Tilastotieteen johdantokurssi -luentokalvot

Yliopistonlehtori Juho Kopra

Itä-Suomen yliopisto, Tietojenkäsittelytieteen laitos

19.09.2023

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.

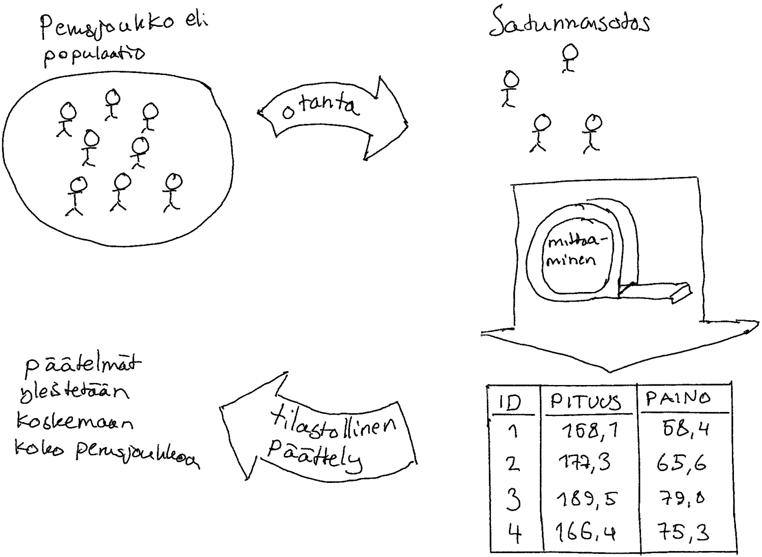
Sattuma ja satunnaisuus (jatkuu)

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	158,1	58,4
2	177,3	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

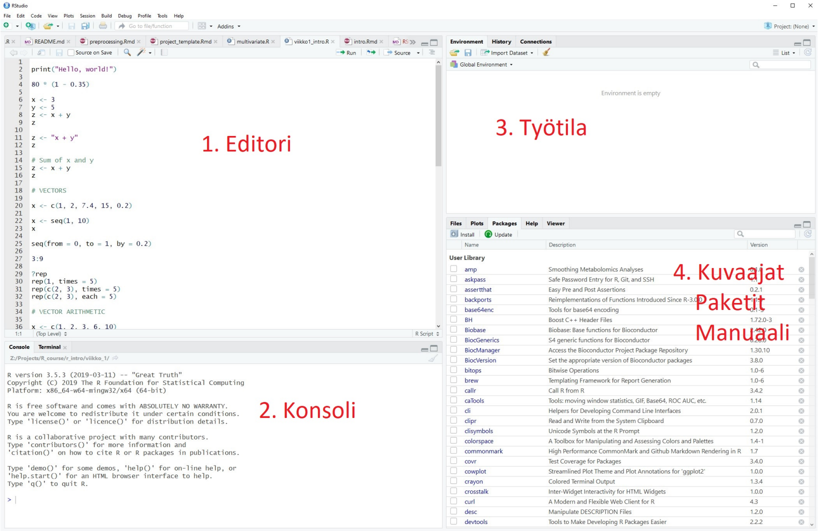
```r
x <- 1.5 # sijoittaa luvun 1.5 objektiin x
x # tulostaa x:n tiedot konsoliin
y <- 1:5 # sijoittaa luvut 1-5 objektiin y
y2 <- y*2 # y2 on kaksi kertaa y:n arvot
?sample # avaa sample-funktion ohjesivun
sample(y,2) # poimii 2 havainnon satunnaisotoksen y-objektista
```

- 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

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] "Jaska" "Niina"
```

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

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

Otanta R-kielellä jatk.

Yksinkertainen satunnaisotanta palauttaen

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

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

