu valimi t-test:**
   - Kasutatakse kahe sõltumatu valimi keskmiste võrdlemiseks, kui populatsiooni standardhälbed ei ole teada.
   - Näiteks saab t-testi abil kontrollida, kas kahe erineva grupi keskmised väärtused erinevad oluliselt.
3. **Kahe sõltuva valimi t-test:**
   - Kasutatakse kahe sõltuva valimi (näiteks enne ja pärast mõõtmised samade isikute peal) keskmiste võrdlemiseks.
   - Näiteks saab seda kasutada ravimi efektiivsuse hindamiseks, võrreldes patsiendi seisundit enne ja pärast ravi.
4. **Regressioonanalüüs:**
   - Kasutatakse regressioonimudeli parameetrite olulisuse testimiseks.
   - T-jaotust kasutatakse standardiseeritud regressioonikordaja (t-väärtuse) jaotuse hindamisel.
5. **Usaldusvahemikud:**
   - Kasutatakse populatsiooni keskmise usaldusvahemiku leidmiseks, kui populatsiooni standardhälve ei ole teada.
   - T-jaotust kasutatakse juhul, kui valimi suurus on väike ja/või standardhälve pole teada, et määrata usaldusvahemikku, millesse tegelik keskmine tõenäosusega jääb.
6. **Hüpoteeside testimine:**
   - T-jaotust kasutatakse hüpoteeside testimisel, kui valimi suurus on väike või populatsiooni standardhälve ei ole teada.
   - T-jaotust kasutatakse kriitilise väärtuse määramiseks ja otsustamaks, kas nullhüpotees lükatakse tagasi või mitte.
7. **Korrelatsioonianalüüs**:
   - Kasutatakse korrelatsioonikordaja olulisuse testimiseks..
   - T-jaotust kasutatakse, et testida, kas leitud korrelatsioonikordaja on statistiliselt oluline ehk erineb oluliselt nullist.

T-jaotuse ja normaaljaotuse võrdlus

- **Sarnasus:** Mõlemad on sümmeetrilised, kellakujulised ja pidevad jaotused. Suurte vabadusastmete korral läheneb t-jaotus normaaljaotusele.
- **Erinevus:** T-jaotus on lamedama tipu ja raskemate sabadega kui normaaljaotus, eriti väikeste vabadus-

astmete korral. See kajastab suuremat ebakindlust, mis tekib, kui populatsiooni standardhälve on teadmata. Normaaljaotust kasutatakse juhul, kui populatsiooni standardhälve on teada.

T-jaotuse kriitilised väärtused

- Hüpoteeside testimisel kasutatakse t-jaotuse kriitilisi väärtusi, et määrata, kas valimi tulemused on statistiliselt olulised.
- Kriitilised väärtused leitakse t-jaotuse tabelitest või statistikatarkvara abil, võttes arvesse vabadusastmeid ja olulisuse nivoo (α).
- Olulisuse nivoo (α) on tõenäosus, et nullhüpotees lükatakse tagasi, kui see on tegelikult õige.
- Kui teststatistiku väärtus on suurem kui kriitiline väärtus, siis nullhüpotees lükatakse tagasi.

Kokkuvõttes, t-jaotus on oluline statistiline tööriist, mida kasutatakse juhul, kui populatsiooni standardhälve on teadmata ja valimi suurus on väike. See võimaldab teha statistilisi järeldusi ja testida hüpoteese usaldusväärselt.

Järgmise sammuna soovitaksin uurida, kuidas t-jaotust kasutatakse erinevates hüpoteeside testimise stsenaariumides ja kuidas arvutada usaldusvahemikke, et paremini mõista selle praktilist rakendust.

- **Näide**: pt(2.26, 10) arvutab tõenäosuse, et t-jaotuse väärtus on väiksem või võrdne 2.26, kui vabadusastmete arv

- **Näide**: dt(2.26, 10) arvutab t-jaotuse tõenäosustihedust väärtuses 2.26, kui vabadusastmete arv on 10.
- **Näide**: qt(0.975, 10) arvutab t-jaotuse ülemise kvantiili 0.975 väärtuse, kui vabadusastmete arv on 10.

Exceli keskkonnas t-jaotusega seotud arvutuste tegemiseks on olemas järgmised funktsioonid:

1. **T.DIST**:

- **Kirjeldus**: T.DIST funktsioon arvutab Studenti t-jaotuse kumulatiivse tõenäosuse.
- **Kasutamine**: T.DIST(x; vabadusastmed; kumulatiiv) kus x on t-jaotuse väärtus, mille kumulatiivset tõenäosust soovite arvutada, vabadusastmed tähistavad t-jaotuse vabadusastmete arvu, ja kumulatiiv on loogiline väärtus, mis määrab, kas funktsioon tagastab tõenäosuse (TRUE) või tiheduse (FALSE).
- **Näide**: T.DIST(2,26; 10; TRUE) annab tõenäosuse, et t-jaotuse väärtus on väiksem või võrdne 2,26, kui vabadusastmete arv on 10.

2. **T.DIST.RT**:

- **Kirjeldus**: T.DIST.RT funktsioon arvutab ühepoolse tõenäosuse paremast sabast Studenti t-jaotuses.
- **Kasutamine**: T.DIST.RT(x; vabadusastmed) kus x on t-jaotuse väärtus, mille ühepoolset tõenäosust paremast sabast soovite arvutada, ja vabadusastmed on t-jaotuse vabadusastmete arv.
- **Näide**: T.DIST.RT(2,26; 10) annab tõenäosuse, et t-jaotuse väärtus on suurem kui 2,26, kui vabadusastmete arv on 10.

3. **T.DIST.2T**:

- **Kirjeldus**: T.DIST.2T funktsioon arvutab kahepoolse tõenäosuse Studenti t-jaotuses.
- **Kasutamine**: T.DIST.2T(x; vabadusastmed) kus x on t-jaotuse väärtus, mille kahepoolset tõenäosust soovite arvutada, ja vabadusastmed on t-jaotuse vabadusastmete arv.
- **Näide**: T.DIST.2T(2,26; 10) annab tõenäosuse, et t-jaotuse absoluutne väärtus on suurem kui 2,26, kui vabadusastmete arv on 10.

4. **T.INV**:

- **Kirjeldus**: T.INV funktsioon arvutab Studenti t-jaotuse kvantiili antud kumulatiivse tõenäosuse jaoks.
- **Kasutamine**: T.INV(tõenäosus; vabadusastmed) kus tõenäosus on soovitud kumulatiivne tõenäosus, ja vabadusastmed tähistavad t-jaotuse vabadusastmete arvu.
- **Näide**: T.INV(0,975; 10) annab t-jaotuse kvantiili, mis vastab 0,975 kumulatiivsele tõenäosusele, kui vabadusastmete arv on 10.

5. **T.INV.2T**:

- **Kirjeldus**: T.INV.2T funktsioon arvutab kahepoolse testi kriitilise väärtuse Studenti t-jaotuses.
- **Kasutamine**: T.INV.2T(tõenäosus; vabadusastmed) kus tõenäosus on soovitud kahepoolse testi tõenäosus, ja vabadusastmed on t-jaotuse vabadusastmete arv.
- **Näide**: T.INV.2T(0,05; 10) annab kahepoolse test

5.6 F-jaotus

F-jaotus on statistikas oluline pidev tõenäosusjaotus, mida kasutatakse peamiselt dispersioonanalüüsis (ANOVA) ja hüpoteeside testimisel, kui on vaja võrrelda kahe või enama valimi dispersioone. See jaotus aitab kindlaks teha, kas valimite varieeruvus on oluliselt erinev. F-jaotus on asümmeetriline ja selle kuju sõltub kahest parameetrist: loendaja vabadusastmetest ($\nu 1$) ja nimetaja vabadusastmetest ($\nu 2$).

F-jaotuse omadused

1. **Pidevus:** F-jaotus on pidev, mis tähendab, et muutuja võib omandada suvalise väärtuse teatud vahemikus.
2. **Mitte-negatiivsus:** F-jaotuse väärtused on alati positiivsed või null.
3. **Asümmeetria:** F-jaotus on parempoolse asümmeetriaga.
4. **Kuju:** F-jaotuse kuju sõltub loendaja ($\nu 1$) ja nimetaja ($\nu 2$) vabadusastmetest.
5. **Vabadusastmed:** Vabadusastmed määravad jaotuse kuju. Suurte vabadusastmete korral läheneb F-jaotus normaaljaotusele.
6. **Keskmine ja dispersioon:** F-jaotuse keskmine on ligikaudu võrdne $\nu 2 / (\nu 2 - 2)$ (kui $\nu 2 > 2$) ja dispersioon sõltub vabadusastmetest.

F-jaotuse kasutusalad

F-jaotust kasutatakse peamiselt järgmistes statistilistes meetodites:

1. **Dispersioonanalüüs (ANOVA):**
   - Võrreldakse kahe või enama valimi keskmiste erinevusi.
   - Testitakse, kas valimite keskmised erinevad oluliselt, analüüsides dispersiooni.
   - F-statistik arvutatakse valimitevahelise dispersiooni ja valimisisese dispersiooni suhtena.
2. **Hüpoteeside testimine dispersioonide kohta:**
   - Võrreldakse kahe populatsiooni dispersioone.
   - Testitakse, kas kahe valimi dispersioonid on võrdsed.
3. **Regressioonanalüüs:**
   - Testitakse regressioonimudeli sobivust.
   - Kontrollitakse, kas mudel on statistiliselt oluline.
4. **Mudeli võrdlus:**
   - Võrreldakse statistilisi mudeleid, et testida, kas keerukam mudel sobib andmetele paremini.
5. **Mitme võrdluse probleemid:**
   - Lahendatakse mitme võrdluse probleem, et vältida I tüübi vea suurenemist.

F-jaotuse kriitilised väärtused

- Määratakse, kas teststatistiku väärtus on statistiliselt oluline.
- Leitakse F-jaotuse tabelitest või statistikatarkvara abil.

F-jaotuse ja teiste jaotuste seosed

- **Seos t-jaotusega:** Kui võtta t-jaotusega muutuja ruut, siis see allub F-jaotusele.
- **Seos χ^2-jaotusega:** F-statistikut saab vaadelda kui kahe sõltumatu χ^2-jaotusega suuruse suhet.
- **Näide**: pf(3.84, 1, 30) annab tõenäosuse, et F-jaotuse väärtus on väiksem või võrdne 3.84, kui esimese
- **Näide**: df(3.84, 1, 30) annab F-jaotuse tõenäosustihedust väärtuses 3.84, kui esimese andmestiku vabadusastmete arv on 1 ja teise andmestiku vabadusastmete arv on 30.
- **Näide**: qf(0.95, 1, 30) annab F-jaotuse ülemise kvantiili 0.95 väärtuse, kui esimese andmestiku vabadusastmete arv on 1 ja teise andmestiku vabadusastmete arv on 30.

Excelis F-jaotusega seotud arvutuste tegemiseks Euroopa piirkondades, kus kümnendkoha eraldajaks on koma ja argumentide eraldajaks on semikoolon, saate kasutada järgmisi funktsioone:

- **Näide**: F.DIST(3;10;5;TRUE) annab tõenäosuse, et F-jaotuse väärtus on väiksem või võrdne 3, kui esimese valimi vabadusastmete arv on 10 ja teise valimi oma on 5.

- **Näide**: F.DIST.RT(3;10;5) annab tõenäosuse, et F-jaotuse väärtus on suurem kui 3, kui esimese valimi vabadusastmete arv on 10 ja teise valimi oma on 5.
- **Näide**: F.INV(0,975;10;5) annab F-jaotuse kvantiili, mis vastab 0,975 kumulatiivsele tõenäosusele, kui esimese valimi vabadusastmete arv on 10 ja teise valimi oma on 5.
- **Näide**: F.INV.RT(0,025;10;5) annab F-jaotuse kvantiili, mis vastab 0,025 tõenäosusele paremast sabast

5.7 Hii-ruut-jaotus

Hii-ruut-jaotus (χ^2-jaotus) on pidev tõenäosusjaotus, mida kasutatakse statistikas peamiselt kategooriliste andmete analüüsimiseks ja hüpoteeside testimiseks. See on eriti kasulik, kui soovitakse uurida, kas empiirilised andmed vastavad teoreetilisele jaotusele või kas kahe või enama kategoorilise muutuja vahel esineb seos. Hii-ruut-jaotus on mitte-negatiivne ja asümmeetriline jaotus, mille kuju sõltub ühest parameetrist – vabadusastmete arvust.

Hii-ruut-jaotuse omadused:

1. **Pidevus:** Hii-ruut-jaotus on pidev jaotus, mis tähendab, et muutuja võib omandada suvalisi väärtusi teatud vahemikus.
2. **Mitte-negatiivsus:** Hii-ruut-jaotuse väärtused on alati positiivsed või null. See tuleneb sellest, et hii-ruut-statistik arvutatakse ruutude summana, mis on alati mitte-negatiivsed.
3. **Asümmeetria:** Hii-ruut-jaotus on parempoolse asümmeetriaga, mis tähendab, et selle pikem saba asub suuremate väärtuste pool. Mida väiksem on vabadusastmete arv, seda suurem on asümmeetria. Suuremate vabadusastmete arvu korral läheneb hii-ruut-jaotus normaaljaotusele.
4. **Kuju:** Hii-ruut-jaotuse kuju sõltub vabadusastmete arvust (ν). Vabadusastmed määravad jaotuse kuju ja varieeruvuse.
5. **Vabadusastmed:** Vabadusastmed on seotud andmete struktuuriga. Näiteks, kui testitakse hüpoteesi, kas empiirilised sagedused vastavad oodatavatele sagedustele, siis on vabadusastmete arv võrdne kategooriate arvuga, millest on lahutatud üks (või mitu, sõltuvalt piirangute arvust).
6. **Keskmine ja dispersioon:** Hii-ruut-jaotuse keskmine võrdub vabadusastmete arvuga ($\mu = \nu$) ja dispersioon on kaks korda vabadusastmete arv ($\sigma^2 = 2\nu$).

Hii-ruut-jaotust kasutatakse peamiselt järgmistes statistilistes meetodites:

1. **Sobivustest (Goodness-of-fit test):**
   - Kasutatakse selleks, et testida, kas empiirilised sagedused (vaadeldud andmed) vastavad teatud teoreetilisele jaotusele (näiteks normaaljaotus, ühtlane jaotus).
   - Hii-ruut-statistik arvutatakse võrreldes vaadeldud ja oodatavaid sagedusi.
   - Kui hii-ruut-statistik on suur, siis tähendab see, et vaadeldud andmed erinevad oluliselt oodatavast jaotusest.

