Tilastotieteen johdantokurssi -luentokalvot

Yliopistonlehtori Juho Kopra

Itä-Suomen yliopisto, Tietojenkäsittelytieteen laitos

30.08.2023

1. Tilastotiede
2. Tilastotiedettä soveltavat tieteet
3. Miksi tilastotiedettä kannattaa opiskella?
4. Sattuma ja satunnaisuus
5. Tilastollisen päättelyn perusidea
6. R-kieli tilastotieteen työkaluna
7. Otanta
8. Mittaaminen
9. Muuttujat
10. Aineisto ja havaintomatriisi
11. Jakaumat
12. Kuvaajat
13. Tilastolliset tunnusluvut
14. Summamerkintä ja keskiarvon laskeminen
15. Hajontaluvut
16. Varianssin ja keskihajonnan laskeminen
17. Kvantiilit ja persentiilit
18. Riippuvuuden tarkastelu
19. Pearsonin tulomomentti-korrelaatiokerroin
20. Spearmanin järjestyskorrelaatiokerroin
21. Ristiintaulukko eli kontingenssitaulu
22. Run käyttäminen

Tilastotiede

Tilastotiede

"Statistics is the grammar of science." - Karl Pearson

Mitä tilastotiede on?

"Statistics is a branch of applied mathematics that involves the collection, description, analysis, and inference of conclusions from quantitative data." (Investopedia)

- Tilastotiede on tieteenala, joka käsittelee numeerisen aineiston hankintaa, kuvailua, analysointia, tulkintaa ja esittämistä.
- Käydään seuraavaksi läpi tilastotiede -sanan ja statistics -sanan alkuperää
- Tarkastellaan tutkimusprosessia ja tilastotieteen merkitystä suhteessa tutkimukseen.
- Tarkastellaan yleisimpiä tutkimustyyppejä

Tilasto -sanan alkuperä

- Kantasana tila, eli maatila
- Mainittu ensimmäisen kerran teoksessa Suomen Suuriruhtinaskunnan nykynen tilasto (1848)
- Kyse on ollut maatilojen ja alueiden luetteloinnista ja niitä koskevien tietojen kuvailusta sanallisesti ja keskiarvon avulla.
- Vedetään yhteen tietoja, saadaan aikaan tilasto (esim. montako asukasta maailmassa on peninkulmaa kohden)

Statistics -sanan alkuperä

- Tilastotiede on englanniksi statistics, joka tulee sanasta state (valtio) vuodelta 1791. Statistics on tarkoittanut valtion hallinnolliseen käyttöön kerättyä yleensä numeromuotoista aineistoa, jota voidaan esittää myös kuvien avulla.
- Suomeksi statistics kääntyy myös muotoon tunnusluvut. Tunnusluvuilla on paljon käyttöä tilastotieteessä.
- Suomessa viralliset tilastot mm. valtion käyttöön tuottaa Tilastokeskus (www.stat.fi). Ks. myös www.findikaattori.fi.

Tiede

Tiede on (wikipedia.org)

“todellisuuden ilmiöiden ja niiden välisten suhteiden järjestelmällistä ja arvostelevaa tutkimista”

“sekä sen avulla saatua tietojen jäsentynyttä kokonaisuutta”

Eli kun opiskelemme tilastotiedettä, niin opiskelemme tiedettä. Olemme oppimassa tutkimisen taitoja ja samalla tietojen kokonaisuutta. Tilastotiede on luonteeltaan menetelmätiede, eli oppi menetelmistä, joilla voidaan tehdä tutkimusta koskien yleensä numeerisia aineistoja.

Tutkimusprosessi

1. Ongelman asettaminen
2. Ongelman täsmentäminen ja tutkimusstrategian laatiminen
3. Aineiston kerääminen
4. Aineiston kuvaaminen
5. Aineiston analyysi
6. Johtopäätösten teko
7. Tutkielman tai raportin laatiminen
8. Tutkimustulosten julkaiseminen

- Kohdat 1 ja 2: sovellusalan osaaminen ja tutkimuskirjallisuuden tuntemus
- kohdat 3-6: Puhuttaessa määrällisestä tutkimuksesta, vaiheet 3-6 ovat oleellisesti TILASTOTIEDETTÄ riippumatta siitä, mistä sovellusalasta tutkimus on peräisin
- Kohdat 7 ja 8: kirjoittamisen taito

Tutkimusprosessi ja tilastolliset menetelmät

3. Aineiston keräämisen tekniikat: otantateoria, koesuunnittelun teoria, mittaaminen
4. Aineiston kuvaamisen tekniikat: laadullisen aineiston luokittelu, numeeristen aineistojen muodostaminen, näiden tilastollinen kuvailu jakaumien ja tunnuslukujen avulla, graafinen esitys
5. Johtopäätösten teon tekniikat: tilastollisten testien teoria, tilastollinen päättely, tulosten tulkinta

Tilastotieteen menetelmät ovat hyvin tärkeitä soveltavien tieteiden aloilla. Ilman tilastomenetelmien tuntemista ei usein voi toteuttaa tutkimusta soveltavilla aloilla, kuten lääketiede, terveystieteet, hoitotiede, farmasia, biolääketiede, metsätiede, ympäristötiede, kauppatiede tai tietojenkäsittelytiede.

Tutkimustyyppejä

- Tilastotieteen tutkimus voidaan jakaa teoreettiseen ja soveltavaan
  - **teoreettinen tilastotiede** kehittää tilastomenetelmiä matematiikan avulla
  - **soveltava tilastotiede** käyttää tilastomenetelmiä jonkin toisen tieteenalan tutkimuksessa aineiston analyysissä
- Soveltavat tutkimukset voidaan jaotella sen mukaan, miten aineisto kerätään
  - **kokeellinen tutkimus**: tutkija kontrolloi ainakin joitakin muuttujia
  - **havainnoiva tutkimus**: tutkija seuraa ilmiötä tehden mittauksia, kuitenkaan määräämättä muuttujien arvoja
    - **poikkileikkaustutkimus**: mitataan tiettynä ajankohtan tutkittavan ilmiön tilaa (satunnaisotanta)
    - **pitkittäistutkimus**: samoille yksilöille toistetaan mittauksia eri ajanhetkinä (esim. lasten pituuden seuranta)
  - Tutkimustyyppejä on paljon muitakin, erityisesti terveystieteissä, joita ei tässä käsitellä.

Esimerkkejä soveltavista tutkimuksista

- **kokeellinen tutkimus**: tutkija kontrolloi ainakin joitakin muuttujia
  - esim. halutaan tutkia, miten eri tekijät vaikuttavat mansikan taimien kasvuun. Tutkija kontrolloi mullan laatua, lannoitusta, valon määrää ja kastelua.
- **havainnoiva tutkimus**: tutkija seuraa ilmiötä tehden mittauksia, kuitenkaan määräämättä muuttujien arvoja
  - **poikkileikkaustutkimus**: mitataan tiettynä ajankohtana tutkittavan ilmiön tilaa (satunnaisotanta)
  - esim. FINTERVEYS selvittää suomalaisen aikuisväestön terveydentilaa. Mitataan pituus, paino, BMI, verenpaine, kysytään ravitsemusta, tupakointia, alkoholinkäyttöä, liikuntaa jne.
  - haastattelututkimus eli gallup: esim. vaaligallup
  - **pitkittäistutkimus**: samoille yksilöille toistetaan mittauksia eri ajanhetkinä (esim. lasten pituuden seuranta)

Tilastotiedettä soveltavat tieteet

Tilastotiedettä soveltavat tieteet

- **Seuraavaa kuutta kalvoa (tämä ml.) ei käsitellä luennolla, mutta silmäile ne läpi ja lue oman alasi osalta tarkemmin.**
- Tilastotiede on tieteenala, joka kehittää menetelmiä, joiden avulla tutkitaan ilmiöitä ja niiden säännönmukaisuuksia numeerisen aineiston perusteella.
- Tilastotiede auttaa tekemään johtopäätöksiä ja ennusteita ilmiöistä, jotka ovat epävarmoja tai satunnaisia. Esimerkiksi tilastotiedettä voidaan käyttää arvioimaan Suomen väestön ominaisuuksia, kuten työttömien lukumääriä maakunnittain.