```r
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

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

```r
tab
```

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

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

```r
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

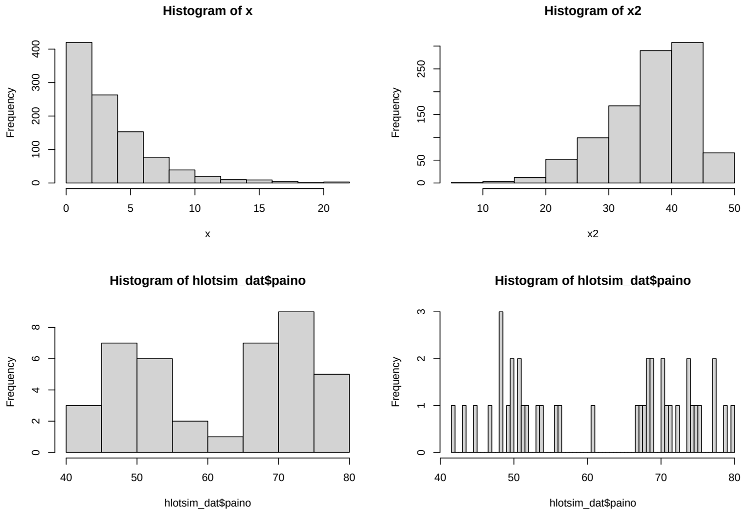
- Luokitellaan väleille 4-6, 6-8, 8-10, 10-12, 12-14.
    - Luokan alaraja ei kuulu luokkaan, yläraja kuuluu (puoliavoin väli).

Histogrammin tulkintaa

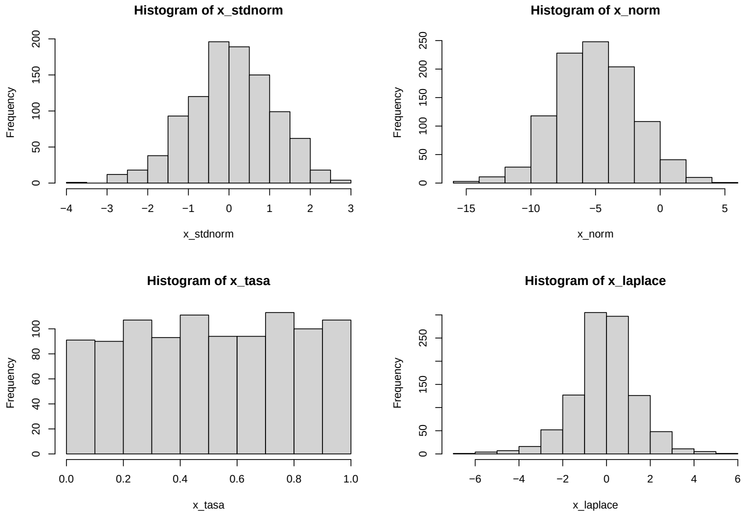
Histogrammista voidaan tulkita ainakin seuraavat seikat (ks. seuraavia kalvoja):

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

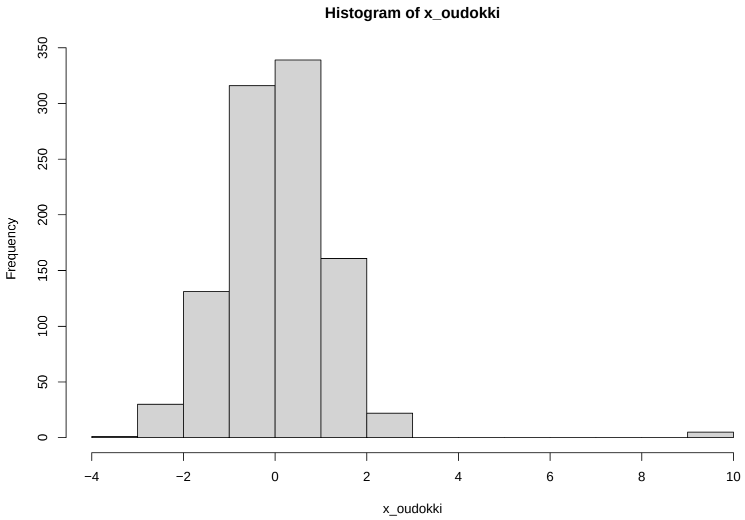
Vinous


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