- Näiteks saab kontrollida, kas mündiviske tulemused (kull või kiri) vastavad oodatavale 50/50 jaotusele.

2. **Sõltumatuse test (Test of independence):**
   - Kasutatakse selleks, et testida, kas kaks kategoorilist muutujat on üksteisest sõltumatud või mitte.
   - Hii-ruut-statistik arvutatakse võrreldes vaadeldud ja oodatavaid sagedusi risttabelis.
   - Kui hii-ruut-statistik on suur, siis tähendab see, et muutujad ei ole sõltumatud, vaid esineb seos.
   - Näiteks saab kontrollida, kas sugu ja erakondlik kuuluvus on omavahel seotud.
3. **Homogeensuse test (Test of homogeneity):**
   - Kasutatakse selleks, et testida, kas mitmest erinevast populatsioonist saadud valimite jaotused on sarnased.
   - Hii-ruut-statistik arvutatakse sarnaselt sõltumatuse testile, võrreldes vaadeldud ja oodatavaid sagedusi.
   - Kui hii-ruut-statistik on suur, siis tähendab see, et populatsioonide jaotused on erinevad.
   - Näiteks saab kontrollida, kas erinevates haiglates on patsientide jaotus sarnane erinevate haiguste järgi.
4. **Variatsiooni testimine:**
   - Hii-ruut-jaotust saab kasutada ka dispersioonide võrdlemiseks, kuigi F-jaotus on selleks otstarbeks tavaliselt eelistatud. Hii-ruut-testi dispersioonide võrdlemiseks on vähem efektiivne, aga võib olla kasulik, kui eeldused F-testi jaoks ei ole täidetud.
   - Hii-ruut-jaotust saab kasutada, et testida kas valimi dispersioon erineb oluliselt etteantud populatsiooni dispersioonist.
5. **Informatsiooni kriteeriumid:**
   - Hii-ruut-jaotust kasutatakse ka infokriteeriumite (nt AIC ja BIC) arvutamisel, mis aitavad valida sobivaimat mudelit antud andmete jaoks.

Hii-ruut statistiku valem on järgmine:

$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$

Kus:

- χ^2 on hii-ruut-statistik
- $\sum$ on summa
- O_i on vaadeldud sagedus
- E_i on oodatav sagedus

Hii-ruut-jaotuse kriitilised väärtused:

- Hüpoteeside testimisel kasutatakse hii-ruut-jaotuse kriitilisi väärtusi, et määrata, kas teststatistiku väärtus on statistiliselt oluline.

- Kriitilised väärtused leitakse hii-ruut-jaotuse tabelitest või statistikatarkvara abil, võttes arvesse vabadusastmete arvu ja valitud olulisuse nivoo (α).
- Olulisuse nivoo (α) on tõenäosus, et nullhüpotees lükatakse tagasi, kui see tegelikult on õige.
- Kui teststatistiku väärtus ületab kriitilise väärtuse, siis nullhüpotees lükatakse tagasi.

Peatükk 6

Statistilised hinnangud

6.1 Hinnangud

6.1.1 Punkthinnang

Üldkogumi tunnuseid analüüsides huvitavad uurijat tavaliselt järgnevad probleemid:

- üldkogumi tunnuse jaotusseadus;
- selle jaotuse arvkarakteristikute ehk parameetrite arvulised väärtused (hinnangud);
- otsustused üldkogumi parameetrite kohta (hüpoteesid).

Statistilises analüüsis vaadatakse sageli esimesena, kas uuritavate andmete jaotus meenutab normaaljaotust, mis on teatud tüüpi jaotus, kus enamik väärtusi koondub keskmise ümber ja vähem levinud väärtused jäävad äärtesse. Mõnikord eeldatakse, et andmed järgivad normaaljaotust ilma seda eraldi kontrollimata, sest varasemad uuringud või teoreetilised põhjendused viitavad sellele. Kui aga selgub, et andmed ei järgi normaaljaotust, tuleks kasutada erilisi, niinimetatud jaotusvabu ehk mitteparameetrilisi meetodeid. Need meetodid võimaldavad teha järeldusi andmekogumi kohta ka siis, kui eeldatav jaotus ei ole normaalne, tagades seeläbi täpsemad tulemused.

Üldkogumi parameetri punkthinnang on number, mille saame valimi andmetest, kasutades kindlat arvutusviisi. See number annab meile aimu, mis võiks olla mingi omaduse tegelik väärtus kogu rühmas, mida me uurime. Punkthinnangut näidatakse sageli sama sümboliga, mida kasutatakse selle omaduse kirjeldamiseks kogu grupis, lisades sellele väikese erimärgi nagu katuse või tärni. Kuna valim võetakse juhuslikult, sõltub punkthinnangu väärtus juhusest ja see võib iga kord erineda, kui võtame uue valimi. Seetõttu käsitleme punkthinnangut kui juhuslikku suurust, millel on oma jaotus ja parameetrid nagu keskmine väärtus või hajuvus. See tähendab, et punkthinnangu abil püüame mõista kogu grupi omadusi, lähtudes vaid osast sellest grupist.

Üldkogumi parameetri, näiteks keskväärtuse, hindamiseks erinevatest valimi andmetest võib kasutada mitmesuguseid arvutusreegleid. Üks lihtne viis keskväärtuse määramiseks on võtta valimi kõige väiksema ja suurima väärtuse keskmine, st. liita need kaks väärtust kokku ja jagada kahega. See meetod on eriti kasulik, kui valim

on väike ja soovitakse kiiresti saada ligikaudset hinnangut keskväärtusele. Kui aga valim on suur ja sisaldab äärmuslikke väärtusi, mis võivad tulemust moonutada, on mõistlik kasutada aritmeetilist keskmist, kust on välja jäetud kõige väiksem ja suurim väärtus. Selline lähenemine aitab vähendada üksikute ebatavaliste väärtuste mõju keskmisele, andes seeläbi usaldusväärsema ülevaate üldkogumi keskväärtusest.

Punkthinnangute arvutuseeskirjade analüüsiga tegelevat matemaatilise statistika haru nimetatakse **hinnangute teooriaks**.

Nihketa hinnang tähendab, et kui me arvutame keskmise väärtuse kõikidest võimalikest hinnangutest, mida me saame valimi andmete põhjal, siis see keskmine peaks langema kokku tegeliku väärtusega, mida me üritame hinnata. See on nagu täpselt sihtpunkti tabamine. Kui hinnang on nihkega, siis see tähendab, et meie hinnangud kipuvad süstemaatiliselt olema kas liiga kõrged või liiga madalad võrreldes tegeliku väärtusega.

Hinnangu headust mõjutab oluliselt ka selle hajuvus, mida näitab standardhälve – see ütleb meile, kui palju hinnangud tegelikust väärtusest keskmiselt kõiguvad. Mida väiksem on hajuvus, seda usaldusväärsem on hinnang. Hajuvus sõltub suuresti valimi suurusest: mida suurem on valim, seda väiksem on hajuvus. See tähendab, et hinnangu täpsuse suurendamiseks, näiteks kahekordistamiseks, peame valimi suurust suurendama neli korda. Teisisõnu, suurema valimi korral saame täpsemaid hinnanguid, mis aitab meil teha paremaid järeldusi üldkogumi kohta.

Kui kasutame sama valimit, võime saada erinevaid punkthinnanguid, olenevalt sellest, millist arvutusreeglit rakendame. Näiteks keskväärtuse hindamisel võib aritmeetiline keskmine anda kõige väiksema hajuvusega tulemuse, samas kui valimi minimaalse ja maksimaalse väärtuse keskmise kasutamine toob kaasa kõige suurema hajuvuse. Seega, kui valime hindamiseeskirja, peaksime otsima sellise, mis annab võimalikult väikese hajuvusega hinnangu.

6.1.1.1 Keskväärtus

Keskväärtus ehk keskmine on statistiline näitaja, mis annab aritmeetilise keskmise kõikidest väärtustest mõõtmisel. See võimaldab kokku võtta suure hulga andmeid üheks esinduslikuks näitajaks, mis näitab, kui suur on väärtuste kogusumma ja kui palju on neid väärtusi.

Keskväärtuse leidmiseks tuleb kõik andmed kokku liita ja seejärel jagada andmete arvu ehk elementide arvuga. Valem keskmise leidmiseks: Keskväärtus = (x1 + x2 + ... + xn) / n, kus x1, x2, ..., xn on andmed ja n on andmete arv.

Näide R-is:

```
andmed <- rnorm(1000, 10, 2)
# Arvutame keskmise
mean(andmed)
```

```
[1] 10
```

Näide Excelis:

```
Lahtrisse B1 sisesta järgnev valem: `=AVERAGE(andmeplokk)`
```

Excel arvutab nüüd nende andmete keskmise, näidates tulemust lahtris B1.

6.1.1.2 Dispersioon

Dispersioon on statistiline mõiste, mis kirjeldab, kui palju hajuvad andmed keskmisest väärtusest. Dispersioon annab aimu sellest, kui palju erinevad üksikandmed keskmisest ning kui suur on nende vahemik. Mida suurem on dispersioon, seda ebaühtlasemalt on andmed hajunud keskmise ümber.

R näide:

```r
# Leiame andmehulga dispersiooni funktsiooni var()
var(andmed)
```

```
[1] 3,93
```

Näide Excelis:

```
`=VAR(andmeplokk)`
```

Mida väiksem on dispersioon, seda kitsam on andmehulga hajuvus. Dispersiooni arvutamise valem võib Excelis varieeruda sõltuvalt sellest, kas soovite arvutada dispersiooni populatsiooni (VAR.P) või valimi (VAR.S) põhjal. R-is kasutatakse funktsiooni `var()` populatsiooni hinnangu jaoks ning `var(x, y = 1)` valimi hinnangu jaoks, kus `x` on andmehulk ja `y` on valimi põhjal arvutatava dispersiooni korrigeerimise parameetrit väljendav arv.

6.1.1.3 Dispersiooni viga

Dispersiooni viga (inglise keeles dispersion error) on statistiline mõiste, mis mõõdab hajuvust või ebatäpsust andmete hulgas. See näitab, kui palju erinevad individuaalsed andmed keskmisest ning omab olulist tähtsust hinnangute usaldusväärsuse hindamisel.

Dispersiooni viga arvutatakse tavaliselt järgmise valemiga:

$$SE = \sqrt{\frac{s^2}{n}}$$

kus SE on dispersiooni viga, s^2 on proovikogumi dispersioon (standardhälve ruudus) ja n on proovikogumi suurus.

Näide R-is:

```r
# Arvutame keskmise ja standardhälbe
keskmine <- mean(andmed)
dispersioon <- sd(andmed)
```

```r
# Arvutame dispersiooni vea
sqrt(dispersioon^2 / length(andmed))
```

```
[1] 0,0627
```

Näide Excelis:

```
=SQRT(STDEV(A1:A10)^2/COUNT(A1:A10)) # dispersioonivea arvutamine
```

Nii R-is kui ka Excelis arvutatud dispersiooni viga võimaldab hinnata, kui hästi keskmine esindab tervet andmestikku ning milline võiks olla keskmise usaldusväärsus.

6.1.1.4 Standardhälve

Standardhälve (või standardhälve, lühendina SD) on statistiline mõõde, mis näitab, kui palju erinevad andmed üldisest keskmisest. Mida suurem on standardhälve, seda laiem on andmehulkude ja keskmise hajuvus. Standardhälve on oluline näitaja, kui soovitakse analüüsida andmete levikut ning võrrelda erinevaid andmehulki.

Näide R-is:

```r
# Arvutame standardhälbe
sd(andmed)
```

```
[1] 1,98
```

Näide Excelis:

```
=STDEV(A1:A10)
```

Näiteks standardhälve arvutamine saab olla vajalik, kui soovitakse teada saada, kui palju erinevad metsanduses mõõdetud puude kõrgused üksteisest ning millised puud on silmatorkavalt kõrgemad või madalamad.

6.1.1.5 Standardhälbe viga

Standardhälve viga on statistiline mõõde, mis näitab, kui palju keskmisest erinevad kogumandmete proovide keskmised tulemused. Standardhälbe viga arvutatakse kogu andmestiku standardhälbe ja proovide arvu ruutjuure suhtes ning seda väljendatakse tavaliselt protsentides.

Standardhälbe vea valem on järgmine:

$$SE = \frac{s}{\sqrt{n}}$$

Kus: - SE on standardhälbe viga - s on proovi standardhälve - n on proovide arv

Näide R-is:

```r
# Arvuta standardhälve
sd_andmed <- sd(andmed)
```

```
# Arvuta standardhälbe viga
sd_andmed / sqrt(length(andmed))
```

```
[1] 0,0627
```

Näide Excelis:

- Arvuta standardhälbe viga: `=STDEV.S(A1:A10)/SQRT(COUNT(A1:A10))`

Saadud tulemus annabki standardhälbe vea, mis aitab hinnata, kui täpselt proov keskmisega kogu populatsiooni keskmine kokku langeb. Mida väiksem on standardhälbe viga, seda täpsem on proovi keskmine hinnang kogu populatsiooni keskmisele.

6.1.1.6 Standardviga

Standardviga on statistiline mõiste, mis näitab, kui suur on keskmise arvutatud vaatlusväärtuse kõikumine tegeliku keskmise hinnangu ümber. Standardviga väljendatakse tavaliselt samades ühikutes nagu uuritav tunnus. Mida väiksem on standardviga, seda usaldusväärsemat hinnangut saab anda populatsiooni keskmisele.

Näide R-is: Eeldame, et meil on andmestik veekogude sügavuste kohta:

```
keskmine <- mean(andmed)
sd(andmed) / sqrt(length(andmed))
```

```
[1] 0,0627
```

Antud näites arvutatakse andmestiku standardiga funktsiooniga `sd` ning seejärel jagatakse see juur arvutatud andmete arvuga, et saada standardviga.

Näide Excelis:

```
=STDEV(A1:A10) / SQRT(COUNT(A1:A10))
```

Seejärel saab Excelis automaatselt arvutada standardvea vastavalt antud andmestikule.

6.1.1.7 Katsetäpsus

Katsetäpsus või eksperimentaaltäpsus kirjeldab seda, kui hästi korduvad mõõtmised või katsetulemused on omavahel kooskõlas. See mõiste aitab hinnata katse tulemuste usaldusväärsust ja täpsust. Kõrge katsetäpsuse puhul on korduvate katsete tulemused sarnased, samas kui madala katsetäpsuse korral võivad tulemused varieeruda suurel määral.