Luonnontieteet

- Luonnontieteissä tilastotieteen menetelmiä käytetään mm. seuraaviin tarkoituksiin:
  - Hypoteesien testaaminen kokeellisilla tai havainnollisilla aineistoilla
  - Mallintaminen ja simulointi fysikaalisista, kemiallisista tai biologisista ilmiöistä
  - Mittausten ja kokeiden suunnittelu ja optimointi
  - Virheiden ja epävarmuuksien arviointi ja hallinta
  - Monimuuttuja-analyysi ja koneoppiminen suurten ja monimutkaisten aineistojen käsittelyssä

Yhteiskuntatieteet

- Yhteiskuntatieteissä tilastotieteen menetelmiä käytetään mm. seuraaviin tarkoituksiin:
  - Kyselytutkimusten suunnittelu, toteutus ja analysointi
  - Tilastollinen päättely väestöllisistä tai yhteiskunnallisista ilmiöistä
  - Regressioanalyysi ja muut riippuvuussuhteiden tutkimisen menetelmät
  - Faktori- ja klusterianalyysi ja muut latenttien rakenteiden paljastamisen menetelmät
  - Aikasarja-analyysi ja ennustaminen taloudellisista tai sosiaalisista muuttujista

Lääketiede

- Lääketieteessä tilastotieteen menetelmiä käytetään mm. seuraaviin tarkoituksiin:
  - Kliinisten tutkimusten suunnittelu, toteutus ja analysointi
  - Lääkekehitys ja lääkevaikutusten arviointi
  - Epidemiologia ja tartuntatautien leviämisen mallintaminen
  - Biostatistiikka ja geneettisten tai molekyylitason aineistojen analysointi
  - Etiikka ja tilastollinen merkitsevyys lääketieteellisessä päätöksenteossa

Taloustiede

- Taloustieteessä tilastotieteen menetelmiä käytetään mm. seuraaviin tarkoituksiin:
  - Makro- ja mikrotaloudellisten ilmiöiden kuvaaminen, selittäminen ja ennustaminen
  - Taloudellisten teorioiden testaaminen empiirisillä aineistoilla
  - Taloudellisten mallien estimointi ja validointi
  - Ekonometria ja taloudellisten riippuvuussuhteiden tutkiminen
  - Tilinpito ja rahoitusmarkkinoiden analysointi

Tekniikka

- Tekniikassa tilastotieteen menetelmiä käytetään mm. seuraaviin tarkoituksiin:
  - Laadunvalvonta ja prosessien parantaminen
  - Suunnittelukokeet ja tuotekehitys
  - Luotettavuus- ja riskianalyysi
  - Signaalinkäsittely ja kuvantaminen
  - Teollisuusmatematiikka ja optimointi

Miksi tilastotiedettä kannattaa opiskella?

Miksi tilastotiedettä kannattaa opiskella?

- Tilastotiedettä kannattaa opiskella yliopistossa sivuaineena, koska se:
  - Antaa sinulle vahvan metodologisen osaamisen ja kriittisen ajattelun taidon
  - Parantaa sinun mahdollisuuksiasi työllistyä ja edetä urallasi
  - Laajentaa sinun näkökulmaasi ja ymmärrystäsi eri alojen ilmiöistä ja ongelmista
  - Mahdollistaa sinulle monitieteisen yhteistyön ja verkostoitumisen

Sivuaineen valinta

- Tilastotiede sopii hyvin sivuaineeksi minkä alan tutkintoon tahansa
- Voit valita tilastotieteen perus- tai aineopinnot tai edetä pidemmälle tilastollisen koneoppimisen suuntaan
- Voit opiskella tilastotiedettä eri ohjelmistoilla, kuten R tai SPSS
- Voit saada tilastollista neuvontaa opintojesi aikana

Sattuma ja satunnaisuus

Sattuma ja satunnaisuus

- Arkipuheessa, kun jotain tapahtuu sattumalta, on se jotain mitä ei voinut arvata ennalta.
- **Satunnaisuus** on keskeinen termi tilastotieteessä ja todennäköisyyslaskennassa.
- Kun tieteessä jotain tapahtumaa pidetään täysin satunnaisena, se tarkoittaa, että kyseistä tapahtumaa ei voida mitenkään ennustaa.
  - Esim. nopanheiton silmäluku on satunnainen.
- Satunnaisuus ei tarkoita, että kaikki mahdolliset arvot olisivat yhtä todennäköisiä.
- Satunnaisen tapahtuman eri arvojen yleisyyttä voidaan jäsentää **todennäköisyyden** avulla.

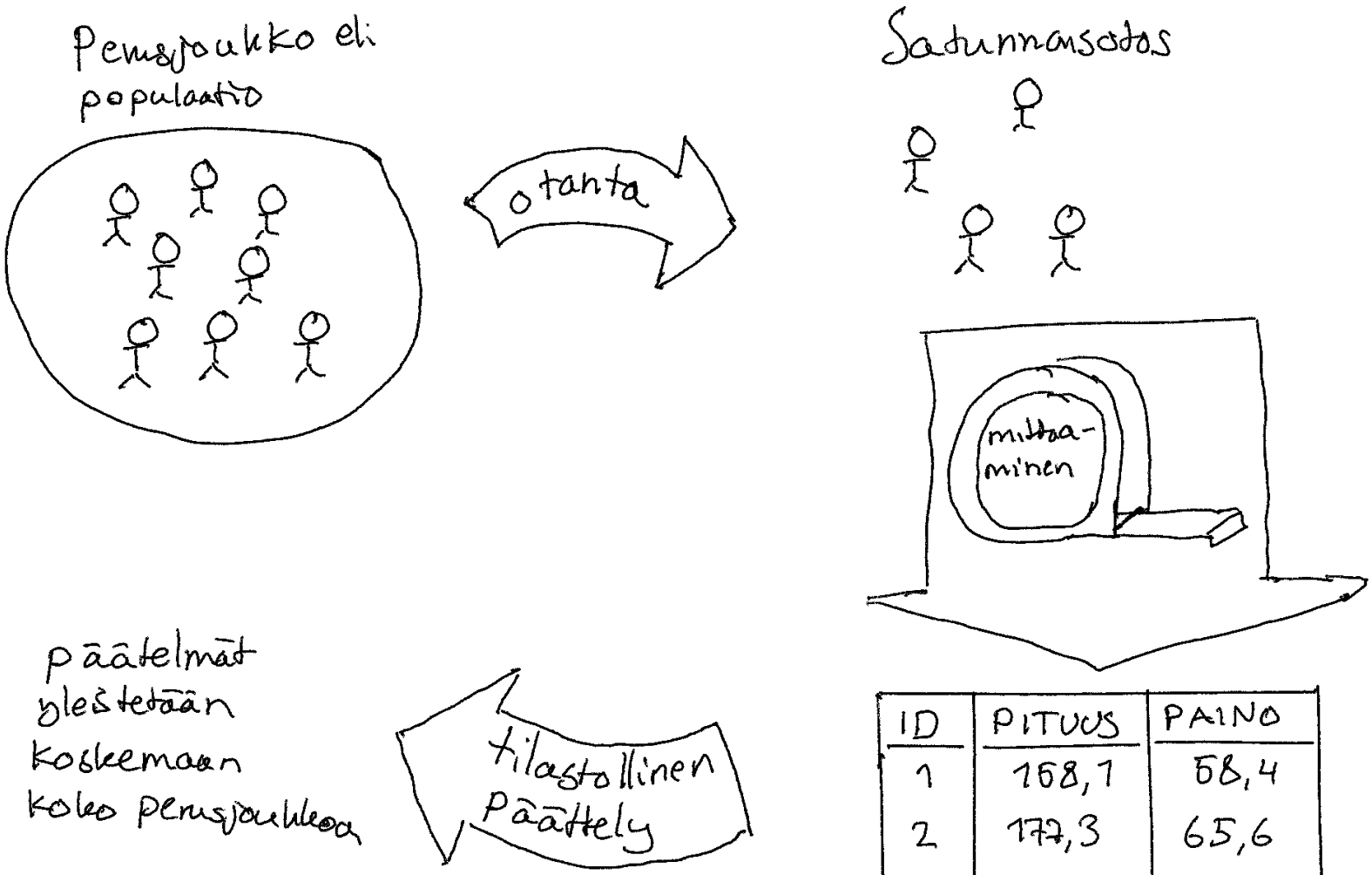
Sattuma ja satunnaisuus (jatkuu)

Tilastotieteessä satunnaisuutta käytetään tilanteesta riippuen hyödyksi tai sitä pyritään kontrolloimaan.

- tilastollinen mallintaminen pyrkii löytämään aineistosta ei-satunnaisen (ns. signaalin) ja erottamaan tästä satunnaisvaihtelun
- tilastollinen hypoteesintestaus selvittää, johtuuko aineistossa havaittu vaihtelu sattumasta, vai löytyykö näyttöä tutkittavan hypoteesin puolesta tai vastaan
- satunnaisotanta mahdollistaa tulosten yleistämisen perusjoukkoon
- koesuunnittelussa satunnaistamisen avulla voidaan erottaa lääkkeen todellinen vaikutus lumevaikutuksesta eli placebosta

Tilastollisen päättelyn perusidea

Tilastollisen päättelyn perusidea



ID	PITUUS	PAINO
1	168,1	58,4
2	172,3	65,6
3	189,5	79,0
4	166,4	75,3

R-kieli tilastotieteen työkaluna

R-kieli tilastotieteen työkaluna

- Vaikka tilastotiede on pitkälti matematiikkaa ja sen hyödyntämistä, käytännön työ on järkevintä tehdä tietokoneella. Kaikki kiinnostavat aineistot ovat sen verran suuria, että minkä tahansa analyysin tekeminen käsin on turhan aikaaviepää.
- Tällä kurssilla käytämme R-kieltä ja RStudiota työkaluina. Emme syvenny kieleen kovin syvällisesti, vaan tutustumme siihen pikkuhiljaa. Jo tässä vaiheessa on kuitenkin hyvä puhua hieman ohjelmoinnista, jotta voimme totutella antamaan tietokoneelle komentoja koodin muodossa.