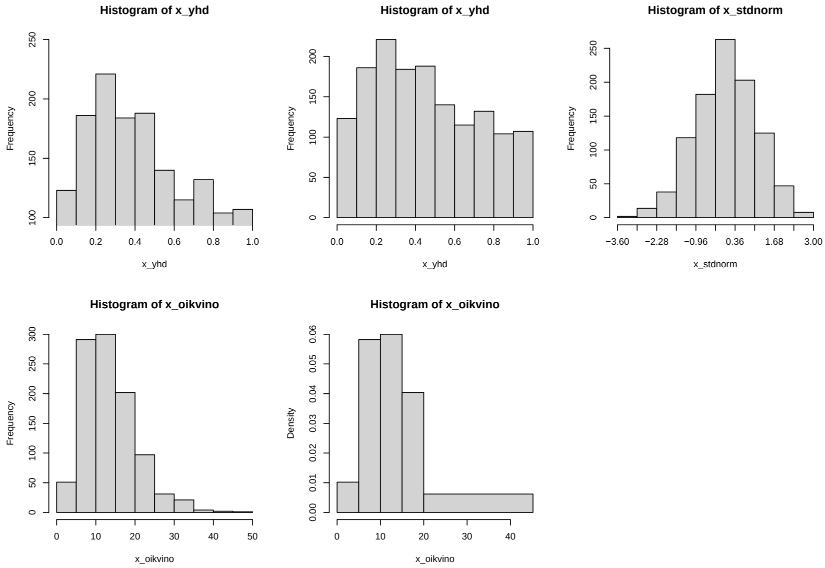
Poikkeava havainto eli oudokki



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

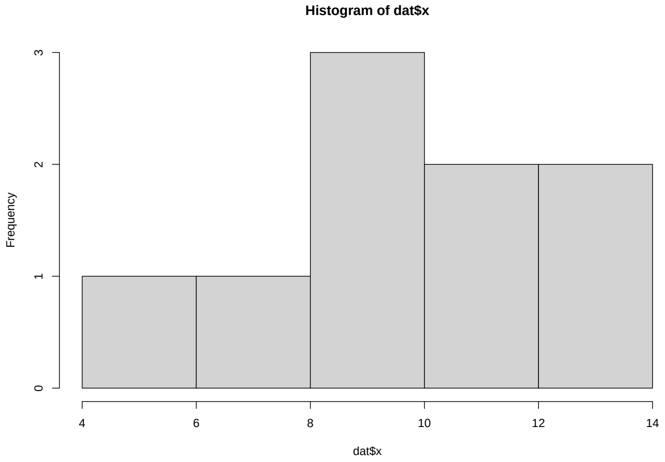


Histogram of x_yhd
Frequency
x_yhd
0.0 0.2 0.4 0.6 0.8 1.0
100 150 200 250
Histogram of x_yhd
Frequency
x_yhd
0.0 0.2 0.4 0.6 0.8 1.0
0 50 100 150 200
Histogram of x_stdnorm
Frequency
x_stdnorm
−3.60 −2.28 −0.96 0.36 1.68 3.00
0 50 100 150 200 250
Histogram of x_oikvino
Frequency
x_oikvino
0 10 20 30 40 50
0 50 100 150 200 250 300
Histogram of x_oikvino
Density
x_oikvino
0 10 20 30 40
0.00 0.01 0.02 0.03 0.04 0.05 0.06

Histogrammi R-kielellä

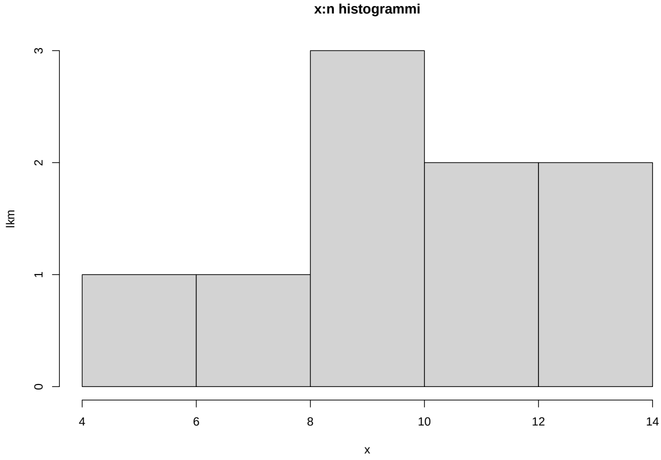
Histogrammi aineistosta `dat` muuttujalle `x` oletusasetuksin

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