Näide R-is:

```
# Loome andmestiku

# Leiame keskmise ning standardhälbe
keskmine <- mean(andmed)
sd <- sd(andmed)
```

```
# Arvutame katsetäpsuse
sd / keskmine
```

```
[1] 0,198
```

Näide Excelis:

```
`=STDEV(A1:A10)/AVERAGE(A1:A10)` (katsetäpsus)
```

Sellisel juhul väljastavad R ja Excel katsetäpsuse väärtuse, mis aitab hinnata antud mõõtmiste korduvust ja usaldusväärsust.

6.1.1.8 Variatsioonikordaja

Variatsioonikordaja on statistiline mõõde, mis väljendab protsendina muutuvust numbriliste andmete hulgas. See võimaldab hinnata, kui suur on andmete hajuvus keskmisest väärtusest ning võrrelda erinevate muutujate hajuvust.

Variatsioonikordajat saab arvutada järgmise valemiga:

$$CV = \left(\frac{sd}{\bar{x}}\right) \times 100\%,$$

kus: - CV on variatsioonikordaja, - sd on standardhälve ehk standardne kõrvalekalle ja - $\bar{x}$ on aritmeetiline keskmine.

Näide R-is:

```
# Arvutame keskmise ja standardhälbe
keskmine <- mean(andmed)
sd <- sd(andmed)

# Arvutame variatsioonikordaja
(sd / keskmine) * 100
```

```
[1] 19,8
```

=STDEV(A1:A10)/AVERAGE(A1:A10) * 100

Loodetavasti aitas see selgitus ja näited mõista variatsioonikordaja mõistet ja selle arvutamist nii R-is kui ka Excelis.

6.1.1.9 Variatsioonikordaja viga

Variatsioonikordaja viga on statistiline mõiste, mis näitab suurust, millega proov valimina võetud andmed erinevad populatsiooni üldisest muutujast. See arvutatakse suhtelise veana, võttes arvesse andmete keskmist ja standardhälvet.

Variatsioonikordaja vigade arvutusvalemit saab defineerida järgnevalt:

$$CV = \frac{s}{\bar{x}} \times 100\%$$

kus CV - variatsioonikordaja viga, s - standardhälve ja $\bar{x}$ - keskmine.

Kui variatsioonikordaja viga on väike, siis võib järeldada, et andmed on stabiilsed ja usaldusväärsed. Kui see aga on suur, siis võib see tähendada, et andmed on dispersiooniga ebaühtlased.

Näide R-keskkonnas:

```
# Andmestik
andmed <- c(10, 15, 20, 25, 30)

# Keskmine ja standardhälve
keskmine <- mean(andmed)
sd <- sd(andmed)

# Variatsioonikordaja viga
sd / keskmine * 100
```

Excelis saab variatsioonikordaja vea lihtsalt arvutada järgmise valemi abil: `=STDEV(A2:E2)/A4*100`, kus A2:E2 on andmehulk ja A4 on keskmine.

Variatsioonikordaja viga aitab hinnata andmete usaldusväärsust ja stabiilsust ning on oluline tööriist statistiliste analüüside tegemisel.

6.2 Vahemikhinnangud

Üldkogumi parameetri vahemikhinnanguks nimetatakse valimi põhjal määratud vahemikku, kuhu see üldkogumi parameeter kuulub teatud (küllalt suure) tõenäosusega. Seda uurija poolt etteantavat tõenäosust nimetatakse usaldusnivooks ja tähistatakse traditsiooniliselt 1-α. Usaldusnivoo väärtuseks võetakse metsanduslikes uurimustes tavaliselt 0,95, kuid vastavalt uurija kaalutlustele võib selleks olla ka 0,90; 0,99 või mingi muu suur tõenäosus. Vastavalt usaldusnivoo etteantud väärtusele arvutatakse parameetri usalduspiirid, s.o kaks arvu, mille vahel asub üldkogumi parameeter tõenäosusega 1-α. Teades, et parameetri hinnang on juhuslik suurus, võib alumist usalduspiiri vaadelda kui parameetri hinnangu $\alpha/2$-kvantiili ja ülemist usalduspiiri kui parameetri hinnangu 1-$\alpha/2$-kvantiili.

6.2.1 Keskväärtuse usalduspiirid

Keskväärtuse usalduspiirid on statistiline mõiste, mis aitab hinnata keskmise väärtuse usaldusväärsust. Keskväärtus ehk tõenäosusega 1-α keskmisele vastav populatsiooni parameeter paikneb usalduspiiride vahemikus keskmise ise ja sellest suurema või väiksema arvu vahel.

Usalduspiirid aitavad mõõta seda, kui kindlalt võime väita, et hinnatud keskmine väärtus on populatsiooni tegelik keskmine väärtus. Usalduspiiride arvutamiseks kasutatakse tavaliselt keskväärtuse ja standardvea su-

het ning t-tabelist leitakse vastav kriitiline tõenäosus. Näiteks 95% usaldusvahemik tähendab, et oleme 95% kindlad, et populatsiooni keskmine jääb antud vahemiku sisse.

Keskväärtuse usalduspiirid on olulised, kui soovime teha järeldusi populatsiooni keskmise kohta, kuid ei saa seda täpselt kindlaks teha. Need aitavad meil hinnata tulemuste usaldusväärsust ja teha otsuseid põhinedes piisavalt kindlatel andmetel.

6.2.2 Dispersiooni usalduspiirid

Dispersiooni usalduspiirid on statistilised piirid, mis näitavad, kui palju võib populatsiooni dispersioon, standardhälve või varieerumine erineda, kui võtame arvesse valimi dispersiooni. Usalduspiirid annavad meile võimaluse teha järeldusi populatsiooni dispersiooni kohta, kui meil on vaid valimi andmed.

Dispersiooni usalduspiirid arvutatakse tavaliselt usaldusvahemikuna, mis näitab, milline on tõenäosus, et populatsiooni dispersioon jääb kindla vahemiku piiresse. Näiteks 95% usalduspiirid tähendavad, et 95% tõenäosusega langeb populatsiooni dispersioon selle vahemiku sisse.

Usalduspiiride arvutamine toimub tavaliselt tundmatu populatsiooni dispersiooni põhjal, kasutades valimi dispersiooni ja valimi suurust. Usalduspiiride laiust mõjutavad valimi suurus, usaldusvahemik ning populatsiooni dispersiooni enda suurus.

Dispersiooni usalduspiirid on olulised tööriistad statistikas, sest need aitavad meil teha järeldusi populatsiooni varieeruvuse kohta ning hinnata, kui usaldusväärsed on meie tulemused. Mõistes dispersiooni usalduspiiride kontseptsiooni, saavad metsandustudengid paremini analüüsida ja tõlgendada metsanduslike andmete varieeruvust ning teha ratsionaalseid otsuseid metsandusalastes uurimustes.

6.2.3 Binoomjaotuse usalduspiirid

Binoomjaotus on tõenäosusteoorias kasutatav jaotus, mis kirjeldab kahe võimaliku tulemusega eksperimentide tulemuste jaotust. Näiteks võib binoomjaotust kasutada olukorras, kus uuritakse mingi sündmuse (nt müntide viskamine, taimede ellujäämine) kordaminekut (edu) või ebaõnnestumist (ebaedu). Binoomjaotuse parameeter p tähistab tõenäosust, et sündmus toimub ühel katsekorral.

Usalduspiirid on hinnangulised piirid, mille sisaldamisel on kindel usaldustase, et parameetri tegelik väärtus jääb nende piiride vahele. Binoomjaotuse puhul tähendab see, et soovime leida piirid, mille vahel paikneb tõenäosus, et sündmus toimub ühel katsekorral.

Binoomjaotuse puhul saab leida 100(1-α)% usaldusvahemiku soovitud usaldustasemega α. Usalduspiirid leitakse tavaliselt normaaljaotuse abil, kasutades normaaljaotuse ligilähedast seost binoomjaotusega (kui n on piisavalt suur). Kui binoomjaotuse põhjal antakse 100(1-α)% usaldusvahemik, siis näeb see välja järgmine:

$$\hat{p} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

kus $\hat{p}$ on valimi proportsioon, $Z_{1-\frac{\alpha}{2}}$ on normaaljaotuse kvantiil vastava usaldustasemega ja n on valimi suurus.

Kokkuvõttes on binoomjaotuse parameetri p usalduspiirid hinnangulised piirid, mille vahel on kindel usaldustase, et parameetri tegelik väärtus paikneb. Need piirid aitavad teha järeldusi tõenäosuse kohta, et sündmus toimub ühel katsekorral.

6.2.4 Standardhälbe usalduspiirid

Standardhälbe usalduspiirid on statistiline mõiste, mis aitab hinnata andmeid või tulemusi ning nende usaldusväärtust. Standardhälve on mõõtühik, mis näitab, kui palju erinevad keskmised tulemused koondumisest keskmise ümber.

Standardhälbe usalduspiirid näitavad, kui kindlalt võib olla keskmise tulemuse usaldusväärsus teatud vahemikus. Enamasti kasutatakse 95% usalduspiire, mis tähendab, et 95% tõenäosusega jääb tulemus antud usalduspiiride vahemikku. Usalduspiirid arvutatakse keskmise tulemuse ja standardhälbe põhjal.

Standardhälbe usalduspiiride kasutamine aitab hinnata andmete usaldusväärsust ning teha järeldusi selle kohta, kui suures määras võivad andmed hajuda. See on oluline statistiline tööriist, mis aitab metsandustudengitel analüüsida andmeid ja teha tõenduspõhiseid otsuseid metsanduse valdkonnas.

Peatükk 7

Mitme tunnuse koosanalüüs

7.1 Lineaarne regressioon

Lineaarne regressioon on üks viis, kuidas uurida, kuidas kaks või enam asja on omavahel seotud. Selle abil saame aru saada, kuidas ühe tunnuse muutumine mõjutab teist. Näiteks, kui me teame, kuidas ühe tunnuse väärtus muutub, saame ennustada, mis võib juhtuda teise tunnusega.

Lineaarne regressioon lähtub eeldusest, et kahe asja vaheline suhe on sirgejooneline. See tähendab, et kui üks tunnus muutub, siis teine asi muutub kindlal viisil, mida saab joonistada sirgjoonena. Selle meetodi eesmärk on leida kõige sobivam sirgjoon, mis näitab, kuidas need kaks tunnust on omavahel seotud. Lineaarse regressioonimudeli võib esitada võrrandina:

$$Y = a + bX + \epsilon$$

Kus:

- Y on funktsioontunnus
- X on argumenttunnus
- a on lõikepunkt, mis näitab, kui suur on Y väärtus, kui X väärtus on o
- b on regresseerimiskordaja, mis näitab, kui palju muutub Y väärtus, kui X väärtus suureneb ühe ühiku võrra
- ϵ on veaväärtus, mis tähistab juhuslikke kõrvalekaldeid mudelist

Lineaarse regressiooni puhul kasutatakse mitmesuguseid meetodeid, et hinnata, kui hästi meie sirgjoon mudelina toimib. Üks selline meetod on vähim ruutude meetod. See tähendab, et me otsime sirgjoont, mis käitub nii, et meie ennustuste ja tegelikkuses mõõdetud väärtuste erinevused oleksid võimalikult väikesed. Lihtsamalt öeldes, me püüame leida sirgjoont, mis sobib kõige paremini meie andmetega, minimeerides vigu ennustustes.

Lineaarse regressiooni usaldusväärsed tulemused sõltuvad teatud eeldustest. Need eeldused aitavad tagada, et meie analüüs peegeldaks tegelikkust võimalikult täpselt. Siin on need eeldused lihtsustatult:

- **Lineaarsus.** Eeldatakse, et kahe tunnuse vaheline suhe on sirgjooneline. See tähendab, et kui üks tunnus muutub, siis teise tunnus vastus on alati sellega proportsioonis, sõltumata olukorrast.
- **Normaaljaotus.** Hälve ehk ennustuste ja tegelike tulemuste erinevused järgivad kindlat mustrit, mis vastab normaaljaotusele.
- **Homoskedastilisus.** Sõltuva tunnuse väärtuste hajuvus on ühtlane kogu sõltumatu tunnuse ulatuses, mis tähendab, et erinevused ei suurene ega vähene sõltumatu tunnuse väärtuse kasvades.
- **Autokorrelatsiooni puudumine.** Vaatluste vead on omavahel sõltumatud, st et ühe vaatluse viga ei mõjuta teise vaatluse viga.
- **Multikollineaarsuse puudumine.** Analüüsitavad tunnused on sõltumatud, st ühe tunnuse väärtus ei sõltu teise tunnuse väärtusest ega ole nendega liiga tihedalt seotud.

Need eeldused aitavad tagada, et lineaarne regressioonimudel oleks usaldusväärne ja et selle abil tehtud järeldused peegeldaksid tegelikke seoseid.