Ohjelmoinnin perusperiaatteet

- Ohjelmointikielissä, kuten R-kieli, on tietyt perussanat, jotka on tiedettävä, jotta kieltä pystyy käyttämään. Samoin kielessä on oma ns. kielioppinsa, jota kutsutaan syntaksiksi. Sekä käytettävien sanojen että syntaksin on oltava täysin oikein, jotta tietokone suostuu tekemään mitään. Pienetkin piste- tai pilkkuvirheet tai väärä tai puuttuva kirjain aiheuttaa yleensä virheilmoituksen.
- Ohjelmointikielissä tietokoneelle annetaan ohjeita (siksi kai sanakin ohjelmointi?), joita tietokone noudattaa annetussa järjestyksessä. Siksi on väliä, että monivaiheiset komennot annetaan oikeassa järjestyksessä!
- Tässä vaiheessa ei ole tarpeen opetella kaikkea R:ään liittyvää ulkoa, vaan voit katsoa materiaalista oikeat komennot. Mikäli haluat oppia R:ää syvemmin, ei kuitenkaan ole haitaksi jos joitain keskeisimpiä työkaluja jää jo muistiin.

R:n peruskäyttöä

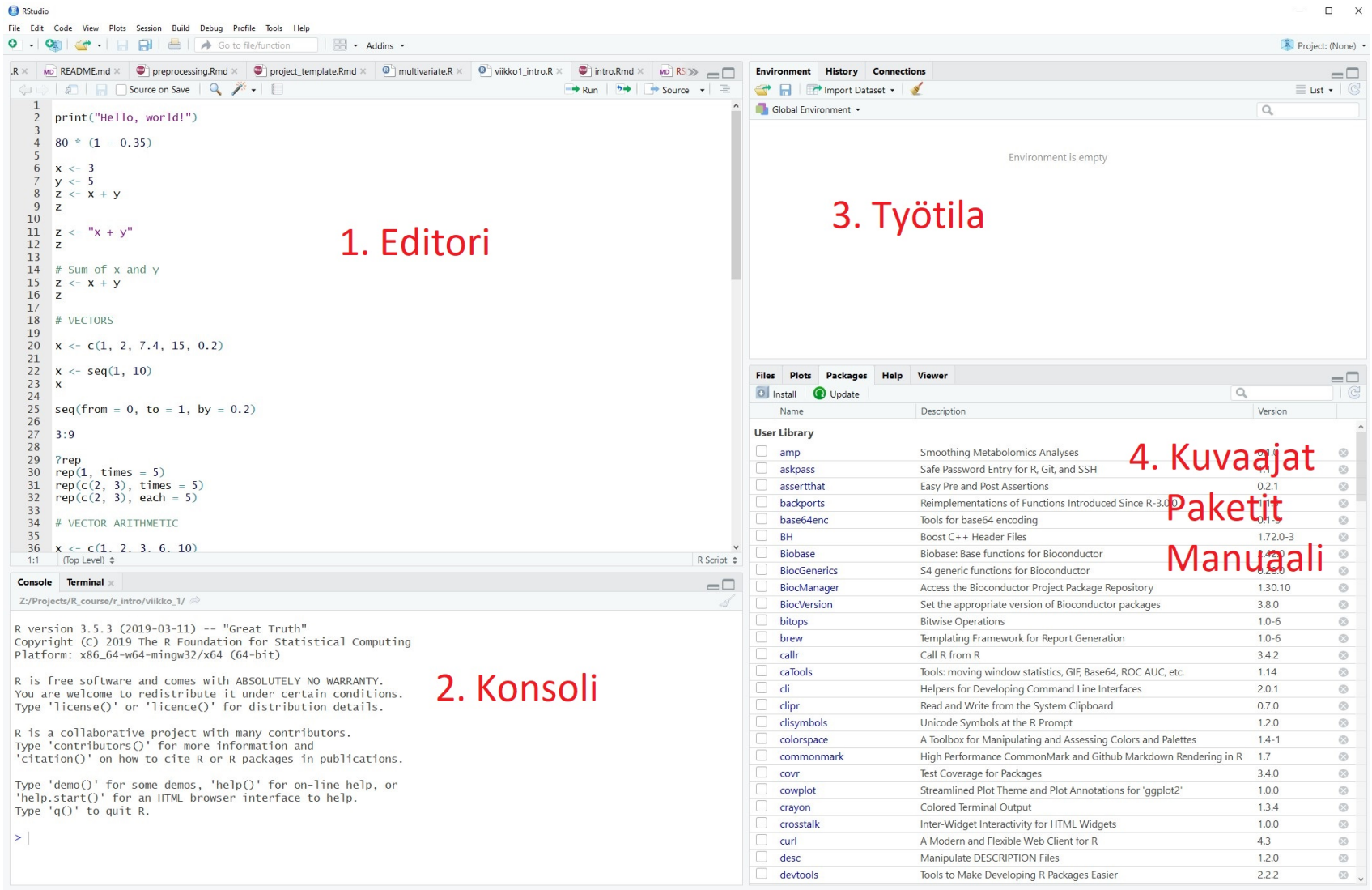
```
x <- 1.5 # sijoittaa luvun 1.5 objektiin x
x # tulostaa x:n tiedot konsoliin
y <- 1:5 # sijoittaa luvut 1-5 objektiin y
y2 <- y*2 # y2 on kaksi kertaa y:n arvot
?sample # avaa sample-funktion ohjesivun
sample(y,2) # poimii 2 havainnon satunnaisotoksen y-objektista
```

- Peruslaskutoimitukset: + - * / ^
- Vertailuoperaattorit: == != <= >=

R ja RStudio

- R ja RStudio ovat eri ohjelmia. RStudio on graafinen ympäristö R-kielen käyttöön.
- R:ää voi käyttää jopa komentoriviltä tai jonkin muun graafisen ympäristön kautta. Windows-tietokoneilla R:n mukana asentuu R Gui, jota ei suositella käytettävän.

RStudio



Pakettien asentaminen ja lataaminen

- R koostuu perusosasta (base) sekä lisäosista eli paketeista, joilla R:n toiminnallisuutta voidaan laajentaa. Asennetaan nyt remotes -niminen paketti, jota tarvitaan toisen paketin asentamista varten.

```
install.packages("remotes")
```

- Tyypillisesti saat asennettua tarvitsemasi paketit samalla tavalla kuin yllä asennettiin remotes.
- Jos haluamaasi pakettia ei ole lähetetty R:n yleiseen pakettiarkistoon (Comprehensive R Archive Network, CRAN) vaan se on githubissa, niin sen voi asentaa seuraavasti. Asennetaan siis kurssilla tarvittava paketti datas4uef.

```
remotes::install_github("jukop/datas4uef")
```

Objektien luominen

- Kun mitä tahansa tietoja halutaan työstää R:ssä, on se ladattava muistiin jollekin nimelle, jonka käyttäjä voi itse valita. Nimen ei tule kuitenkaan olla mikään varattu sana R-kielessä, joten jos tiedät jo joitakin varattuja sanoja, niin vältä niiden käyttöä objektien niminä.
- Esimerkiksi voin luoda objektin nimeltä `x`, jossa on luvut 1, 5 ja 6:
  - `x <- c(1, 5, 6)`
- Edellä käytimme sijoitusoperaattoria <- sekä yhdistimme luvut funktiolla c (c kuten combine)
  - Sijoitusoperaattorin merkkien välissä ei saa olla välilyöntiä!
  - Funktiokutsun tunnistaa siitä, että sen jälkeen on sulut ja sulkujen väliin tulee funktioiden argumentit eroteltuna pilkuilla. Tässä tapauksessa argumentit olivat arvot 1, 5 ja 6.

Lisätietoa

- Käsittelemme kulloisenkin aiheen yhteydessä miten asioita tehdään R:llä. Tällä tavoin R tulee ainakin vähän tutuksi, vaikkei sitä tarvitsekaan tällä kurssilla osata tentissä itse käyttää (mutta sitä saa käyttää). Osassa harjoitustehtäviä kuitenkin tarvitaan R:n käyttöä.
- Jos haluat paneutua heti enemmän R:n käyttöön, voit hyödyntää R-kieli -kurssin materiaalia: https://vm3751.kaj.pouta.csc.fi/shiny/_book/r-kurssi.html

Tilastotieteen termit

- Kurssin alkuvaiheessa esintyy paljon tilastotieteen termejä, jotka on syytä opetella tuntemaan.
- Mikäli luet englanninkielistä tutkimus- tai oppikirjallisuutta, niin suomalainen tilastotieteen sanasto kannattaa ottaa avuksi: https://sanasto.tilastoseura.fi/
  - Luentokalvoilla en esittele englanninkielistä termistöä, mutta *verkkoluentomonisteessa* on usein mainittu myös englanninkielinen termi.
- Uusimmille menetelmille ei välttämättä ole vakiintunutta suomenkielistä nimeä. Siksi englanninkielisiä termejä käytetään edelleen paljon.