Muutetaan x- ja y-akseleiden selitteet sekä otsikko:

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



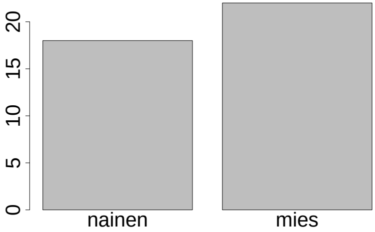
Kuvaajat

Kuvaajat

- Tutustutaan seuraaviin kuvaajiin tai kuvioihin:
  - pylväskuvaaja eli palkkikuvio
  - piirakkakuvio eli sektorikuvio
  - laatikkokuvio
  - histogrammi (käsiteltiin aiemmin)
  - sirontakuvio eli hajontakuvio
- Opetellaan tulkitsemaan ja piirtämään kyseiset kuviot
- Muita kuvioita voi opetella halutessaan osion viimeisen kalvon linkkien avulla

Pylväskuvio eli palkkikuvio

- luokitteluasteikollisen muuttujan frekvenssijakauman tai prosenttijakauman esittämiseen



Pylväskuvio R-kielellä

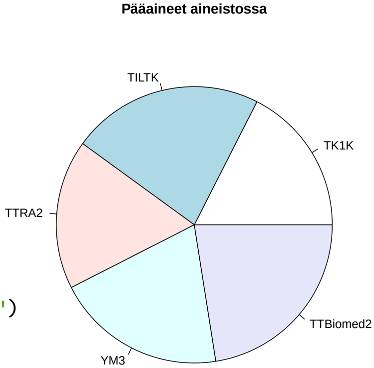
```r
barplot(table(hlotsim_dat$sukupuoli))
```