7.1.1 Lineaarne regressioon R keskkonnas

```
# Näide R keskkonnas
# Laadime peatükis kasutatavad paketid
library(dplyr)
library(ggplot2)
library(readxl)
library(car)
library(ggpmisc)
```

Loeme andmed failist naited.xlsx töölehelt “prt_andmed” ja salvestame selle tabelisse:

```
# Näide R keskkonnas
andmed <- read_excel("data/naited.xlsx","prt_andmed") %>% filter(h > 0)
```

Teeme lineaarse regressiooni tunnuste D ja H vahel:

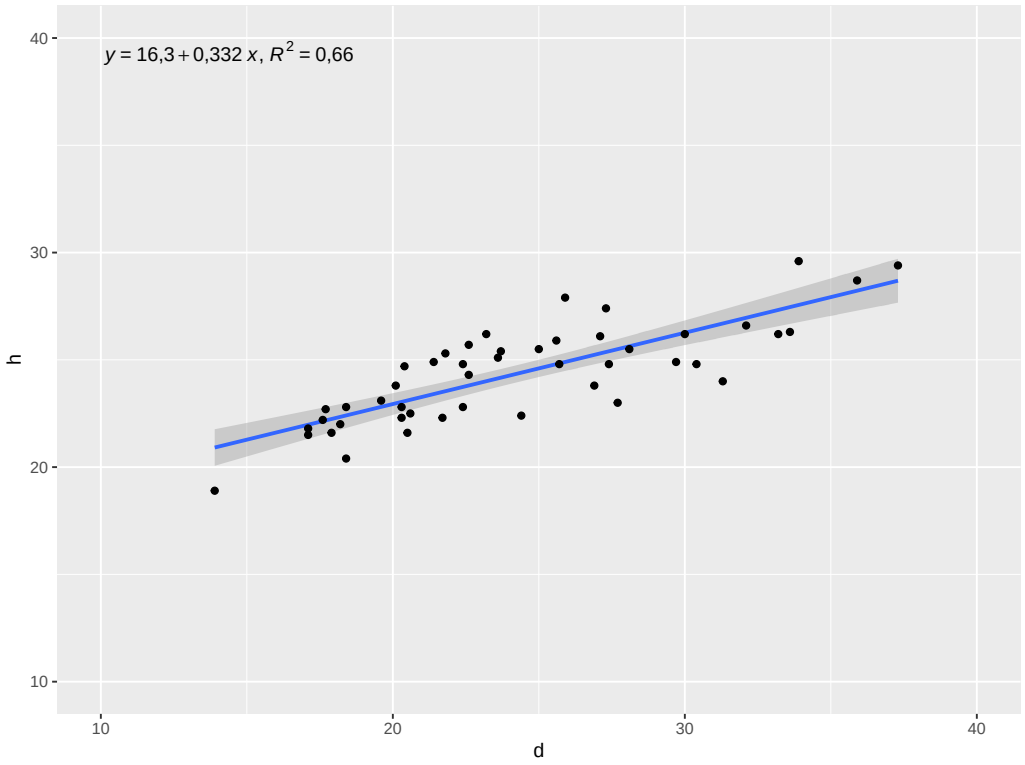
```
# Näide R keskkonnas
print(lm_model <- lm(h ~ d, data = andmed))
```

```
Call:
lm(formula = h ~ d, data = andmed)

Coefficients:
(Intercept)            d
     16,295        0,332
```

Järgnevalt visualiseerime nii andmeid kui ka mudelit funktsiooni `ggplot()` abil.

```
# Näide R keskkonnas
ggplot(andmed, aes(x = d, y = h)) +
  stat_poly_line() +
  stat_poly_eq(use_label(c("eq", "R2"))) +
  geom_point() +
  xlim(c(10, 40)) + ylim(c(10, 40))
```

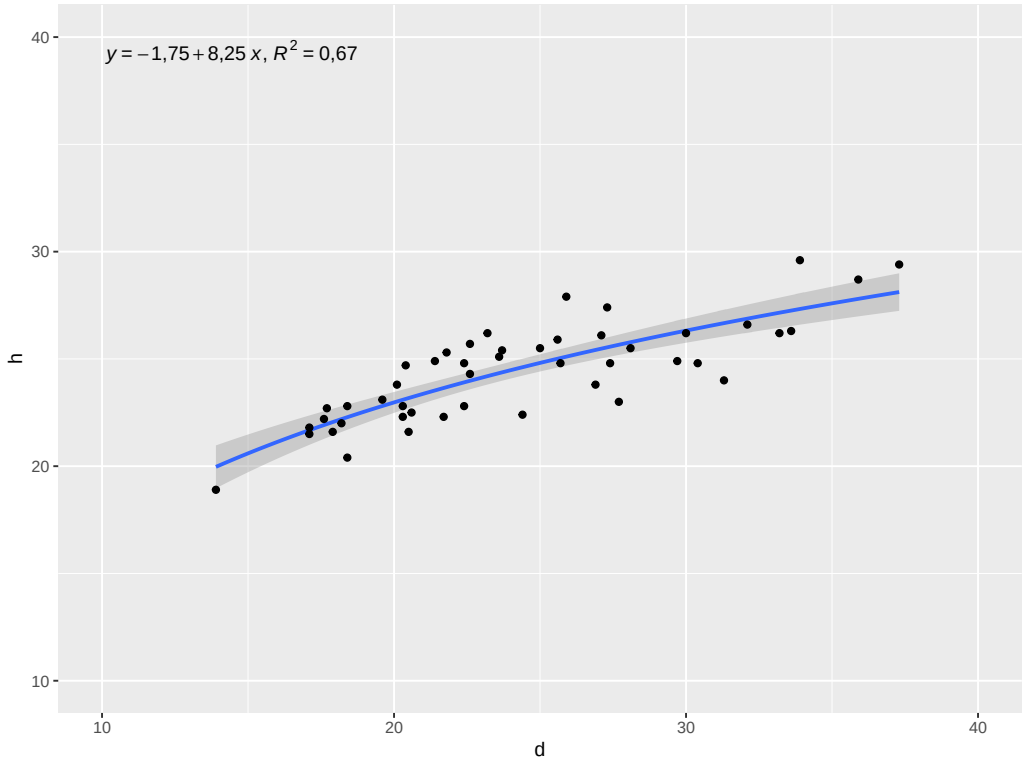


Joonis 7.1. Lineaarne regressioon

Järgnevalt käsu selgitus:

- `ggplot(andmed, aes(x = d, y = h))`: See algatab ggplot-objekti, määrates andmestiku (data) ja esteetilised vastavused. See tähendab, et andmestiku veerg `d` vastendatakse x-teljega ja veerg `h` y-teljega.
- `stat_poly_line()`: See lisab graafikule polünoomilise regressioonijoone. Vaikimisi kasutab see lineaarset mudelit (astmega 1), kuid seda saab muuta, kasutades argumenti `formula`. Näiteks `stat_poly_line(formula = y ~ x^2)` lisaks graafikule ruutvõrrandi.
- `stat_poly_eq(use_label(c("eq", "R2")))`: See lisab graafikule polünoomilise regressioonivõrrandi ja R-ruudu väärtuse. `use_label()` funktsiooniga saab määrata, milliseid tunnuseid kuvatakse. Antud juhul kuvatakse võrrand (`eq`) ja R-ruut (`R2`).
- `geom_point()`: See lisab graafikule punktid, mis tähistavad andmepunkte.
- `xlim(c(10, 40))`: See seab x-telje piirid vahemikku 10 kuni 40.
- `ylim(c(10, 40))`: See seab y-telje piirid vahemikku 10 kuni 40.

```
# Näide R keskkonnas
ggplot(andmed, aes(x = d, y = h)) +
  stat_poly_line(formula = y ~ log(x)) +
  stat_poly_eq(formula = y ~ log(x), use_label(c("eq", "R2"))) +
  geom_point() +
  xlim(c(10, 40)) + ylim(c(10, 40))
```



Joonis 7.2. Lineaarne regressioon argumenttunnuse logistilise teisendusega

R-is on võimalik lihtalt lisada mudelisse täiendavad tunnuseid, mis arvutakse otse mudeli arvutamise käigus ning seetõttu ei ole vajalik eraldi eelnevalt täiendavat tunnust välja arvutada. Funktsiooniga `I()` saab luua uusi tunnuseid olemasolevate tunnuste põhjal. Uus tunnus arvutatakse iga andmerea jaoks mudeli arvutamise käigus.

Järgnevas näites on muudetud mudelit: `lm(h ~ log(d)+I(pl == "MA"), data = andmed)`

- **`log(d)`**: Funktsioon `log()` arvutab `d` naturaallogaritmi, mis aitab lineariseerida seost diameetri ja kõrguse vahel.
- **`I(pl == "MA")`**: See on teine sõltumatu tunnus, mis luuakse jooksvalt funktsiooni `I()` abil. See on loogiline tunnus, mis on `TRUE`, kui tunnus `pl` väärtus on "MA" , ja `FALSE` vastasel juhul. Sisuliselt lisab see mudelisse teavet selle kohta, kas puu on mänd. Arvutustes kasutatakse antud tunnust vastavalt 1 või 0 väärtusena.

```r
# Näide R keskkonnas
# täiendatud argumenttunnustega regressioon
summary(lm_model_pl <- lm(h ~ log(d)+I(pl == "MA"), data = andmed))
```

```
Call:
lm(formula = h ~ log(d) + I(pl == "MA"), data = andmed)

Residuals:
   Min     1Q Median     3Q    Max
-2,169 -0,857  0,048  0,665  3,421

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         -3,583      2,725   -1,31    0,195
log(d)               8,924      0,872   10,24  3,2e-13 ***
I(pl == "MA")TRUE   -0,978      0,430   -2,27    0,028 *
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Residual standard error: 1,27 on 44 degrees of freedom
Multiple R-squared:  0,707, Adjusted R-squared:  0,694
F-statistic: 53,2 on 2 and 44 DF,  p-value: 1,81e-12
```

Antud juhul on näha, et mänd on teistest statistiliselt oluliselt erinev.

7.1.2 Lineaarne regressioon MS Exceli keskkonnas

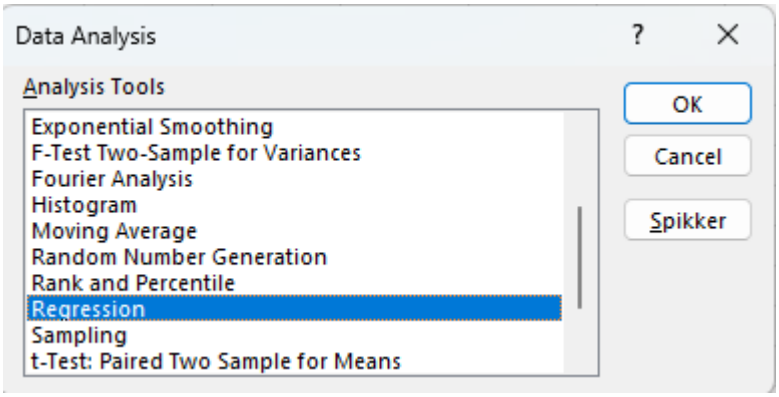
MS Exceli keskkonnas saab lineaarse regressiooni parameetreid arvutada mitmel viisil.

- Kui argumenttunnuseid on vaid üks, siis saab kasutada funktsioone `intercept()` ja `slope()`.
- Kui argumenttunnuseid on mitu (või ka ainult üks), siis saab kasutada funktsiooni `linest()`, mis annab tulemuseks juba ka täpsemad mudeli parameetrite hinnangud.
- Ühe ja mitme argumenttunnuse puhul saab kasutada *Analysis Toolpak* protseduuri *Regression*.

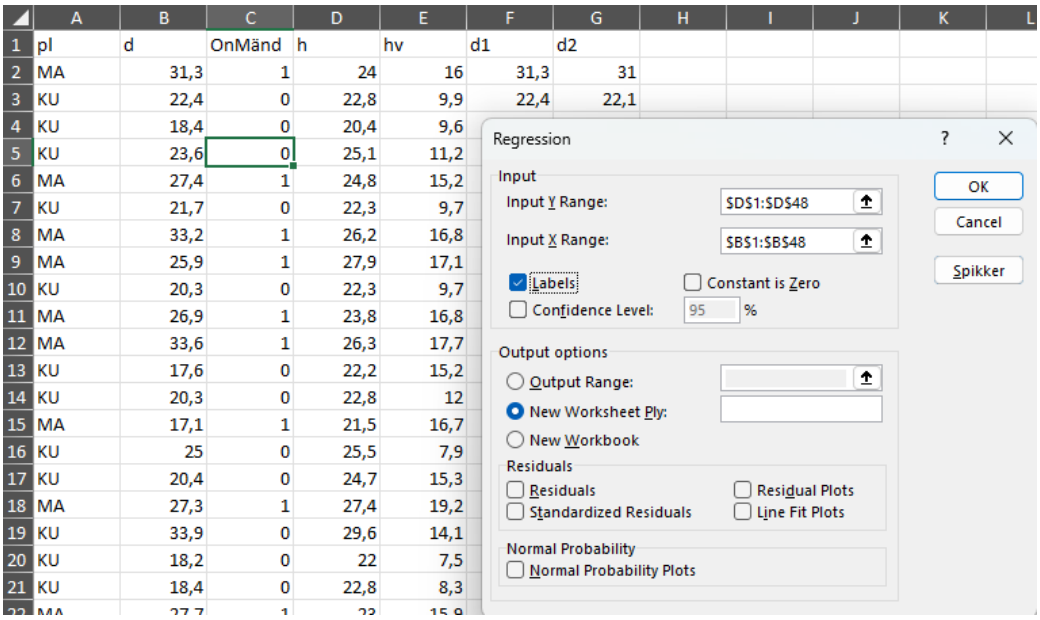
Avame faili naited.xlsx töölehel “linreg”, kus on samad andmed nagu “prt_andmed”, aga sellel on juba välja filtreeritud ainult mudelpuude andmed. MS Exceli keskkonnas on üldiselt vajalik filtreeritud andmed viia eraldi tabelisse, et regressiooni käigus muid andmeid ei kasutataks.

Analysis Toolpak abil regressioonanalüüsi tegemiseks tuleb menüüs “Andmed” valida “Data Analysis”, mis järel avaneb analüüsivahendite aken.

Kordame samasugust analüüsi ka MS Exceli keskkonnas nagu R-is, arvutame diameetrist naturaallogaritmi funktsiooniga `ln()` ning kasutame `OnMänd` tunnust.

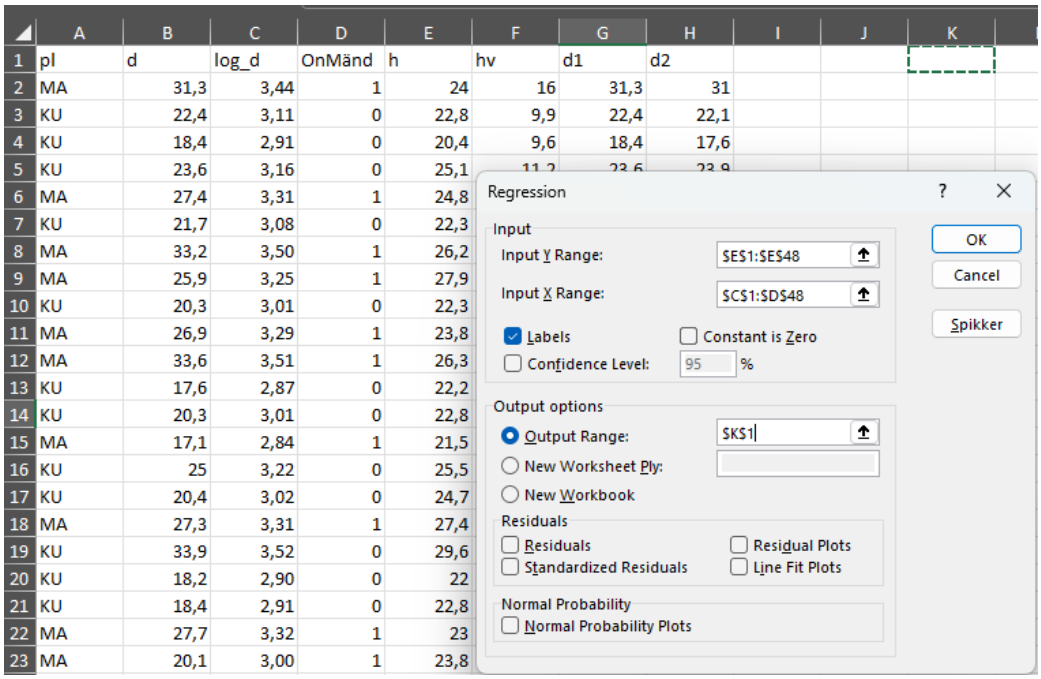


Joonis 7.3. Data Analysis aken MS Excelis.