Otanta

Otanta

- Otanta on tilastotieteen osa-alue, joka tarkastelee kuinka havaintoyksiköt kannattaa poimia perusjoukosta
- Otanta on tärkeää, koska sen avulla voidaan tehdä tilastollista päättelyä koko perusjoukosta otoksen perusteella
- Otannassa pyritään poimimaan edustava otos perusjoukosta.
- Yleisimpiä otantamenetelmiä ovat yksinkertainen satunnaisotanta palauttaen ja palauttamatta
  - Riippuen tutkimuksen tavoitteista voidaan käyttää myös monimutkaisempia otantamenetelmiä, jotka voivat olla tehokkaampia tiettyyn kysymykseen vastattaessa.

Perusjoukko ja otos

- Perusjoukko on tilastollisen tutkimuksen kohdejoukko eli se joukko yksiköitä, joista halutaan saada tietoa. Esimerkiksi jos halutaan tutkia suomalaisten mielipiteitä EU:sta, niin perusjoukkona ovat kaikki Suomen kansalaiset.
- Otos on perusjoukon osajoukko eli se joukko yksiköitä, joilta kerätään aineisto tilastollista analyysia varten. Esimerkiksi jos halutaan tutkia suomalaisten mielipiteitä EU:sta, niin otoksena voi olla 1000 satunnaisesti valittua Suomen kansalaista.

Satunnaisotanta ja satunnaisotos

- Tyypillinen vaatimus hyvälle otokselle on että se on poimittu nk. satunnaisotantaa käyttäen
  - Ts. tutkija tai mikään taustatekijä ei määrää sitä, kuka tulee poimituksi otokseen. Jos näin olisi, niin on vaikeaa tehdä päätelmiä perusjoukkoa koskien.
  - Jos otos ei ole satunnaisotos, siitä käytetään nimeä näyte.
- Otoksen tulee olla edustava eli sen tulee heijastaa perusjoukon ominaisuuksia mahdollisimman hyvin.
- Suurempi otos mahdollistaa tarkemman käsityksen saamisen perusjoukosta, mutta yleensä aineiston kerääminen on kallista. Siksi otoksen koko on painottelua tulosten tarkkuuden ja tutkimuksen kustannusten kanssa.

Yksinkertainen satunnaisotanta

- Yksinkertainen satunnaisotanta on yleisin ja yksinkertaisin otantamenetelmä
- Siinä poiminta tehdään suoraan arpomalla alkiot perusjoukosta
- Jokaisella alkiolla on sama todennäköisyys tulla valituksi otokseen

Palauttaen vai palauttamatta?

- Yksinkertainen satunnaisotanta voidaan tehdä joko palauttaen tai palauttamatta valittu alkio takaisin perusjoukkoon
- Tehtäessä otanta palauttaen sama alkio voi tulla otokseen monta kertaa
- Tehtäessä otanta palauttamatta jokainen alkio voidaan valita vain kerran

Otanta R-kielellä

Yksinkertainen satunnaisotanta palauttamatta

```
# luodaan vektori porukka, josta poimitaan otos
porukka <- c("Erkki", "Jaska", "Niina")
# porukka -vektorissa on kolme alkiota, joten
# voidaan poimia kokoa 1, 2 tai 3 oleva otos
sample(porukka, size=2)
```

```
## [1] "Niina" "Jaska"
```

```
# uusi funktiokutsu poimii uuden otoksen
sample(porukka, size=2)
```

```
## [1] "Erkki" "Niina"
```

Otanta R-kielellä jatk.

Yksinkertainen satunnaisotanta palauttaen

```
sample(porukka, size=4,replace=TRUE)
```

```
## [1] "Niina" "Niina" "Jaska" "Erkki"
```

- Asettamalla nk. alkuluku (seed) ennen koodin suorittamista arpoo aina saman otoksen
  - Vertaa kaverin kanssa tai suorita useita kertoja peräkkäin

```
set.seed(42)
sample(porukka, size=2)
```

```
## [1] "Erkki" "Niina"
```

Mittaaminen

Johdanto

- Jotta tutkittavasta ilmiöstä saadaan muodostettua numeerinen/määrällinen aineisto, on tehtävä mittauksia.
- Mittaamisen kohteena ovat tilastoyksiköt eli ne yksiköt, joista halutaan saada tietoa. Tilastoyksikkö voi olla esimerkiksi henkilö, organisaatio, tapahtuma tai asiakirja.
- Mittaamisen avulla voidaan määrittää tutkittavien tilastoyksiköiden ominaisuuksia numeerisesti.
- Mittaamisen tuloksena syntyy muuttujia, jotka kuvaavat tilastoyksiköiden ominaisuuksia. Muuttujan arvo kullekin tilastoyksikölle kertoo tilastoyksikölle tehdyn mittaustuloksen.

Esimerkkejä mittaamisesta

- pituus mitattuna mittanauhalla, punnittu paino
- henkilön oma käsitys painostaan (eri kuin edellä)
- ajan mittaaminen sekuntikellolla
- liikenneonnettomuuksien määrä Valtatie 9:llä vuoden aikana
- formulakuljettajan sijoitus kilpailussa
- lämpötila $^\circ C$ (vrt. $^\circ F$)
- silmien väri (laatuero)
- auton merkki, lintulaji
- älykkyysosamäärä

Mittaustapoja

Mittaustapoja on ainakin

- mittalaitteella mittaaminen (esim. vaaka)
- henkilöltä kysyminen (kasvotusten, puhelimitse, netissä tai kirjeellä)
- tietokannasta virallisen tilaston tms. mukaan, rekisteriaineistot
- havainnoimalla ja kirjaamalla havainnot ylös
- kyselylomakkeella

Muuttujien tyypit

- Muuttujat voidaan luokitella eri tavoin niiden ominaisuuksien perusteella. Yksi tapa on luokitella muuttujat niiden arvojen tyypin mukaan numeerisiin ja luokkamuuttujiin.
- **Numeeriset muuttujat** ovat sellaisia, joiden arvot ovat lukuja. Numeeriset muuttujat voidaan jakaa edelleen jatkuviin ja diskreetteihin muuttujiin. Jatkuvilla muuttujilla on teoriassa äärettömän monta mahdollista arvoa tietyllä välillä, esimerkiksi pituus tai paino. Diskreeteillä muuttujilla on vain rajallinen määrä mahdollisia arvoja, esimerkiksi lasten lukumäärä tai silmien väri.
- **Luokkamuuttujat** ovat sellaisia, joiden arvot ovat sanoja tai merkkejä. Luokkamuuttujat voidaan jakaa edelleen dikotomisiin ja moniarvoisiin muuttujiin. Dikotomisilla muuttujilla on vain kaksi mahdollista arvoa, esimerkiksi sukupuoli tai kyllä/ei-vastaus. Moniarvoisilla muuttujilla on useampia mahdollisia arvoja, esimerkiksi poliittinen kanta tai ammatti.

Muuttujien mitta-asteikot

Muuttuja kuuluu johonkin seuraavista mitta-asteikoista:

1. Laatueroasteikko
2. Järjestysasteikko
3. Välimatka-asteikko
4. Suhdeasteikko

Mitta-asteikkojen ominaisuudet

1. Laatueroasteikko
   - toisensa poissulkevat luokat
   - esim. mies/nainen, Volvo/Audi/BMW, talitiainen/sinitiainen/...
   - voidaan sanoa, mihin luokkaan havainto kuuluu, mutta suuruusjärjestys ei ole mielekäs
2. Järjestysasteikko
   - myös toisensa poisulkevat luokat
   - eri arvojen suuruusjärjestystä voidaan vertailla
   - esim. Likert-asteikko täysin eri mieltä/jokseenkin eri mieltä/ei samaa eikä eri mieltä/jokseenkin samaa mieltä/täysin samaa mieltä
   - eri arvojen etäisyyttä ei voida ilmaista numeerisesti

Mitta-asteikkojen ominaisuudet

3. Välimatka-asteikko
   - yleensä jatkuva, mutta mittaustarkkuus voi tehdä myös diskreetin
   - kahden lukuarvon erotus määrittää etäisyyden
   - käytössä on mittayksikkö, esim. $^\circ C$
     - $25^\circ C - 20^\circ C = 5^\circ C$
   - jakolaskua ei pidetä mielekkäänä
4. Suhdeasteikko
   - välimatka-asteikon ominaisuudet ja lisäksi:
   - muuttujan arvon saadessa arvon 0, mitattava ominaisuus häviää
   - esim. lämpötila kelvinasteina ($^\circ K$) mittaa lämpöliikkeen määrää ($0^\circ K$ = lämpöliike lakkaa), etäisyys sentteinä (cm), paino (kg)
   - jakolaskulla on mielekäs tulkinta, esim. 75 cm pitkä lapsi on 1.5-kertaa pidempi kuin 50 cm pitkä lapsi $\frac{75\ cm}{50\ cm} = 1.5$