```r
# suhteelliset osuudet
barplot(prop.table(table(hlotsim_dat$sukupuoli)))
```

Piirakkakuvio

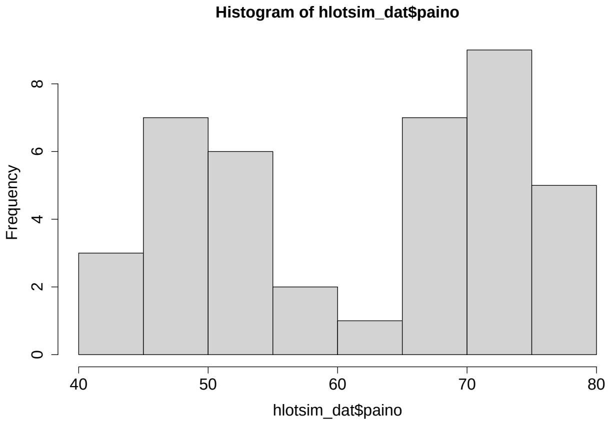
- luokitteluasteikollisen muuttujan lukumäärien tai osuuksien/prosenttien kuvaamiseen
- ei sovellu tieteelliseen esittämiseen
  - käyttöohje: Älä käytä!

```
lkm <- table(hlotsim_dat$pääaine)
pie(lkm,main="Pääaineet aineistossa")
```



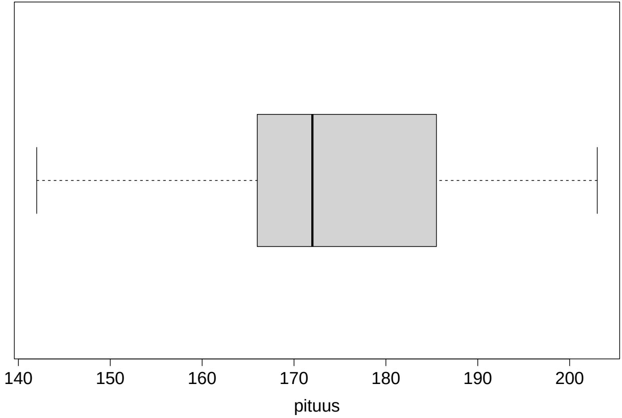
Histogrammi

- vähintään välimatka-asteikollinen muuttuja (diskreetti tai jatkuva)



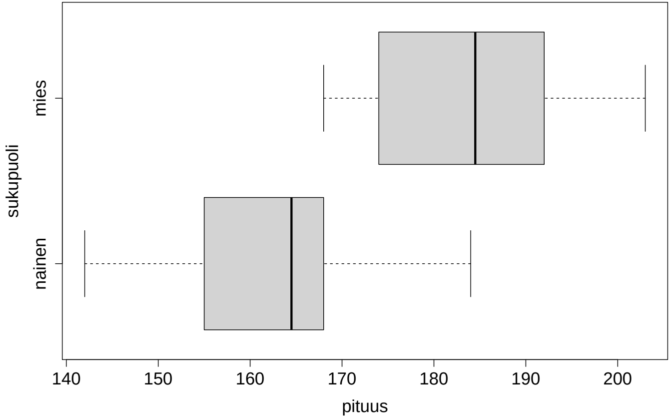
Laatikkokuvio

- vähintään välimatka-asteikollinen muuttuja (diskreetti tai jatkuva)
- voidaan piirtää pysty- tai vaakasuuntaisena




```
boxplot(hlotsim_dat$pituus,xlab="pituus",horizontal=TRUE)
```

Laatikkokuvio ryhmittäin

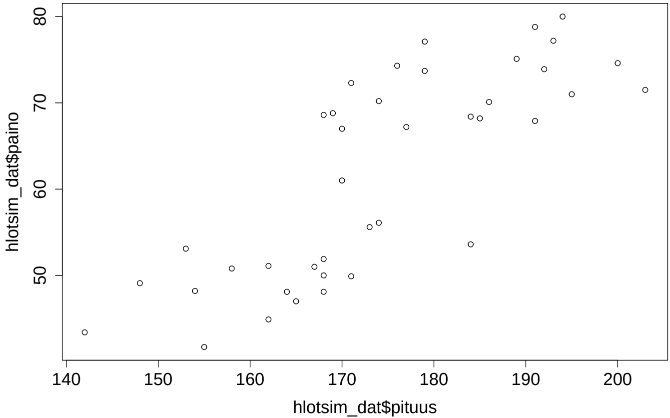


- Ao. koodissa `pituus ~ sukupuoli` on nk. formula, joka ilmaisee että tehdään haluttu asia (boxplot) pituudelle muuttujan sukupuoli suhteen
  - vaatii että aineisto on pitkässä muodossa, eli pituus oltava yhdessä sarakkeessa (ei eri sarakkeita eri sukupuolille)

```
boxplot(pituus ~ sukupuoli,data=hlotsim_dat,horizontal=TRUE)
```

Sirontakuvio eli hajontakuvio

- samoilta tilastoyksiköiltä kaksi vähintään välimatka-asteikollista muuttujaa (mittausparit)