	A	B	C	D	E	F	G
1	pl	d	OnMänd	h	hv	d1	d2
2	MA	31,3	1	24	16	31,3	31
3	KU	22,4	0	22,8	9,9	22,4	22,1
4	KU	18,4	0	20,4	9,6		
5	KU	23,6	0	25,1	11,2		
6	MA	27,4	1	24,8	15,2		
7	KU	21,7	0	22,3	9,7		
8	MA	33,2	1	26,2	16,8		
9	MA	25,9	1	27,9	17,1		
10	KU	20,3	0	22,3	9,7		
11	MA	26,9	1	23,8	16,8		
12	MA	33,6	1	26,3	17,7		
13	KU	17,6	0	22,2	15,2		
14	KU	20,3	0	22,8	12		
15	MA	17,1	1	21,5	16,7		
16	KU	25	0	25,5	7,9		
17	KU	20,4	0	24,7	15,3		
18	MA	27,3	1	27,4	19,2		
19	KU	33,9	0	29,6	14,1		
20	KU	18,2	0	22	7,5		
21	KU	18,4	0	22,8	8,3		

Joonis 7.4. Regressioonanalüüsi aken MS Excelis.



	A	B	C	D	E	F	G	H
1	pl	d	log_d	OnMänd	h	hv	d1	d2
2	MA	31,3	3,44	1	24	16	31,3	31
3	KU	22,4	3,11	0	22,8	9,9	22,4	22,1
4	KU	18,4	2,91	0	20,4	9,6	18,4	17,6
5	KU	23,6	3,16	0	25,1	11,2	23,6	23,9
6	MA	27,4	3,31	1	24,8			
7	KU	21,7	3,08	0	22,3			
8	MA	33,2	3,50	1	26,2			
9	MA	25,9	3,25	1	27,9			
10	KU	20,3	3,01	0	22,3			
11	MA	26,9	3,29	1	23,8			
12	MA	33,6	3,51	1	26,3			
13	KU	17,6	2,87	0	22,2			
14	KU	20,3	3,01	0	22,8			
15	MA	17,1	2,84	1	21,5			
16	KU	25	3,22	0	25,5			
17	KU	20,4	3,02	0	24,7			
18	MA	27,3	3,31	1	27,4			
19	KU	33,9	3,52	0	29,6			
20	KU	18,2	2,90	0	22			
21	KU	18,4	2,91	0	22,8			
22	MA	27,7	3,32	1	23			
23	MA	20,1	3,00	1	23,8			

Joonis 7.5. Regressioonanalüüs kahe tunnuse alusel.

Table: Lineaarse regressiooni tulemus MS Excelis.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,841
R Square	0,707
Adjusted R Square	0,694
Standard Error	1,268
Observations	47

ANOVA

	df	**SS**	**MS**	**F**	**Significance F**
Regression	2	171,139	85,570	53,191	0,000
Residual	44	70,783	1,609		
Total	46	241,922			

	Coefficients	**Standard Error**	**t Stat**	**P-value**
Intercept	-3,5829	2,7247	-1,3150	0,1953
log_d	8,9235	0,8718	10,2358	0,0000
OnMänd	-0,9776	0,4300	-2,2737	0,0279

7.1.3 Regressioonanalüüsi tulemuste tõlgendamine

Lineaarse regressioonimudeli kokkuvõte sisaldab mitmeid olulisi statistilisi näitajaid, mis aitavad hinnata mudeli sobivust andmetele.

Determinatsioonikordaja, tuntud ka kui R^2 (R-ruut), on statistiline mõõdik, mida kasutatakse lineaarse regressiooni kontekstis, et hinnata, kui suures ulatuses argumenttunnuse variatsioon on seletatav sõltumatute tunnuste abil mudelis. R^2 väärtus varieerub vahemikus 0 kuni 1, kus 0 tähendab, et mudel ei seleta argumenttunnuse variatsiooni üldse, samas kui 1 tähendab, et mudel seletab argumenttunnuse variatsiooni täielikult.

R^2 arvutamiseks kasutatakse järgmist valemit:

$$R^2 = 1 - \frac{\mathrm{SS}_{\mathrm{res}}}{\mathrm{SS}_{\mathrm{tot}}}$$

kus:

- SS_{res} on jääkide ruutude summa (residual sum of squares), mis mõõdab mudeli poolt seletamata variatsiooni.
- SS_{tot} on kogu variatsiooni ruutude summa (total sum of squares), mis mõõdab argumenttunnuse variatsiooni võrreldes tema keskmisega.

Kõrge R^2 väärtus (lähedane 1-le) näitab, et mudeli poolt seletatav variatsioon on suur, mis tähendab, et mudel sobib andmetele hästi. See viitab sellele, et suur osa argumenttunnuse variatsioonist on seletatav mudeli sõltumatute tunnustega.

Madal R^2 väärtus (lähedane 0-le) näitab, et mudel seletab argumenttunnuse variatsioonist väikese osa, mis tähendab, et mudeli sobivus andmetele on kehv. See võib viidata sellele, et mudelis kasutatavad sõltumatud tunnused ei ole sõltuva tunnuse variatsiooni seletamisel tõhusad.

Oluline on mõista, et kuigi R^2 on kasulik mõõdik mudeli sobivuse hindamisel, ei tähenda kõrge R^2 automaatselt, et mudel on õige või et sõltumatud tunnusd põhjustavad argumenttunnuse käitumist. Samuti ei pruugi madal R^2 alati tähendada mudeli ebatõhusust, eriti keerukamate või mittelineaarsete suhete puhul. Lisaks tuleks mudeli hindamisel arvesse võtta ka muid tegureid, nagu mudeli tõlgendatavus, sõltumatute tunnuste olulisus ja korrigeeritud R^2.

Lineaarse regressioonimudeli **nihutamata hinnang**, tuntud ka kui korrigeeritud R^2 (inglise keeles "adjusted R^2"), on statistiline mõõdik, mida kasutatakse mudeli seletusvõime hindamiseks, võttes arvesse mudeli sõltumatute tunnuste arvu. Erinevalt lihtsast R^2-st, mis mõõdab, kui suur osa argumenttunnuse variatsioonist on seletatav mudeli abil, arvestab korrigeeritud R^2 mudeli keerukust, arvestab mudeli liigsete sõltumatute tunnuste lisamise eest. See on oluline, kuna mudelisse liiga palju tunnusid lisades võib R^2 väärtus eksitavalt tõusta, viidates paremale mudeli sobivusele, kuigi tegelikult võivad lisatud tunnusd olla statistiliselt mitteolulised.

Korrigeeritud R^2 arvutatakse järgmise valemi järgi:

$$R^2_{adj} = 1 - \left(\frac{(1 - R^2)(n - 1)}{n - k - 1} \right)$$

kus:

- R^2 on mudeli lihtne determinatsioonikoefitsient,
- n on vaatluste arv,
- k on sõltumatute tunnuste arv mudelis.

Valemist nähtub, et R^2_{adj} väheneb, kui lisatakse tunnusid, mis ei paranda mudeli seletusvõimet proportsionaalselt nende arvuga. Seega, erinevalt lihtsast R^2-st, pakub korrigeeritud R^2 ausamat hinnangut mudeli sobivusele, eriti kui võrrelda mudeleid, mis sisaldavad erinevat arvu sõltumatuid tunnusid. Korrigeeritud R^2

on eriti kasulik mudelite võrdlemisel ja mudeli valimisel, aidates vältida ülemudelitamist ja eelistades mudelit, mis annab parima tasakaalu seletusvõime ja keerukuse vahel.

- P-väärtused näitavad, kas argumenttunnus on statistiliselt oluline mudelis. Kui P-väärtus on väiksem kui 0,05, siis võib järeldada, et argumenttunnus mõjutab oluliselt sõltuvat tunnust.
- Regressioonikordaja väärtused annavad teavet sõltumatute tunnuste mõju kohta sõltuvale tunnusle.

Kokkuvõttes aitab väljund mõista, kuidas mudel sobitub andmetega ja millist infot see pakub sõltuvate ja sõltumatute tunnuste seoste kohta.

7.1.4 Homoskedastilisuse testimine

Statistilise regressioonimudeli kontekstis viitab homoskedastiivsus olukorrale, kus jääkliikmete (ehk mudeli veatermini) dispersioon on konstantne sõltumata seletavate muutujate väärtustest. See tähendab, et mudeli ennustusvead on ühtlase hajuvusega kogu vaadeldava andmevahemiku ulatuses, ega ole süstemaatiliselt seotud ennustatud väärtustega. Homoskedastiivsus on kriitiline eeldus mitmetes statistilistes meetodites, eriti vähimruutude meetodil baseeruva lineaarse regressiooni puhul. Selle eelduse rikkumine, tuntud kui heteroskedastiivsus, võib moonutada mudeli parameetrite hinnanguid, viies ebatäpsete usaldusvahemike ja potentsiaalselt vigaste järeldusteni hüpoteeside testimisel.

On mitmeid viise, kuidas testida, kas andmetes esineb homoskedastiivsust:

1. **Visuaalne analüüs:** Lihtsaim viis on vaadata jääkliikmete graafikut seletavate muutujate või ennustatud väärtuste suhtes. Kui jääkliikmed on ühtlaselt hajunud nulljoone ümber, on homoskedastiivsus tõenäoline. Kui tekib mingi muster (näiteks koonusekujuline hajumine), on tõenäoline heteroskedastiivsus.
2. **Formaliseeritud testid:** On olemas mitmed statistilised testid, näiteks Breusch-Pagani test ja White'i test, mis testivad heteroskedastiivsust formaalselt. Need testid kontrollivad, kas jääkliikmete dispersioon on seotud seletavate muutujatega.
3. **Robustsed meetodid:** Kui andmetes esineb heteroskedastiivsust, võib kasutada robustseid regressioonimeetodeid, mis on vähem tundlikud heteroskedastiivsuse suhtes.

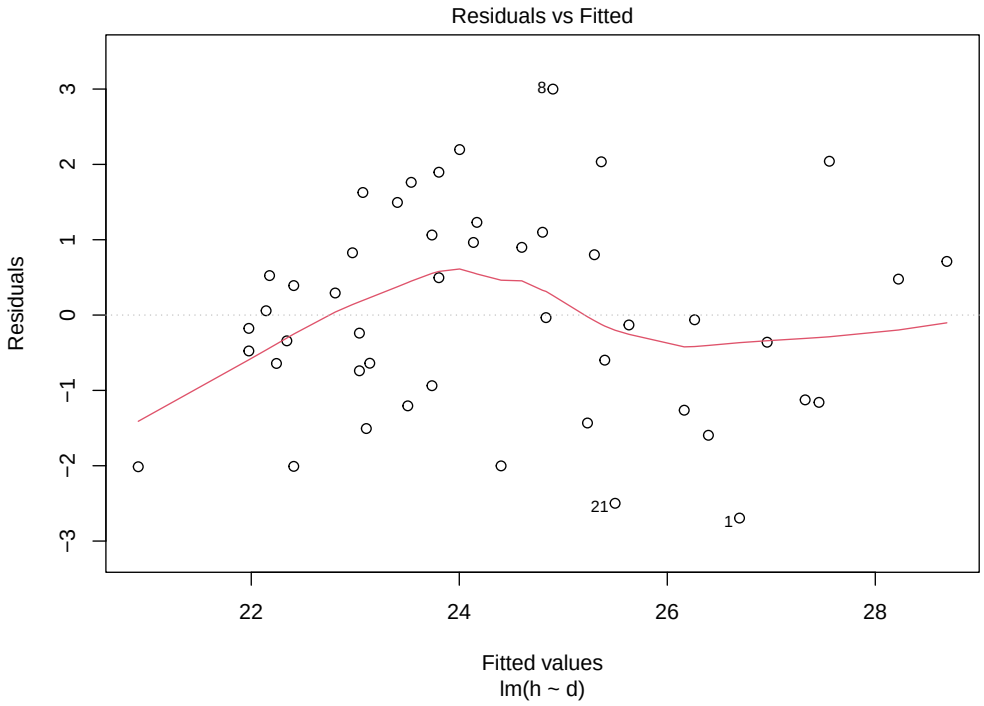
```r
# Näide R keskkonnas
plot(lm_model, which=1, main = "")
```

Lineaarse mudeli homoskedastilisuse testimiseks saab kasutada Breush-Pagan või White'i testi. Need testid hindavad, kas jääkide varieeruvus on seotud sõltumatute tunnustega ning kas see varieeruvus on ühtlane.

```r
# Näide R keskkonnas
lmtest::bptest(lm_model)
```

```
    studentized Breusch-Pagan test

data:  lm_model
```



Joonis 7.6. Jäägid vs prognoositud väärtus

```
BP = 0,9, df = 1, p-value = 0,4
```

Paketi `lmtest` funktsioon `bptest()` testib regressioonimudeli jääkide heteroskedastsuse olemasolu. See tähendab, et jääkide varieeruvus ei ole konstantne üle kõigi sõltumatute tunnuste väärtuste ning võib tekitada probleeme statistiliste hüpoteeside testimisel.

Testi tulemusi võib lugeda järgmiselt:

- Nullhüpotees (H0): jääkide varieeruvus on konstantne (st heteroskedastsust ei esine).
- Alternatiivhüpotees (H1): jääkide varieeruvus ei ole konstantne (st heteroskedastsus esineb).

Testist saadav p-väärtus aitab hinnata, kas on piisavalt tõendeid, et lükata ümber nullhüpotees heteroskedastsuse puudumise kohta. Väike p-väärtus (<0,05) viitab sellele, et saame nullhüpoteesi ümber lükata ja järeldada, et heteroskedastsust esineb.