Mitta-asteikot R-kielessä

- R-kieli olettaa, että kaikki laskutoimitukset ovat sallittuja
  - Ts. jos muuttuja on numeerinen, niin R olettaa suhdeasteikon
- Toisaalta käytännön kannalta ei yleensä ole väliä onko muuttuja suhdeasteikollinen vai välimatka-asteikollinen (jakolasku on harvinainen vaatimus perusmenetelmissä)
  - Siispä keskitytään siihen onko muuttuja numeerinen (vähintään välimatka-asteikollinen) vai järjestysasteikollinen tai laatueroasteikollinen
- Esimerkiksi olkoon muuttuja, joka on koodattu lukuarvoin 1, 2 ja 3 siten että 1 = "Eri mieltä", 2="Samapa tuo" ja 3="Samaa mieltä".

```
x <- c(1,2,2,3)
```

- laatueroasteikolliseksi

```
x_nom <- factor(x,levels=c(1,2,3),
           labels=c("Eri mieltä", "Samapa tuo","Samaa mieltä"))
x_nom
```

```
## [1] Eri mieltä   Samapa tuo   Samapa tuo   Samaa mieltä
## Levels: Eri mieltä Samapa tuo Samaa mieltä
```

- järjestysasteikolliseksi

```
x_ord <- factor(x,levels=c(1,2,3),
           labels=c("Eri mieltä", "Samapa tuo","Samaa mieltä"),
           ordered=TRUE)
x_ord
```

```
## [1] Eri mieltä   Samapa tuo   Samapa tuo   Samaa mieltä
## Levels: Eri mieltä < Samapa tuo < Samaa mieltä
```

Muuttujat

Muuttuja ja arvo

- Muuttujalla viitataan mitattavan kohteen ominaisuuteen, joka vaihtelee yksiköstä tai mittauksesta toiseen. Esimerkiksi henkilön pituus, paino ja poliittinen kanta ovat muuttujia.
- Muuttujan arvo on muuttujan mittaustulos tietyssä yksikössä. Esimerkiksi henkilön pituuden arvo voi olla 170 cm ja poliittisen kannan arvo voi olla vasemmisto.
- Muuttujien arvot syntyvät mittaamisen tuloksena. Arvojen vaihtelusta syntyy muuttujan jakauma, jota voidaan kuvata tilastollisilla tunnusluvuilla ja tilastollisen grafiikan eli kuvaajien avulla.

Aineisto ja havaintomatriisi

Aineisto ja havaintomatriisi

- Tarkastellaan seuraavaksi aineiston ja havaintomatriisin käsitteitä.
- Seuraaviin perussääntöihin on olemassa poikkeuksia, sillä kaikkia tietoja ei voi esittää sujuvasti taulukkomuodossa.
- Taulukkomuotoinen esitystapa kuitenkin toimii tutkimustarkoituksiin erittäin usein.

Aineisto ja havaintomatriisi

- Kun samoista tilastoyksiköistä koostetaan useita muuttujia yhteen saadaan aineisto.
- Yleensä aineisto esitetään havaintomatriisin muodossa.
  - Havaintomatriisin kukin sarake sisältää yhden muuttujan arvot.
  - Havaintomatriisin kukin rivi sisältää mittaustiedot yhdeltä tilastoyksiköltä.
  - Taulukon solu sisältää muuttujan arvon.

sukupuoli	ikä	pituus	paino	pääaine
mies	27	194	80	TTRA2
mies	26	170	67	TTRA2
nainen	23	165	47	TILTK
mies	17	170	61	TK1K
nainen	25	168	50	TILTK

Aineistosta tarkemmin

- Aineistossa on hyvä olla yksi sarake, joka yksilöi tilastoyksiköt.
  - Edelliseltä kalvolta tämä puuttui.
- Jos aineistossa on useita mittauksia samalta tilastoyksiköltä, niin silloin aineistossa voi olla useita rivejä.
  - R-kieltä käytettäessä tämä on suositeltavampi tapa muodostaa aineisto vs. että aineistossa olisi useita sarakkeita eri mittauskertoja varten.
  - Leikitään ajatuksella, että aineistossa on kaksi mittausta jokaiselta henkilöltä vuoden välein

Toistomittausaineisto:

id	sukupuoli	ikä	paino
1	mies	27	80
1	mies	28	82
2	mies	26	67
2	mies	27	68
3	nainen	23	47
3	nainen	24	47
4	mies	17	61
4	mies	18	59
5	nainen	25	50
5	nainen	26	51

Aineiston käyttöönotto R-kielellä

- Tarkemmin ja yleisemmin aineiston käyttöönottoa opetellaan kurssilla R-kieli, 2op.
- Tällä kurssilla tarvittavat aineistot saa käyttöön seuraavasti:
  - Lataa aiemmin asentamasi paketti `datas4uef` komennolla `library(datas4uef)`
  - Sitten komentoa data käyttäen ja tietämällä aineiston nimen saat aineiston käyttöön. Esim. `data(hlotsim_dat)`.

```
library(datas4uef)
data(hlotsim_dat)
```

Objektien tyypit

- Objekteja on eri tyyppisiä, joista tärkeimmät ovat `numeric`, `character` ja `factor`
  - `numeric` ilmaisee välimatka- ja suhdeasteikollisia muuttujia.
  - `factor` ilmaisee luokitteluasteikollisia ja järjestysasteikollisia muuttujia
  - `character` on tekstimuotoinen tieto, joka voidaan muuttaa `numeric` tai `factor`-tyyppiseksi.
- Aineiston objektityyppi on data.frame tai tibble (uudempi)
- Objektin tyypin voi tarkastaa funktiolla `typeof`, esim. `typeof(hlotsim_dat)`
  - Jos objektin tyyppi on `data.frame` eli aineisto niin yhden muuttujan poiminta onnistuu $-merkin avulla (aineiston_nimi$muuttuja):

```
hlotsim_dat$ikä
```

```
##  [1] 27 26 23 17 25 28 20 21 33 32 25 25 21 15 19 25 27 29 20 19 20
## [26] 24 16 26 20 26 23 24 20 31 27 24 25 26 24 28
```

Vektorit

- Aineistot ovat siis data.frame-tyyppisiä tai vastaavia.
- Aineiston sarakkeissa olevat muuttujat ovat R:ssä vektoreita.
  - Vektoreita voi luoda itse `c()` -komennolla, esim. `c(1,5,6)` luo vektorin, jossa on luvut 1, 5, ja 6.
  - Aineiston sarakkeiden ei tarvitse olla keskenään samaa tyyppiä.
- Uuden sarakkeen lisääminen aineistolle `dat` onnistuu `dat$uusi_sarake <- c(1,2,3)`
  - Huomaa, että sijoitusoperaattorin oikealla puolella on oltava sama määrä lukuja kuin data.frame:ssa on rivejä. Rivimäärän saat komennolla `nrow(dat)`
- Saraketta voi muokata ylikirjoittamalla vanhan sarakkeen tiedot. Esim. kerrotaan sarakkeen arvot kahdella `dat$vanha_sarake <- 2*dat$vanha_sarake`
  - Tässä on oltava huolellinen, sillä muutoksia ei voi peruuttaa.

Jakaumat

Jakaumat

- Jakaumia on kahdenlaisia:
  - Empiirinen jakauma eli aineistosta laskettu jakauma
  - Teoreettinen jakauma eli todennäköisyysjakauma

Empiirinen jakauma kuvaa, mitä arvoja muuttuja saa aineistossa ja miten yleisiä mitkäkin arvot ovat. Se voidaan esittää taulukkona tai kuvaajan avulla. Sen tietoa voidaan myös kuvailla tunnusluvuilla.

Todennäköisyysjakauma ilmaisee satunnaismuuttujan arvojen todennäköisyydet. Esim. normaalijakauma (tuttu lukiosta) on todennäköisyysjakauma. Todennäköisyysjakaumiin palataan myöhemmin.

- Tosimaailmaa koskevassa tutkimuksessa ei tyypillisesti tunneta todellista todennäköisyysjakaumaa, vaan sitä pyritään ns. estimoimaan (eli tekemään päätelmiä siitä) aineiston avulla.

Empiirinen jakauma

Diskreetille aineistolle (laatuero- ja järjestysasteikko) voidaan laskea havaintojen lukumäärät kutakin mahdollista arvoa kohti.