```
plot(hlotsim_dat$pituus,hlotsim_dat$paino)
```

Lisätietoa kuvaajista

- Kuvaajia on lukuisia muitakin kuin edellä on esitetty. Ks. https://r-graph-gallery.com/
- Nykyaikainen tapa tehdä kuvaajia R-kielellä (ei käytössä tällä kurssilla): https://ggplot2-book.org/

Tilastolliset tunnusluvut

Tilastolliset tunnusluvut

- Edellä esitettyjen kuvaajien avulla saa nopeasti käsityksen aineiston luonteesta ja jakaumista.
- Tarkempaa analyysiä varten on järkevää laskea tilastollisia tunnuslukuja. Tunnuslukuja tarvitaan myös, jotta osataan piirtää kuvaajia.
- Tunnusluvut ovat keskeisiä myös myöhemmin, kun siirtytään tilastollisen päättelyn pariin.

Johdanto

- Tilastolliset tunnusluvut ovat lukuja, jotka kuvaavat tilastollista aineistoa tiivistetysti ja havainnollisesti.
  - Tunnusluvut siis kuvaavat jakauman ominaisuuksia.
- Tilastollisia tunnuslukuja ovat esimerkiksi keskiarvo, mediaani, moodi, varianssi, keskihajonta ja korrelaatio.
  - Näistä keskilukuja ovat keskiarvo, mediaani, moodi.
  - Hajontalukuja ovat varianssi ja keskihajonta.
- Muuttujan mitta-asteikosta riippuu, mitä tunnuslukuja muuttujalle voidaan laskea.
- Tilastollisia tunnuslukuja voidaan käyttää aineiston analysointiin, vertailuun ja esittämiseen.

Keskiarvo eli otoskeskiarvo

- Keskiarvo (aritmeettinen keskiarvo) kuvaa aineiston keskimääräistä arvoa.
- Vaatii vähintään välimatka-asteikollisen muuttujan, koska perustuu etäisyyksien laskemiseen (yhteen- ja vähennyslaskuun).
- Keskiarvo lasketaan jakamalla aineiston arvojen summa aineiston lukumäärällä.
- Keskiarvon merkintä on $\bar{x}$ ja laskukaava

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i .$$

- Termiä otoskeskiarvo käytetään tähdentämään että tarkoitetaan aineistosta eli otoksesta laskettua keskiarvoa.
- Keskiarvon vastine populaatiossa on odotusarvo μ.

Keskiarvo

- Esimerkki: Jos luokan oppilaiden pituudet ovat 160, 165, 170, 175 ja 180 cm, niin niiden keskiarvo on $\frac{160+165+170+175+180}{5} = 170$ cm.
- Keskiarvo on herkkä poikkeaville arvoille (oudokeille).
  - Esim. lisätään viimemainittuun aineistoon hlö, jonka pituus on 250 cm. Nyt keskiarvo on 183.3 cm.

Mediaani

- Mediaani on tilastollinen tunnusluku, joka kuvaa aineiston keskimmäistä arvoa. Se jakaa aineiston kahteen yhtä suureen osaan: puolet arvoista on mediaania pienempiä ja puolet suurempia.
- Voidaan laskea vähintään järjestysasteikolliselle muuttujalle.
  - Käytetään usein välimatka-asteikolliselle muuttujalle sen robustisuuden vuoksi.
- Mediaani lasketaan järjestämällä aineiston arvot suuruusjärjestykseen ja valitsemalla keskimmäinen arvo. Jos arvoja on parillinen määrä, mediaani on kahden keskimmäisen arvon keskiarvo.
- Mediaanin merkintä on M tai $Q2$.

Mediaani

- Esimerkki: Jos luokan oppilaiden pituudet ovat 160, 165, 170, 175 ja 180 cm, niin mediaani on 170 cm. Jos oppilaiden pituudet ovat 160, 165, 170, 175, 180 ja 185 cm, niin mediaani on $\frac{170+175}{2}$=172.5 cm.
- Poikkeavat havainnot eivät vaikuta herkästi mediaanin arvoon ts. mediaani on robusti tunnusluku.
    - Esim. lisätään viimemainittuun aineistoon hlö, jonka pituus on 250 cm. Nyt mediaani on 175 cm (aiemmin 172.5).

Moodi

- Moodi kertoo aineiston arvon, joka on yleisin.
- Moodi soveltuu laatueroasteikollisten muuttujien kuvailuun.
  - Esim. katsastusasemalla vieraili autoja päivässä seuraavasti. Moodiksi saadaan “Volkswagen”, koska sen frekvenssi on suurin.

```
## autot
##       Audi      Tesla Volkswagen
##          8          5         13
```

- Numeerisille muuttujille voidaan laskea moodiluokka esim. histogrammista. Moodiluokka on silloin se väli, jonka pylväs on korkein.
- Teoreettisille jakaumille on helpompi määrittää moodi jakauman kuvaajan korkeimman kohdan perusteella.

Keskiarvo, mediaani ja moodi R:ssä

Keskiarvo, mediaani ja moodi R:ssä

- Lasketaan aineiston `hlotsim_dat` muuttujan pituus keskiarvo ja mediaani käyttäen funktioita `mean(x)` ja `median(x)`