7.1.5 Jääkide normaalsuse test

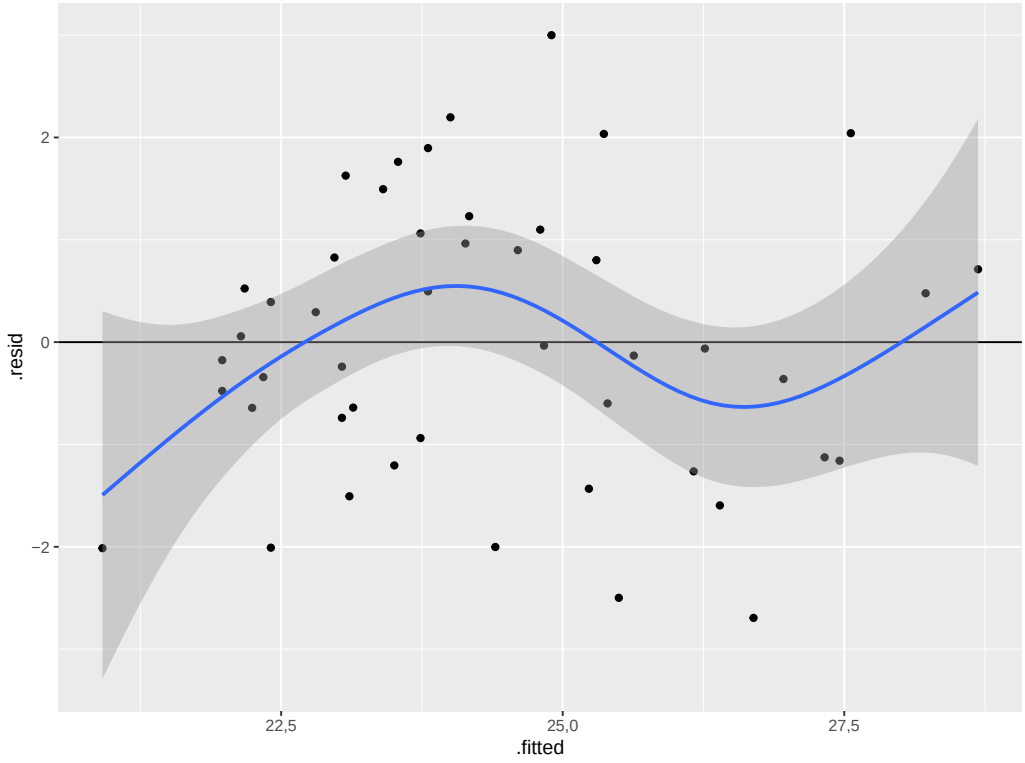
Üks oluline eeldus, mida lineaarses regressioonis kasutatakse, on see, et mudeli jäägid vastaksid normaaljaotusele. Normaaljaotus tähendab, et andmete jaotus on sümmeetriline ning keskmise jaotusega. Kui jäägid on normaaljaotusele lähedased, siis on regressioonimudeli ennustused usaldusväärsemad ja täpsemad. Paljud statistilised testid eeldavad, et andmed vastavad normaaljaotusele. Kui jäägid ei vasta normaaljaotusele, võivad

testide tulemused olla ebatäpsed või valed. Normaaljaotuse eeldustele mitte vastamine võib viidata sellele, et mudel ei sobi antud andmetele hästi ning on vajalik leida alternatiivseid mudeleid.

Normaliseeritud jääkide graafik on üks viis, kuidas hinnata, kas jäägid vastavad normaaljaotusele. See graafik aitab visualiseerida, kas jäägid on juhuslikult hajutatud ümber nulli ja kas nende jaotus on ligilähedane normaaljaotusele.

```
# Näide R keskkonnas

ggplot(lm_model, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0)+
  geom_smooth(method = "gam", formula = y ~ s(x))  # Lisab sujuva joone GAM mudeli abil
```



Koodi selgitus:

`ggplot(lm_model, aes(x = .fitted, y = .resid))`

- See käsk loob aluse graafikule ggplot2 paketiga.
- `lm_model` määrab andmeallika, milleks on lineaarne mudel.
- `aes(x = .fitted, y = .resid)` määrab x-teljele mudeli sobitatud väärtused (.fitted) ja y-teljele jääkväärtused (.resid).

`geom_point()`

- See lisab graafikule punktid, mis esindavad iga andmepunkti jääki ja sobitatud väärtust.

`geom_hline(yintercept = 0)`

- See lisab horisontaalse joone y-telje väärtusel 0. See joon aitab visuaalselt hinnata jääkide jaotust - ideaalses mudelis peaksid punktid jaotuma ühtlaselt joone ümber.

`geom_smooth(method = "gam", formula = y ~ s(x))`

- See lisab graafikule sujuva joone, mis on arvutatud GAM mudeli (Generalized Additive Model) abil.
- `method = "gam"` määrab joone arvutamise meetodiks GAM.
- `formula = y ~ s(x)` määrab mudeli valemi, kus y (jääkväärtused) on seotud x-i (sobitatud väärtused) sujuva funktsiooniga s(x).

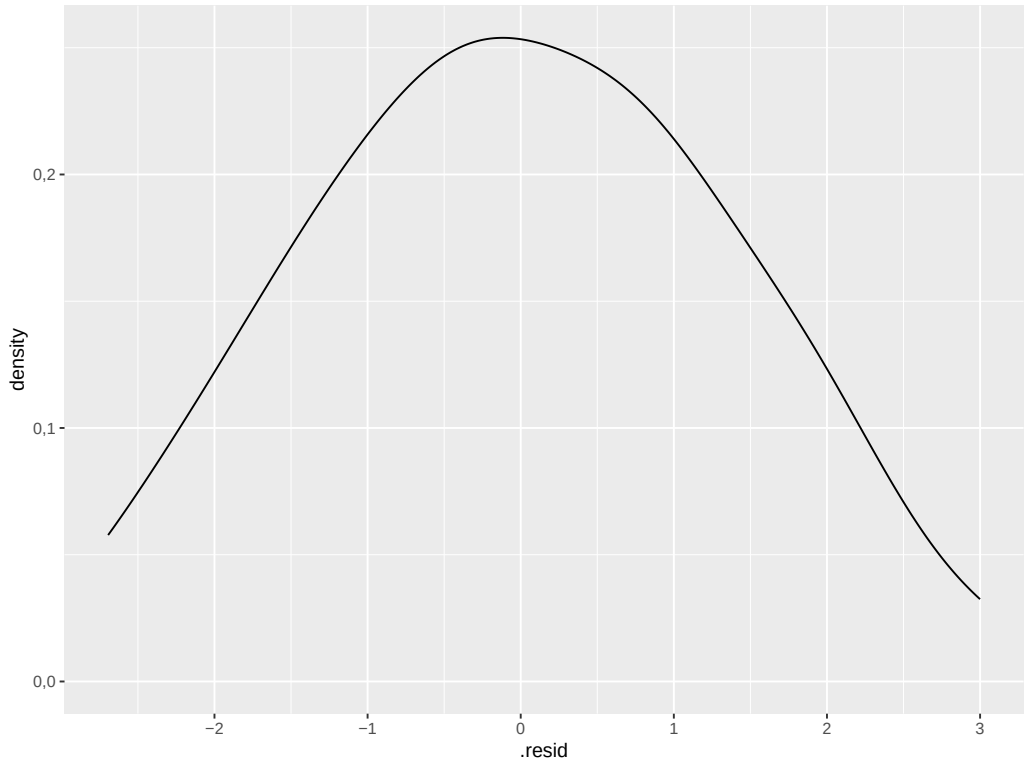
Kokkuvõtteks:

Kood genereerib diagnostilise graafiku, mis aitab hinnata lineaarse mudeli sobivust andmetega. Jääkide jaotus 0-joone ümber ja sujuva joone kuju annavad informatsiooni mudeli eelduste (nt homoskedastsus ja lineaarsus) täidetuse kohta.

Samuti saab vaadata jääkide jagunemist tihedusfunktsiooni abil koostatud graafikul.

```r
# Näide R keskkonnas

ggplot(lm_model, aes(x = .resid)) +
  geom_density()
```



Joonis 7.7. Jääkide tihedusfunktsioon

Sellise protseduuri abil saad hinnata, kui hästi sobib loodud lineaarne regressioonimudel andmetele ning kas

jääkide jaotus vastab normaalsuse eeldusele.

```
# Näide R keskkonnas
# ANOVA
anova(lm_model)
```

```
Analysis of Variance Table

Response: h
          Df Sum Sq Mean Sq F value  Pr(>F)
d          1  158,5   158,5    85,5 5,7e-12 ***
Residuals 45   83,4     1,9
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
```

```
confint(lm_model)
```

```
            2,5 % 97,5 %
(Intercept) 14,49 18,099
d            0,26  0,405
```

7.1.6 Multikollineaarsus

Lineaarses regressioonis tähistab multikollineaarsus olukorda, kus kaks või enam sõltumatut tunnust (selgitajat) on omavahel tugevalt seotud või korreleerunud. See tähendab, et ühe selgitaja väärtusi saab suure täpsusega ennustada teiste selgitajate väärtuste põhjal. Multikollineaarsus võib olla probleemiks, sest:

Mõjutab koefitsientide hinnangute täpsust: multikollineaarsus suurendab koefitsientide hinnangute varieeruvust, mis muudab need ebastabiilseks. Väiksed muudatused andmestikus võivad tuua kaasa suuri muutusi koefitsientide hinnangutes, muutes mudeli interpretatsiooni keeruliseks.

Raskendab tunnuste tähtsuse hindamist: kõrge multikollineaarsuse korral võib olla raske kindlaks teha, milline argumenttunnus mõjutab sõltuvat tunnust, sest statistilised testid ei pruugi olla usaldusväärsed. See võib viia oluliste tunnuste ebaõiglase kõrvalejätmiseni või ebaoluliste tunnuste mudelisse lisamiseni.

Vähendab mudeli interpretatsiooni selgust: kui argumenttunnused on omavahel tugevalt seotud, muutub mudeli interpretatsioon keerulisemaks, kuna keeruline on eristada iga tunnus unikaalset mõju argumenttunnusele .

Multikollineaarsuse tuvastamiseks lineaarses regressioonis kasutatakse tavaliselt järgmisi meetodeid:

- **Korrelatsioonimatriks**: Vaadeldakse sõltumatute tunnuste vahelisi korrelatsioonikoefitsiente. Kõrge korrelatsioon näitab võimalikku multikollineaarsust.
- **Variance Inflation Factor (VIF)**: VIF mõõdab, kui palju tunnus variatiivsus suureneb multikollineaarsuse tõttu. Tavaliselt peetakse VIF väärtust üle 5 või 10 viitavaks tugevale multikollineaarsusele.
- **Tolerants**: Tolerants on VIF-i pöördväärtus ja näitab, kui suur osa tunnus variatsioonist ei ole seletatav teiste mudeli tunnustega. Madal tolerants viitab multikollineaarsusele.

Multikollineaarsuse lahendamiseks võib kasutada mitmeid lähenemisviise, näiteks tunnuste eemaldamist mu-

delist, tunnuste kombineerimist või andmete transformeerimist. Samuti võib kasulikuks osutuda regulaarimismeetodite, nagu ridge regressioon või lasso regressioon, kasutamine, mis aitavad vähendada koefitsientide varieeruvust, karistades suuri koefitsiente.

Vaatame järgmist näidet, kus puude kõrgust ennustame mõlema diameetri alusel. Kui varasemalt oli näha, et üks diameeter on oluline, siis antud mudelid ei ole enam kumbki oluline, sest nende tunnuse vahel on korrelatsioon.

```
# Näide R keskkonnas
summary(lm.col <- lm(h ~ d1 + d2, data = andmed))
```

```
Call:
lm(formula = h ~ d1 + d2, data = andmed)

Residuals:
   Min     1Q Median     3Q    Max
-2,709 -0,893 -0,200  0,787  3,143

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  16,1390     0,8902   18,13   <2e-16 ***
d1           -0,0231     0,2431   -0,09     0,92
d2            0,3643     0,2466    1,48     0,15
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Residual standard error: 1,34 on 44 degrees of freedom
Multiple R-squared:  0,671, Adjusted R-squared:  0,656
F-statistic:   45 on 2 and 44 DF,  p-value: 2,32e-11
```

Tehes täiendavalt ANOVA, on näha, et diameeter, mis lineaarses regressioonimudelis ei olnud oluline, on dispersioonanalüüsi järgi peamine varieeruvuse kirjeldaja.

```
# Näide R keskkonnas
anova(lm.col)
```

```
Analysis of Variance Table

Response: h
          Df Sum Sq Mean Sq F value  Pr(>F)
d1         1  158,5   158,5   87,73 4,8e-12 ***
d2         1    3,9     3,9    2,18    0,15
Residuals 44   79,5     1,8
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
```

```
# Arvutame VIF väärtuse multikollineaarsuse tuvastamiseks
vif(lm.col)
```

```
d1 d2
47 47
```

Kui VIF väärtus on üle 5, siis on tunnuste vahel tugev korrelatsioon, mida saame kontrollida ka korrelatsioonimaatriksiga.

```
# Näide R keskkonnas
cor(andmed[,c("d1","d2")])
```

```
      d1    d2
d1 1,000 0,989
d2 0,989 1,000
```

Sellest on näha, et need on tugevas korrelatsioonis (0,986). See mõjutab ka lineaarse regressiooni parameetrite arvutamist.

7.2 Dispersioonanalüüs

Mitmese regressioonanalüüsi korral on tegemist mudeliga, kus pidev funktsioontunnus avaldub pidevate argumenttunnuste lineaarkombinatsioonina. Dispersioonanalüüsi korral on tegemist mudelitega, kus funktsioontunnuseks on pidev tunnus, kuid argument-tunnused on diskreetsed (võivad olla nii arvulised kui ka mittearvulised). Dispersioon-analüüsi argumenttunnuseid nimetatakse faktoriteks. Ühefaktorilise dispersioonanalüüsi puhul uuritakse funktsioontunnuse sõltuvust ainult ühest faktorist. Faktori võimalikke väärtusi nimetatakse faktori tasemeteks.

Olgu üldkogum jaotatud mittelõikuvateks osadeks ehk rühmadeks faktori tasemete järgi. Öeldakse, et faktor X mõjub tunnusele Y, kui tunnuse Y keskväärtused eri rühmades on erinevad. Uurija käsutuses on valimi mõõtmisandmed, mille funktsioon-tunnuse rühmakeskmised tavaliselt erinevad. Dispersioonanalüüs on meetod, millega otsitakse vastust küsimusele, kas valimi rühmakeskmiste erinevus on põhjustatud uuritava faktori mõjust või valimi juhuslikkusest.

Dispersioonanalüüsi võimalused ja vahendid sõltuvad oluliselt valitud mudelist ja katsekorraldusest. Kui iga faktortunnuse igal tasemel on tehtud ühepalju mõõtmisi, nimetatakse mudelit tasakaalustatuks, vastasel juhul on tegemist tasakaalustamata mudeliga. Kui mõõtmisi on tehtud faktori kõigil võimalikel tasemetel, siis nimetatakse seda fikseeritud faktoriks. Kui faktortunnusel on palju erinevaid väärtusi ja mõõdetud tasemeid vaadeldakse juhusliku valimina faktori võimalike tasemete hulgast, siis nimetatakse seda juhuslikuks faktoriks. Järgnevalt käsitleme ainult fikseeritud faktoritega mudeleid.