Aineisto:

id	sukupuoli
1	mies
2	nainen
3	nainen
4	nainen
5	mies
6	nainen
7	mies

Frekvenssijakauma:

sukupuoli	lkm
mies	3
nainen	4

Prosenttijakauma:

sukupuoli	%
mies	42.9
nainen	57.1

Frekvenssi- ja prosenttijakauma R-kielellä

```
# frekvenssijakauma
tab <- table(dat$sukupuoli)
```

```
tab
```

```
##
##   mies nainen
##      3      4
```

```
# prosenttijakauma
pros_jakauma <- round(prop.table(tab)*100,1)
```

```
pros_jakauma
```

```
##
##   mies nainen
##   42.9   57.1
```

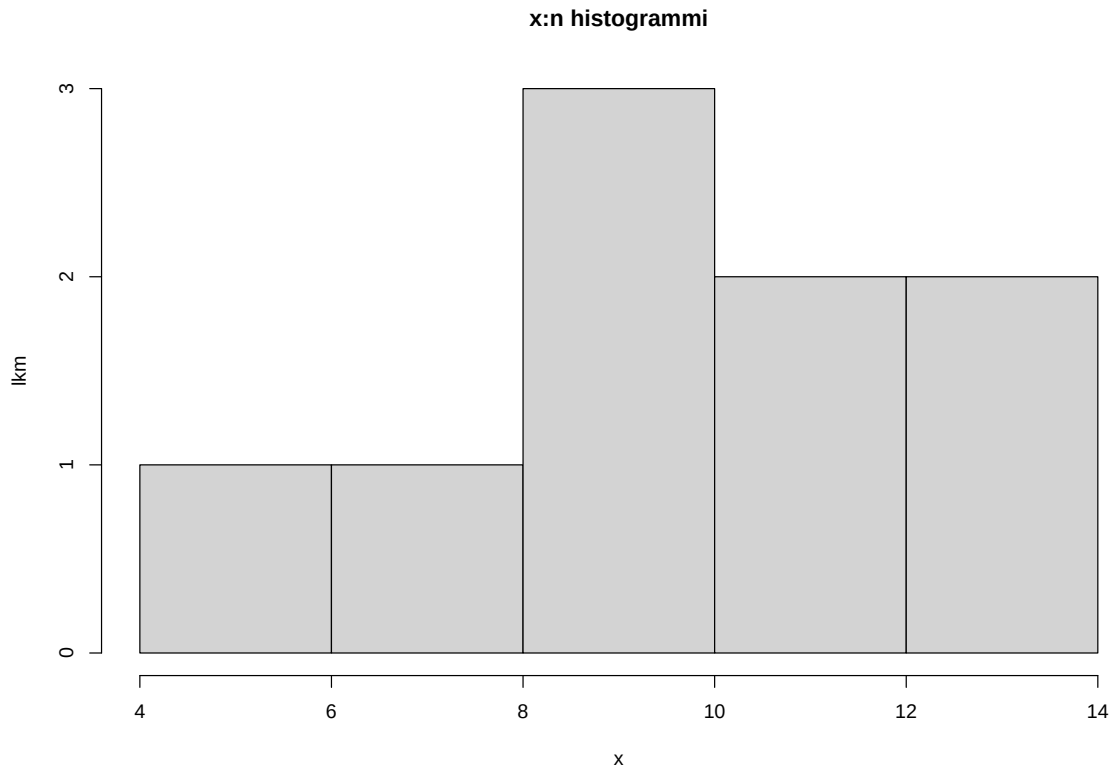
Empiirinen jakauma

- Jatkuvalle muuttujalle arvoja ei kannata taulukoida, sillä taulukosta saattaa tulla liian pitkä luettavaksi.
- Yksi vaihtoehto on luokitella jatkuvan muuttujan aineisto ja esittää se taulukon tai kuvan avulla.
  - Tässä saatetaan kuitenkin menettää arvokasta informaatiota.
- Taulukon sijaan voidaan myös käyttää tunnuslukuja (engl. statistics) kyseisen jakauman kuvailemiseksi.

Aineisto:

id	x
1	5.12
2	6.47
3	8.38
4	8.72
5	9.39
6	10.41
7	10.92
8	12.07
9	13.79

- Luokitellaan väleille 4-6, 6-8, 8-10, 10-12, 12-14.
  - Luokan alaraja ei kuulu luokkaan, yläraja kuuluu (puoliavoin väli).

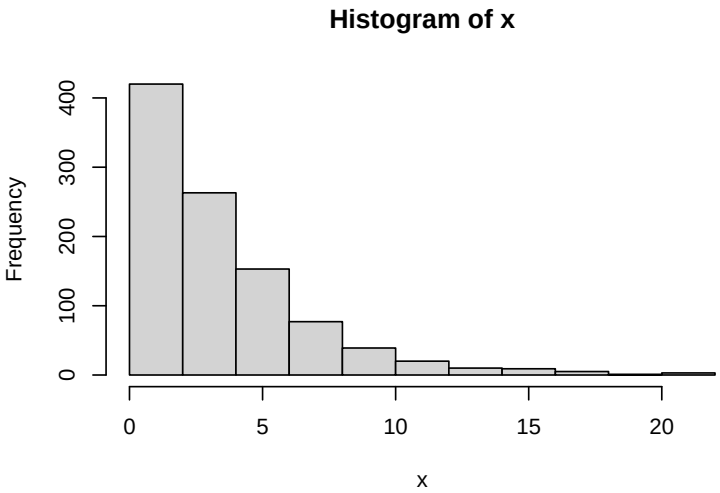


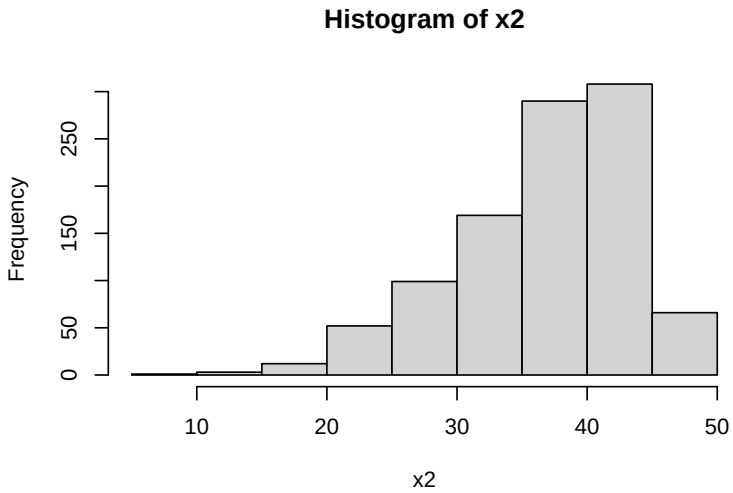
Histogrammin tulkintaa

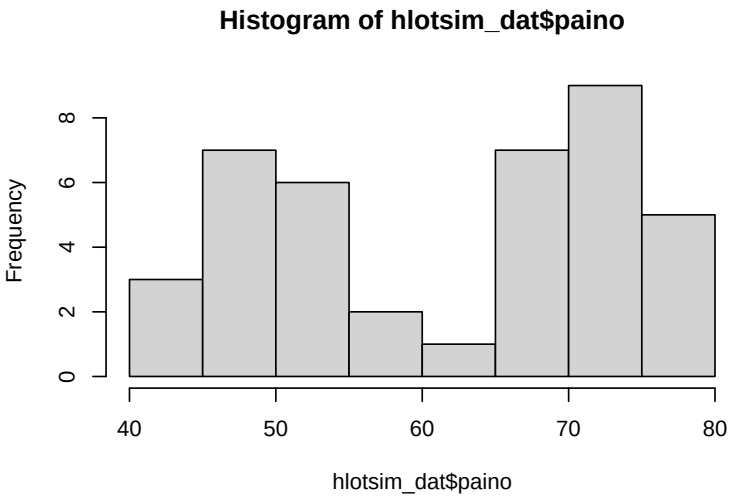
Histogrammista voidaan tulkita ainakin seuraavat seikat (ks. seuraavia kalvoja):

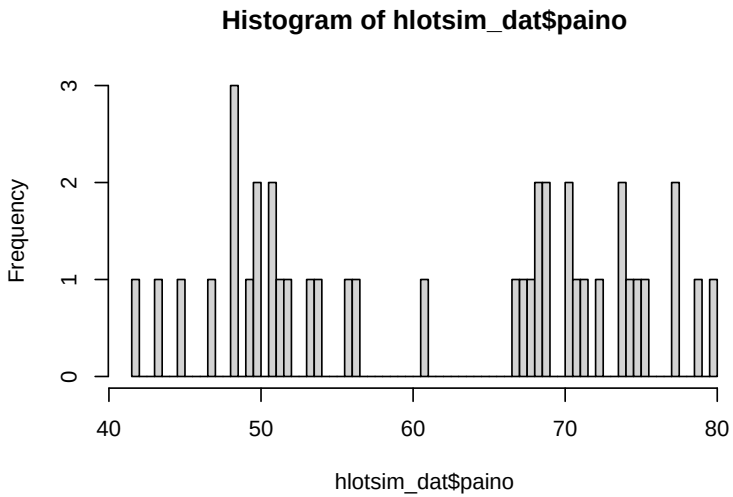
- vinous oikealle tai vasemmalle
- symmetrisyys (vastakkainen vinoudelle)
- monihuippuisuus
- valitun pylväiden määrän vaikutus kuvaajaan
- oudokki eli poikkeava havainto

Vinous

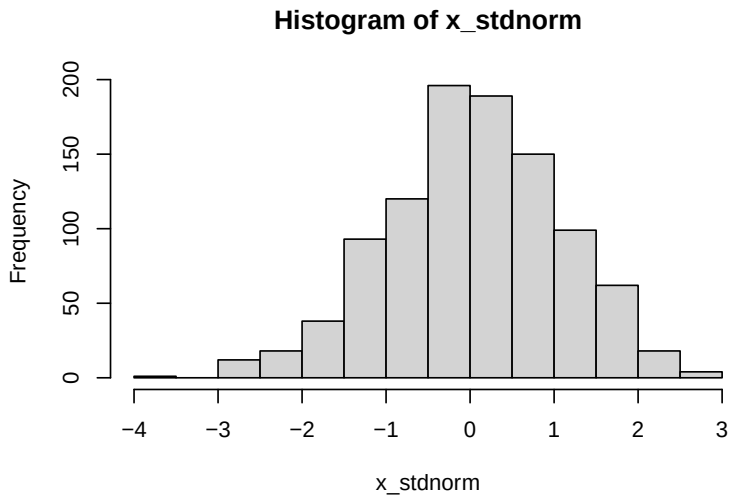

Histogram of x
Frequency
0
100
200
300
400
0
5
10
15
20
x

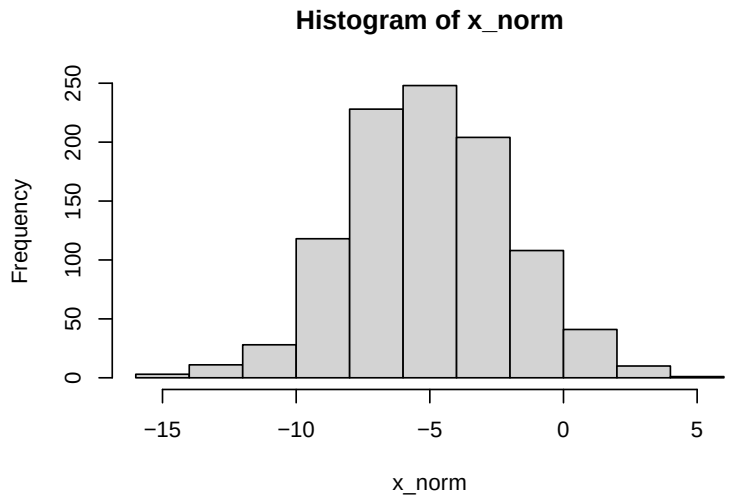

Histogram of x2
Frequency
0
50
150
250
10
20
30
40
50
x2

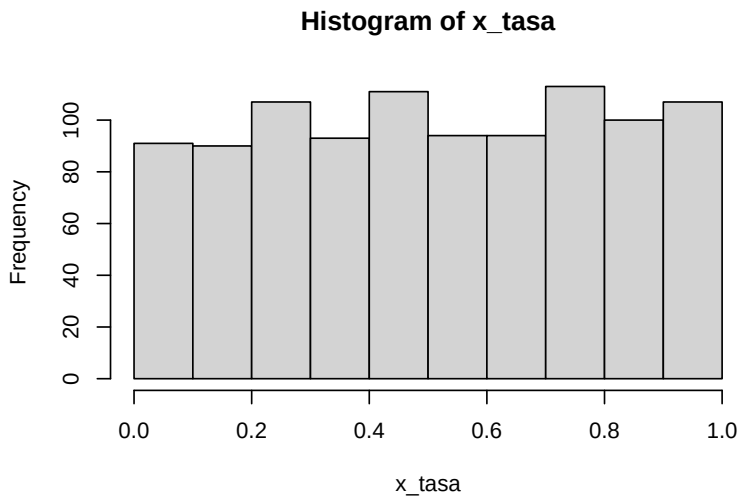

Histogram of hlotsim_dat$paino
Frequency
0
2
4
6
8
40
50
60
70
80
hlotsim_dat$paino

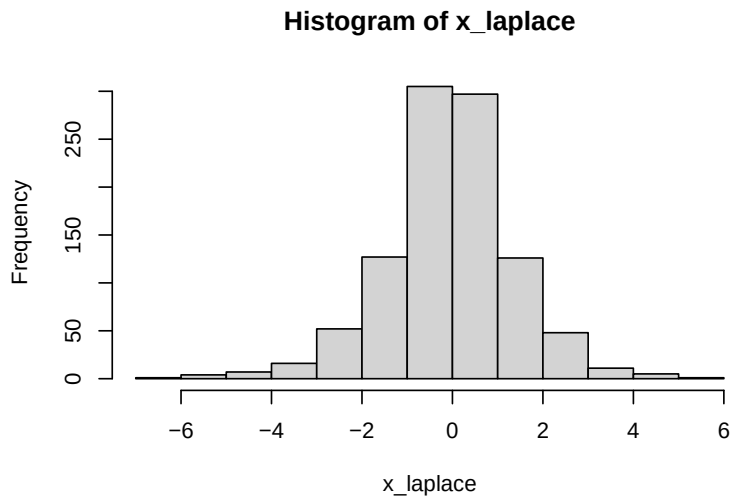

Histogram of hlotsim_dat$paino
Frequency
0
1
2
3
40
50
60
70
80
hlotsim_dat$paino