Peatükk 8

Mittelineaarne regressioon

Mittelineaarne regressioon on statistiline meetod, mida kasutatakse sõltuva muutuja ja ühe või mitme sõltumatu muutuja vahelise mittelineaarse seose modelleerimiseks. Erinevalt lineaarsest regressioonist, mis eeldab sirgjoonelist seost, võimaldab mittelineaarne regressioon paindlikumaid seoseid, näiteks kõverjoonelisi, eksponentsiaalseid või logistilisi seoseid.

Metsanduses esineb sageli mittelineaarseid seoseid. Mittelineaarne regressioon võimaldab neid keerukaid seoseid täpsemalt modelleerida ja paremini mõista.

Mittelineaarne regressioonimudel kirjutatakse üldjuhul kujul:

$$Y = f(X, \beta) + \epsilon$$

kus:

- **Y** on sõltuv muutuja.
- **X** on sõltumatu(te) muutuja(te) vektor.
- $f(X, \beta)$ on mittelineaarne funktsioon, mis kirjeldab seost X ja Y vahel.
- β on mudeli parameetrite vektor, mida tuleb hinnata.
- ϵ on vealiige, mis esindab juhuslikku varieeruvust.

Mittelineaarse regressiooni parameetreid hinnatakse tavaliselt **vähimruutude meetodil**, kus otsitakse parameetrite väärtusi, mis minimeerivad mudeli ennustatud väärtuste ja tegelike väärtuste vaheliste ruutude summat. See on iteratiivne protsess, mis nõuab arvutitarkvara kasutamist.

8.1 R

Näiteks, kui soovime R-is kasutada `nls()` mudelit, mis kirjeldab puu kõrgust sõltuvalt puu diameetrist, võime kasutada järgmist näidiskoodi.

```r
# Laadime peatükis kasutatavad paketid
library(dplyr)
library(readxl)
```

```r
# Laadime andmed sisse failist ning
# filtreerime välja puud, mille h > 0
andmed <- read_excel("data/naited.xlsx","prt_andmed") %>% filter(h > 0)
```

```r
# Määratleme mudeli funktsioon
puu_mudel <- function(D, a, b, c) {
 1.3 + a * (D / (D + b))^c
}
```

```r
# Omista algväärtused parameetritele
alg_param <- list(a = 1, c = 0.5)
```

```r
# Sobita mudel andmetele
mudel <- nls(h ~ puu_mudel(d, a, 5, c), data = andmed, start = alg_param)
```

```r
# Vaata mudeli kokkuvõtet
summary(mudel)
```

```
Formula: h ~ puu_mudel(d, a, 5, c)

Parameters:
  Estimate Std. Error t value Pr(>|t|)
a   34,341      1,430   24,01  < 2e-16 ***
c    2,056      0,216    9,51  2,4e-12 ***
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Residual standard error: 1,32 on 45 degrees of freedom

Number of iterations to convergence: 7
Achieved convergence tolerance: 1,14e-07
```

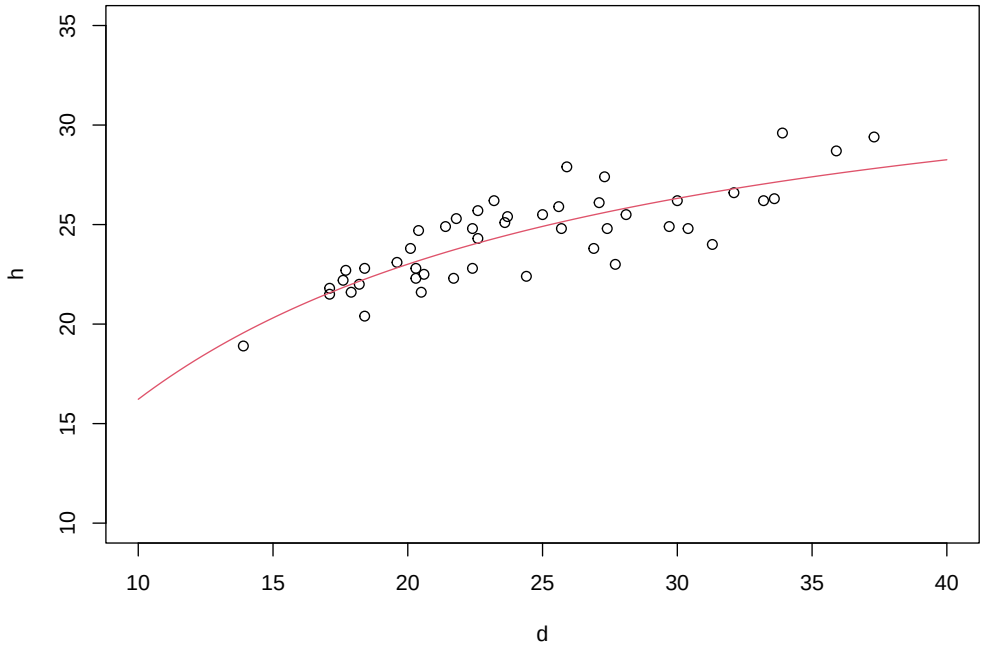
Esitame selle kohta ka graafiku, millele lisame mudeli joone funktsioonidega `curve()` ja `predict()`, milles kasutame loodud mudeli objekti.

```r
plot(h~d, andmed, ylim=c(10, 35), xlim = c(10, 40))
curve(predict(mudel, data.frame(d = x)), 10, 40, add=T, col = 2)
```

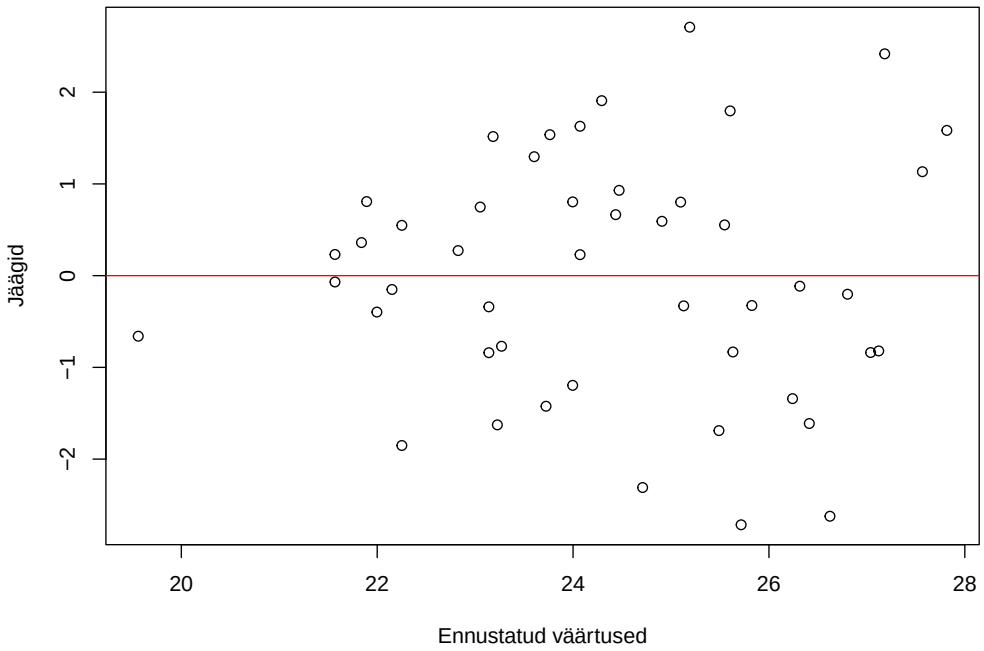
Kontrollime

```r
plot(fitted(mudel), resid(mudel),
 xlab = "Ennustatud väärtused",
 ylab = "Jäägid"
 )
abline(h = 0, col = "red")
```

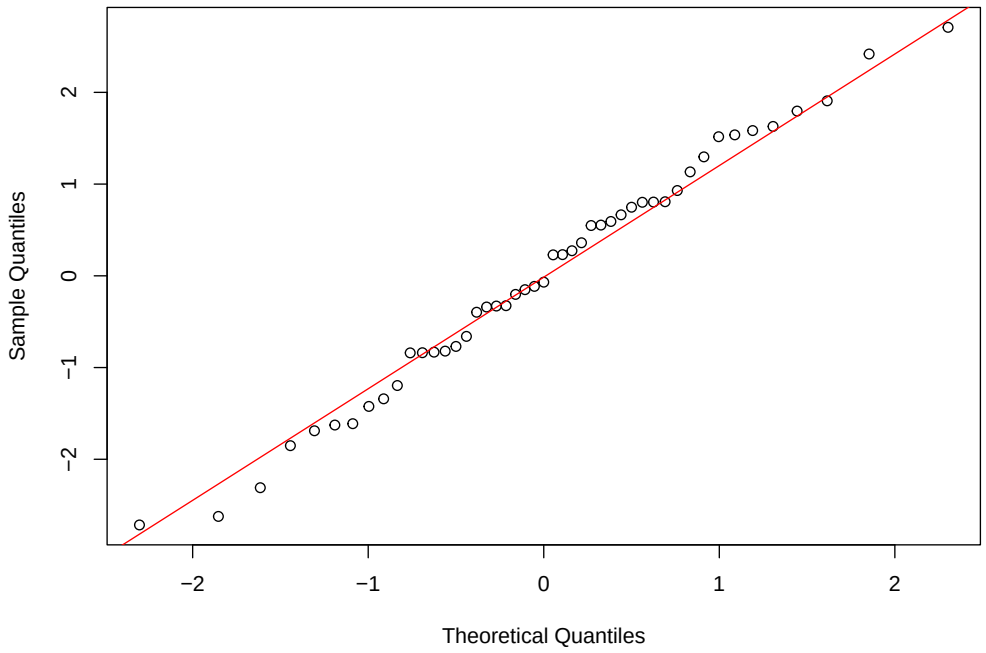
```r
qqnorm(resid(mudel), main = "")
qqline(resid(mudel), col = "red")
```



Joonis 8.1. Mittelineaarne regressioon



Joonis 8.2. Jäägid vs ennustatud



Joonis 8.3. Jääkide normaaljaotuse test

8.2 MS Excel

Andmete sisestamine. Avage fail, kus on puude diameetrid ja kõrgused.

Algparameetrite väärtuste sisestamine: Valige kolm vaba lahtrit ja sisestage neisse algparameetrite väärtused a, b ja c jaoks. Need väärtused võivad olla suvalised, aga parema tulemuse saamiseks peaksid need olema lähedased oodatud väärtustele. Näiteks võite alustada a = 20, b = 10 ja c = 3. Nimetage need lahtrid vastavalt "a", "b" ja "c".

Arvutatud kõrguse veeru loomine: Looge uus veerg nimega "Arvutatud H". Esimesse lahtrisse sisestage valem: `=1.3 + $a * ($D2 / ($D2 + $b))^$c`. Asendage `$a`, `$b` ja `$c` vastavate algparameetrite lahtritega (nt `$F$1`,`$F2`,`F3`).`D2` on esimese diameetri väärtus. See valem arvutab kõrguse vastavalt sisestatud diameetrile ja algparameetrite väärtustele.

Valemi kopeerimine: Lohistage esimese "Arvutatud H" lahtri paremas alumises nurgas olevat ruudukest alla, et kopeerida valem kõikidele diameetri väärtustele.

Ruutkeskmise vea arvutamine: Valige vaba lahter ja sisestage valem `=SUMXMY2(H2:Hn, E2:En)/COUNT(H2:Hn)`. See valem arvutab ruutkeskmise vea (RMSE) mõõdetud kõrguse (H) ja arvutatud kõrguse (Arvutatud H) vahel. `H2:Hn` on mõõdetud kõrguste vahemik ja `E2:En` on arvutatud kõrguste vahemik.

Solveri käivitamine: Avage menüüst "Andmed" -> "Solver". Kui Solverit menüüs ei ole, peate selle lisandmoodulitest aktiveerima (Fail -> Suvandid -> Lisandmoodulid -> Exceli lisandmoodulid -> Mine... -> Märkige ruut "Solver Add-in" ja vajutage "OK").

Solveri seadistamine: * Määrake "Seadista eesmärk" lahtriks RMSE lahter. * Valige "Min", et minimeerida RMSE. * "Muutuvate lahtrite kaupa" lahtritesse sisestage algparameetrite a, b ja c lahtrid. * Vajutage "Lahenda".

Tulemuste analüüs: Solver kohandab parameetrite a, b ja c väärtusi, et minimeerida RMSE. Kui Solver on lahenduse leidnud, kuvatakse leitud parameetrite väärtused ja minimaalne RMSE. Nüüd on teil mittelineaarne regressioonimudel, mis kirjeldab puude kõrguse ja diameetri seost teie andmete põhjal.

Lisamärkused:

- Solver ei pruugi alati leida optimaalset lahendust. Proovige erinevaid algparameetrite väärtusi ja käivitage Solver uuesti.
- Saate visualiseerida oma andmeid ja leitud regressioonikõverat, luues hajuvusdiagrammi, kus x-teljel on diameeter ja y-teljel kõrgus. Lisage diagrammile trendijoon, valides "Kohandatud" ja sisestades regressioonivõrrandi.
- Hinnake mudeli sobivust, arvestades RMSE väärtust ja visuaalset vastavust andmetele.

Peatükk 9

Mitteparameetriline analüüs

9.1 Üldistatud aditiivne mudel

Üldistatud aditiivne mudel (GAM) on statistiline mudel, mis ühendab lineaarse mudeli omadused aditiivsete funktsioonide abil. Lihtsamalt öeldes on GAM mudel, mis modelleerib funktsioontunnust ühe või mitme argumenttunnuse summa abil, kus iga argumenttunnuse mõju funktsioontunnusele modelleeritakse splainfunktsiooni abil.

Splainfunktsioon on matemaatiline funktsioon, mida kasutatakse andmete interpoleerimiseks ja sujuvaks lähendamiseks. See on ühesuguse struktuuriga tükiti polünomiaalne funktsioon, mis on konstrueeritud nii, et see oleks sujuv ja pidev andmepunktide vahel. See tähendab, et spline funktsioonil ei ole järske hüppeid ega nurki, mis teeb selle ideaalseks tööriistaks sujuvate kõverate ja pindade modelleerimiseks.

Selles näites kasutame R-i paketti `mgcv` GAM mudeli sobitamiseks. Eesmärk on modelleerida puu kõrguse ja diameetri vahelist seost, eeldades, et see seos ei pruugi olla lineaarne.