Symmetrisyys

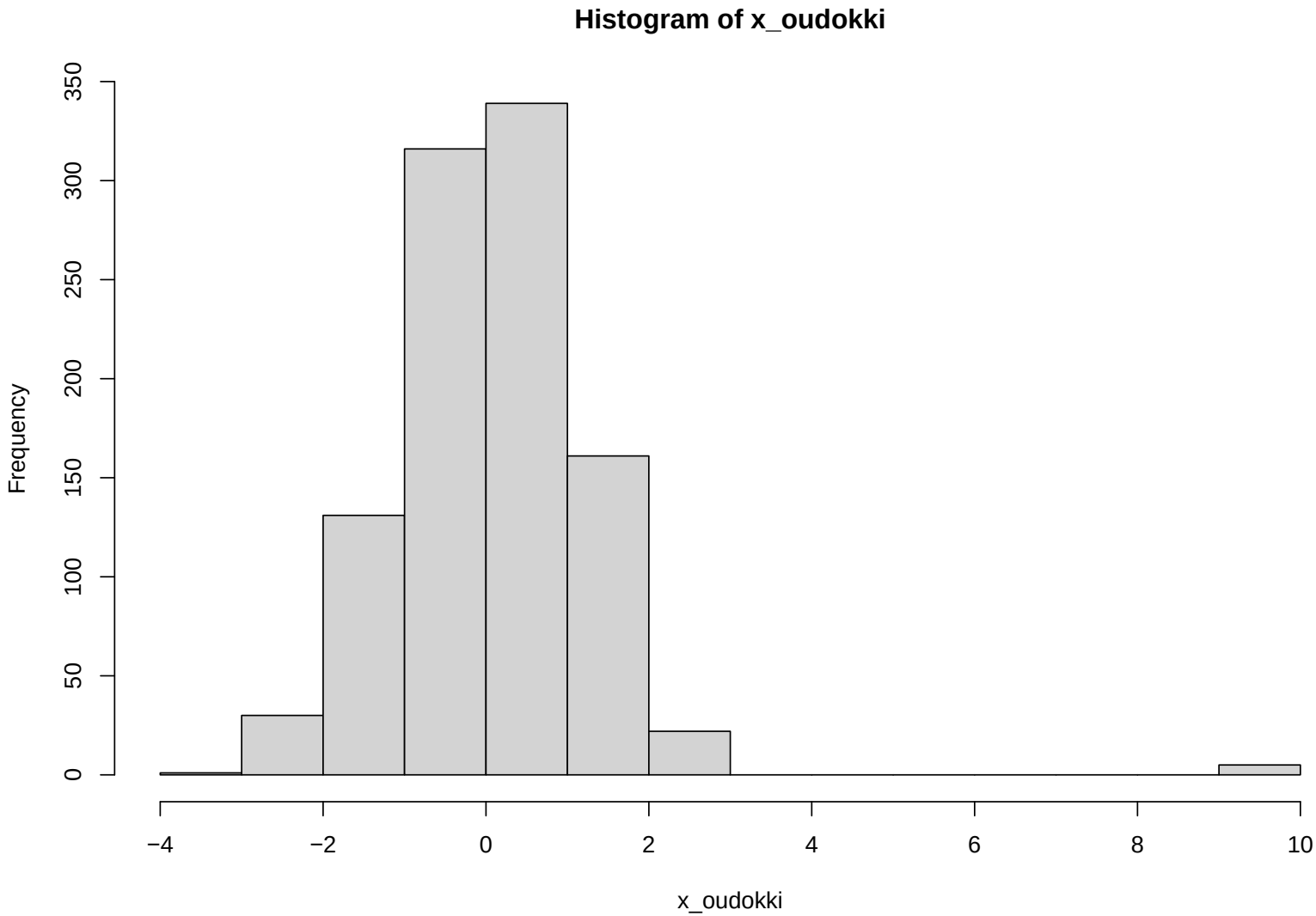








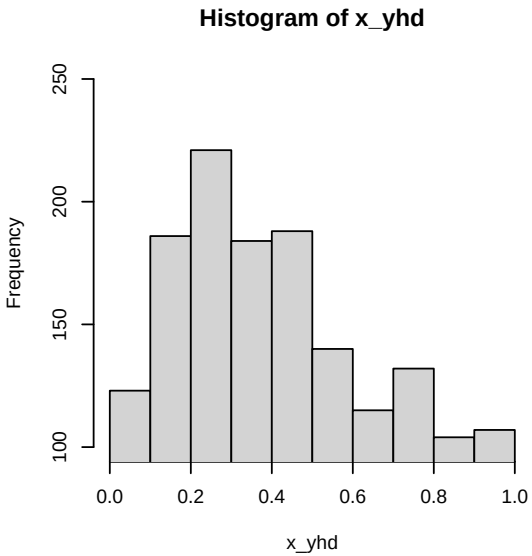
Poikkeava havainto eli oudokki

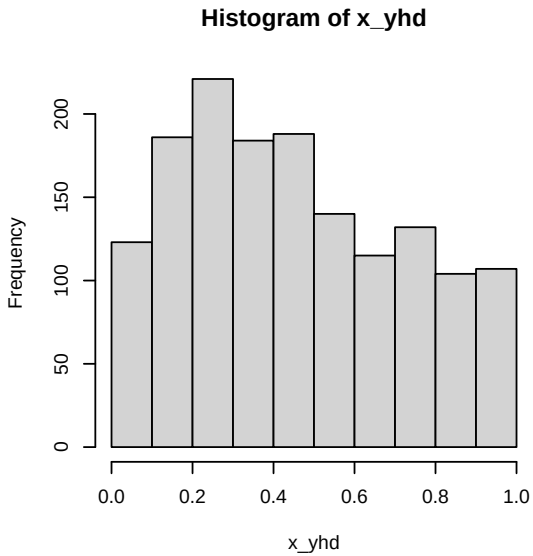


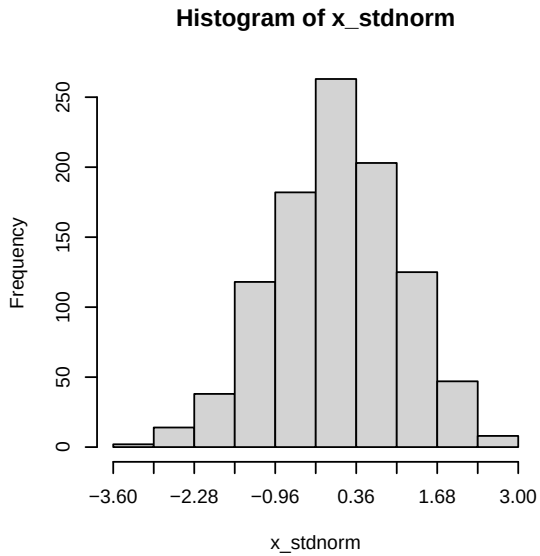
Miten histogrammia ei kannata tehdä

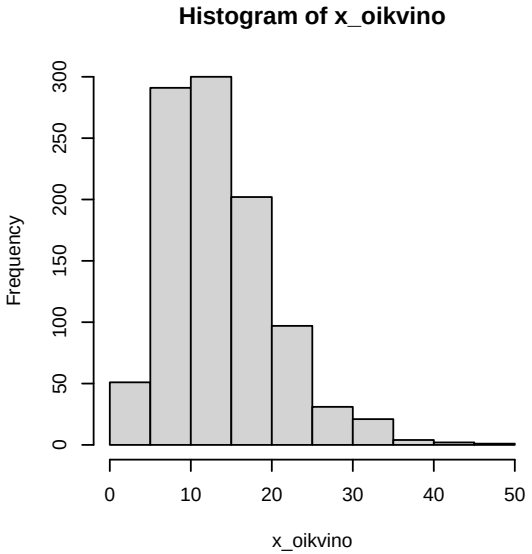
Ei kannata

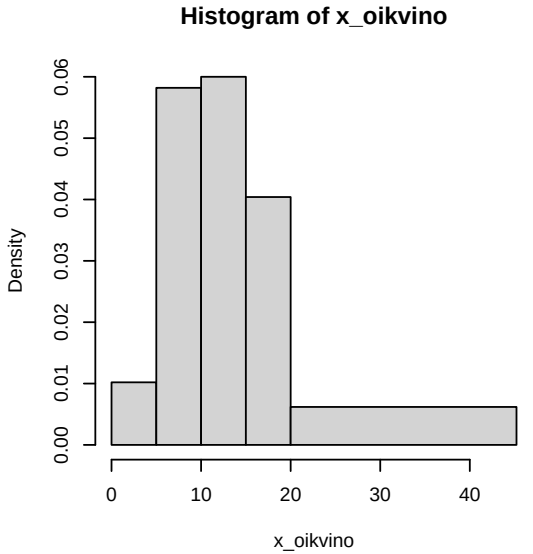
- valita y-akselia siten että nolla ei ole x-akselin risteymässä (osa histogrammia leikkautuu pois)
- valita vaikeasti ymmärrettävää pylväsjakoa
- käyttää epätasavälistä pylväsjakoa


Histogram of x_yhd
Frequency
100
150
200
250
0.0
0.2
0.4
0.6
0.8
1.0
x_yhd


Histogram of x_yhd
Frequency
0
50
100
150
200
0.0
0.2
0.4
0.6
0.8
1.0
x_yhd


Histogram of x_stdnorm
Frequency
0
50
100
150
200
250
−3.60
−2.28
−0.96
0.36
1.68
3.00
x_stdnorm

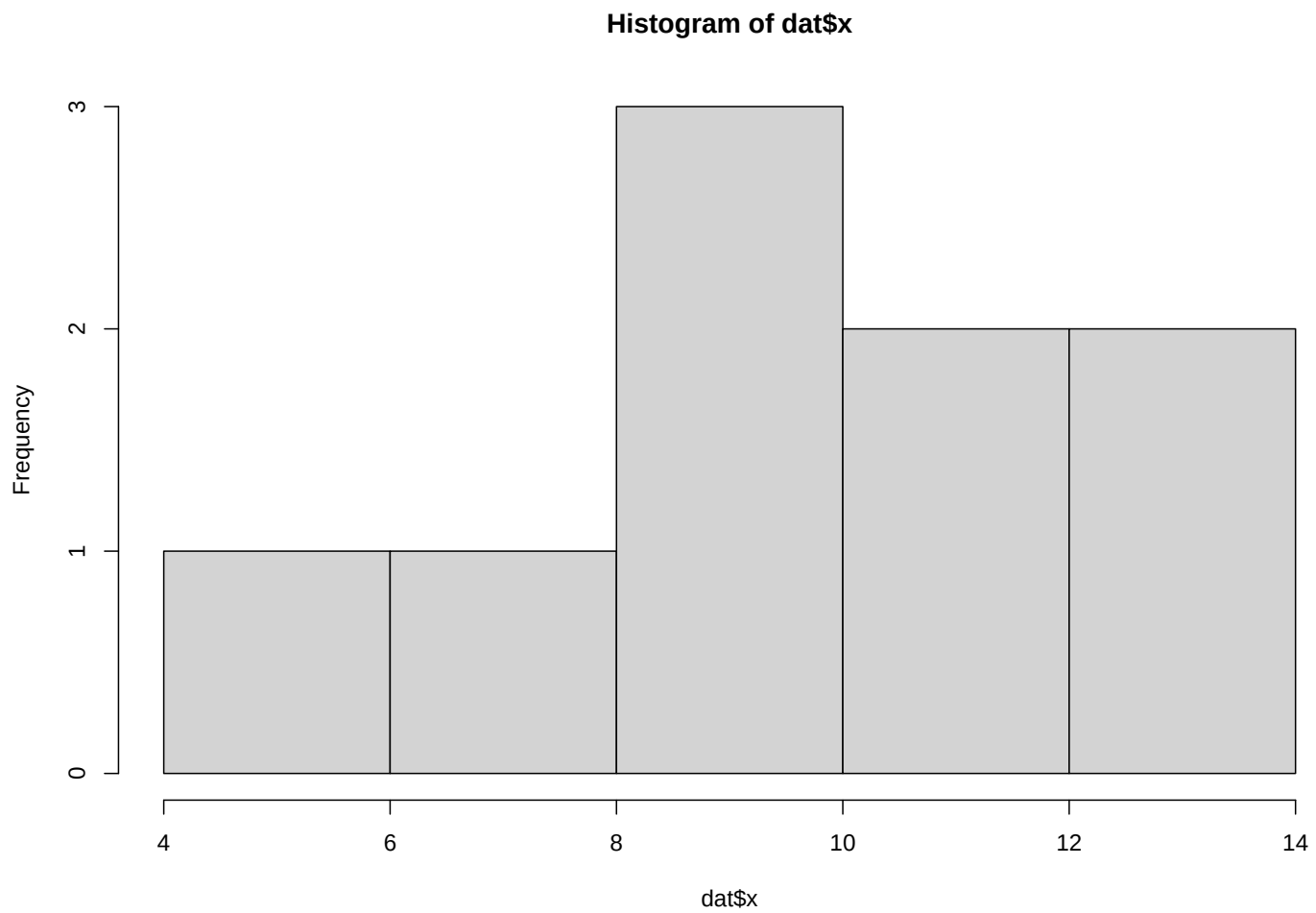

Histogram of x_oikvino
Frequency
0
50
100
150
200
250
300
0
10
20
30
40
50
x_oikvino


Histogram of x_oikvino
Density
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0
10
20
30
40
x_oikvino

Histogrammi R-kielellä

Histogrammi aineistosta `dat` muuttujalle `x` oletusasetuksin

```
hist(dat$x)
```



Muutetaan x- ja y-akseleiden selitteet sekä otsikko:

```
# histogrammi
hist(dat$x,xlab="x",ylab="lkm",main="x:n histogrammi")
```