Kõigepealt laadime andmestiku, mis sisaldab puude kõrguse ja diameetri mõõtmiste andmeid.

```
# Laadime peatükis kasutatavad paketid
library(dplyr)
library(readxl)
library(mgcv)
```

```
# Andmete laadimine
andmed <- read_excel("data/naited.xlsx","prt_andmed") %>% filter(h > 0)
```

GAM mudeli lähendamiseks kasutame funktsiooni `gam()` paketist `mgcv`. Kasutame splainfunktsiooni diameetri muutuja jaoks, et modelleerida mittelineaarset seost.

```
# GAM mudeli sobitamine
mudel <- gam(h ~ s(d), data = andmed)
```

Selles koodis:

- `H ~ s(D)` määrab mudeli valemi. `H` on funktsioontunnus ja `s(D)` määrab, et diameetri (D) mõju kõrgusele modelleeritakse sujuva funktsiooni `s()` abil.
- `data = andmed` määrab andmestiku, mida kasutada.

3. Mudeli kokkuvõtte vaatamine:

Saame mudeli kohta detailse kokkuvõtte funktsiooni `summary()` abil:

```
# Mudeli kokkuvõte
summary(mudel)
```

```
Family: gaussian
Link function: identity

Formula:
h ~ s(d)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   24,368      0,182     134   <2e-16 ***
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Approximate significance of smooth terms:
      edf Ref.df  F p-value
s(d) 3,68   4,61 24  <2e-16 ***
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

R-sq.(adj) =  0,705   Deviance explained = 72,8%
GCV = 1,7242  Scale est. = 1,5525    n = 47
```

Kokkuvõte annab meile informatsiooni mudeli sobivuse kohta, sh:

- **Parametric coefficients:** Kui mudelis on ka lineaarseid komponente, näidatakse siin nende koefitsiendid. Kuna mudelis ühtegi lineaarset tunnust ei ole kasutatud, siis on siin ainult vabaliige `(Intercept)`.
- **Approximate significance of smooth terms:** Siin näeme p-väärtuse hinnangut splainfunktsiooni olulisuse kohta. Madal p-väärtus (<0.05) viitab statistiliselt olulisele mittelineaarsele seosele diameetri ja kõrguse vahel.
- **R-sq.(adj):** Korrigeeritud determinatsioonikordaja näitab, kui suure osa kõrguse varieeruvusest mudel seletab.
- **Deviance explained:** Sarnane determinatsioonikordajale, näitab mudeli poolt seletatud hälbimuse osakaalu.

Kasutades funktsiooni `coef()`, näeme splainfunktsiooni polünoomi parameetrite väärtusi.

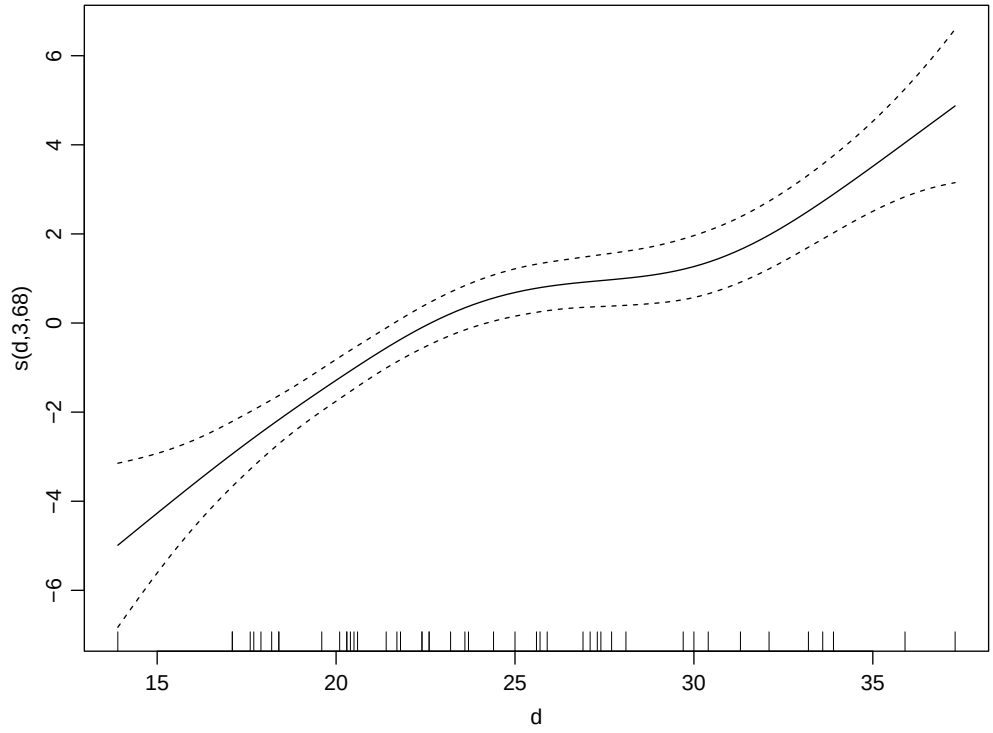
```
coef(mudel)
```

```
(Intercept)      s(d).1      s(d).2      s(d).3      s(d).4      s(d).5
   24,36809   -1,65141   -0,89974   -0,00196    0,03456   -0,04304
     s(d).6      s(d).7      s(d).8      s(d).9
```

```
0,18773    -0,13235    -0,81294     3,44774
```

Saame visualiseerida diameetri ja kõrguse vahelist seost funktsiooni `plot()` abil:

```
# Mudeli visualiseerimine
par(mar=c(4,4,1,1), mgp=c(2, 0.7, 0))
plot(mudel)
```



Joonis 9.1. Kõrguse sõltuvus diameetrist

See kuvab graafiku, mis näitab diameetri mõju puu kõrgusele. Pidevjoon näitab modelleeritud seost ja kriipsjooned näitavad usaldusvahemikku. See graafik aitab visuaalselt hinnata seose kuju ja mittelineaarsust.

Saame kasutada mudelit uute puude kõrguse ennustamiseks, andes ette diameetri väärtused funktsiooni `predict()` abil:

```
# Uute diameetrite väärtused
uued_diameetrid <- data.frame(d = c(12, 22, 32))

# Ennustuste tegemine
ennustused <- predict(mudel, newdata = uued_diameetrid)

# Ennustuste kuvamine
print(ennustused)
```

```
   1    2    3
18,1 24,1 26,3
```

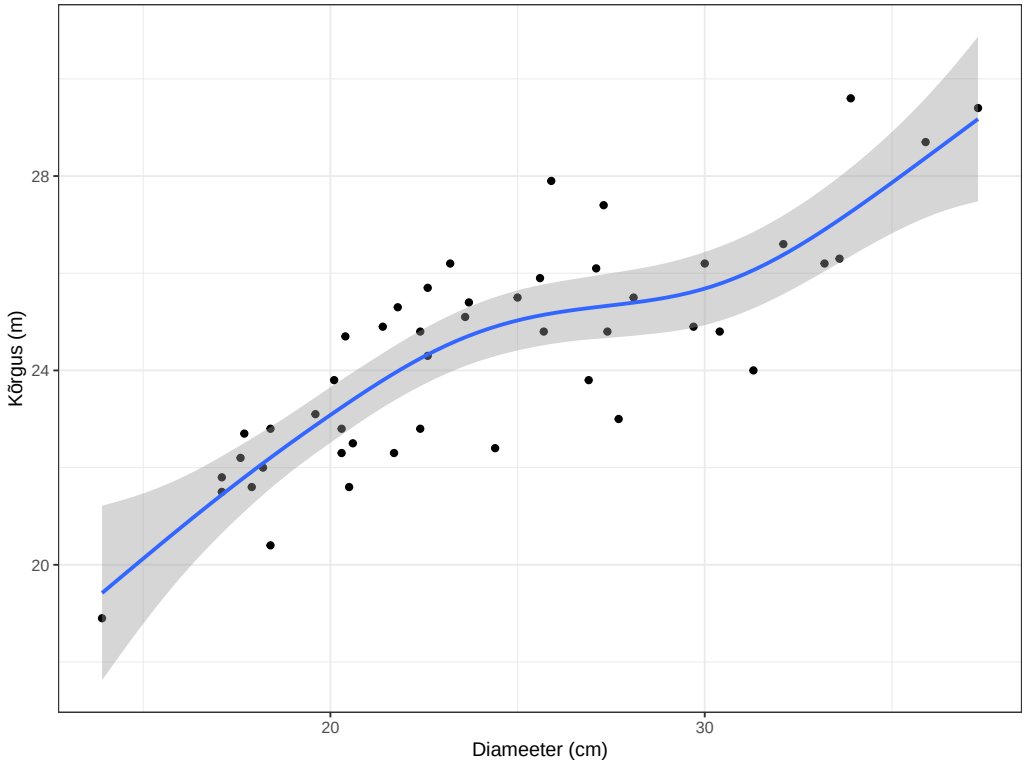
9.2 Visualiseerimine paketiga ggplot2

Oletame, et meil on samad andmed puude kõrguse (h) ja diameetri (d) kohta nagu eelmistes näidetes, aga me tahame luua hajuvusdiagrammi ja lisada sellele sujuva joone otse ggplot2 abil, ilma et peaks eraldi gam() funktsiooni kasutama.

```
# Laadi pakett
library(ggplot2)
```

Kasutame ggplot() funktsiooni koos geom_point() ja geom_smooth() funktsioonidega:

```
ggplot(andmed, aes(x = d, y = h)) +
  geom_point() +  # Joonistab hajuvusdiagrammi
  geom_smooth(method = "gam", formula = y ~ s(x)) +  # Lisab sujuva joone GAM mudeli abil
  labs(x = "Diameeter (cm)", y = "Kõrgus (m)") +
  theme_bw()
```



Joonis 9.2. Kõrguse sõltuvus diameetrist

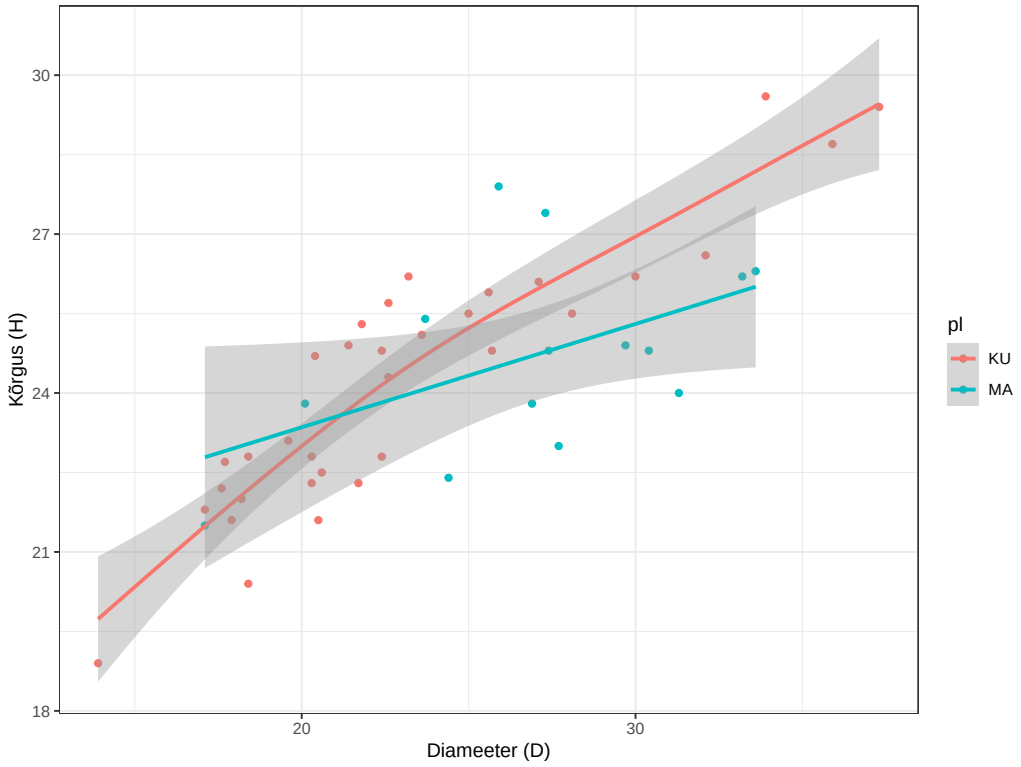
Selgitus:

- ggplot(puud, aes(x = d, y = h)): See rida loob ggplot objekti ja määrab andmestiku (andmed) ning x-telje (diameeter - D) ja y-telje (kõrgus - H) muutujad.
- geom_point(): See funktsioon lisab graafikule punktid, luues hajuvusdiagrammi. Iga punkt esindab ühte puud.
- geom_smooth(method = "gam", formula = y ~ s(x)): See funktsioon lisab graafikule suju-

va joone.
  - `method = "gam"`: Määrab, et sujuva joone sobitamiseks kasutatakse GAM mudelit.
  - `formula = y ~ s(x)`: Määrab mudeli valemi. `y ~ s(x)` tähendab, et y (meie puhul kõrgus H) modelleeritakse x (meie puhul diameeter D) sujuva funktsiooni `s()` abil. See on sama valem, mida kasutasime eelmistes näidetes `gam()` funktsiooni sees.
- `labs(...)`: Määrab telgede sildid ja graafiku pealkirja.
- `theme_bw()`: Muudab graafiku teema must-valgeks.

Sarnaselt eelmisele näitele saate `ggplot2` abil graafikut vastavalt oma soovile kohandada. Näiteks:

```
ggplot(andmed, aes(x = d, y = h, color = pl)) +
  # Muudab punktide värvi, kuju ja suurust
  geom_point() +
  # Eemaldab usaldusvahemiku ja muudab joone värvi
  geom_smooth(method = "gam", formula = y ~ s(x)) +
  labs(x = "Diameeter (D)", y = "Kõrgus (H)") +
  theme_bw()
```



Joonis 9.3. Kõrguse sõltuvus diameetrist

Selles näites:

- `aes(x = D, y = H, color = PL)` on lisatud, et värv sõltub puuliigist.

See näide demonstreerib, kuidas luua hajuvusdiagramm ja lisada sellele sujuv joon, mis on sobitatud GAM mudeli abil, kasutades ainult `ggplot2` funktsioone. See on mugav viis kiiresti visualiseerida mittelineaarseid

seoseid andmetes ilma eraldi GAM mudelit sobitamata. `ggplot2` pakub laialdasi võimalusi graafikute kohandamiseks, võimaldades luua informatiivseid ja visuaalselt meeldivaid visualiseeringuid.