```
mean(hlotsim_dat$pituus)
```

```
## [1] 174.3
```

```
median(hlotsim_dat$pituus)
```

```
## [1] 172
```

- Moodin laskemiseksi ei ole valmista R-funktiota, vaan se tulee päätellä frekvenssitaulukon perusteella. Tehdään frekvenssitaulukko muuttujalle pääaine
  - Kaksi yhtä suurta suurinta frekvenssiä, eli moodeja ovat `TILTK` ja `TTBiomed2`

```
table(hlotsim_dat$pääaine)
```

```
##
##      TK1K     TILTK     TTRA2       YM3 TTBiomed2
##         7         9         7         8         9
```

Summamerkintä ja keskiarvon laskeminen

Johdanto

- Summamerkintä on matemaattinen symboli, jolla ilmaistaan lukujen summaa lyhyesti ja kätevästi.
- Summamerkinnässä käytetään kreikkalaista kirjainta sigma ($\sum$), jonka alapuolella on summauksen alkuarvo ja yläpuolella loppuarvo. Summattavat luvut ilmaistaan yleensä indeksoidulla muuttujalla, jonka indeksi kasvaa yhdellä joka askeleella.
- Esimerkiksi summamerkintä

$$\sum_{i=1}^{5} i$$

tarkoittaa lukujen 1, 2, 3, 4 ja 5 summaa eli 15.
  - Sama summamerkintä voi saada myös muodon $\sum_{i=1}^{5} i$

Keskiarvon laskeminen summamerkinnällä

- Keskiarvon laskemiseen voidaan käyttää summamerkintää apuna. Jos lukujoukon luvut ovat $x_1, x_2, ..., x_n$, niin niiden keskiarvo on $\frac{1}{n}\sum_{i=1}^{n} x_i$.
- Esimerkiksi lukujen 2, 4, 6 ja 8 keskiarvo on

$$\frac{1}{4}\sum_{i=1}^{4} x_i = \frac{1}{4}(2+4+6+8) = \frac{20}{4} = 5$$

.

Hajontaluvut

Johdanto

- Hajontaluvut ovat tilastollisia tunnuslukuja, jotka kuvaavat aineiston hajontaa eli vaihtelua.
- Hajontalukuja ovat esimerkiksi varianssi, keskihajonta, kvartiilit, kvartiiliväli ja variaatiokerroin.
- Hajontalukuja voidaan laskea numeeriselle aineistolle
  - Kvartiilit voidaan laskea myös järjestysasteikolliselle (harvoin käytetty)
- Hajontalukuja voidaan käyttää aineiston vaihtelun arvioimiseen ja vertailuun.

Varianssi

- Varianssi kuvaa aineiston arvojen tyypillistä neliöpoikkeamaa keskiarvosta.
- Varianssi lasketaan neliöimällä aineiston arvojen poikkeamat keskiarvosta ja jakamalla havaintojen määrä vähennettynä yhdellä.
- Otosvarianssin merkintä on s^2 ja sitä vastaava populaation varianssi on σ^2.
- Otosvarianssin laskukaava on

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

missä n on aineiston havaintojen lkm, x_i käy läpi aineiston havainnot ja $\bar{x}$ on keskiarvo.

Varianssi

- Esimerkki: Jos luokan oppilaiden pituudet ovat 160, 165, 170, 175 ja 180 cm, niin varianssi on $\frac{(160-170)^2+(165-170)^2+(170-170)^2+(175-170)^2+(180-170)^2}{5-1} = 62.5$ cm^2.
  - Huomaa mittayksikön muutos! cm $\rightarrow$ cm^2 Siksi varianssin lukuarvoa on vaikea tulkita.

Keskihajonta

- Varianssin tulkinnan vaikeudesta päästään eroon ottamalla varianssista neliöjuuri: $\sqrt{x \text{ cm}^2} = \sqrt{x}$ cm. Näin saadaan keskihajonta.
- Keskihajonta on kuvaa aineiston arvojen keskimääräistä poikkeamaa keskiarvosta.
  - Mittayksikkö on sama kuin alkuperäisillä havainnoilla.
- Keskihajonnan merkintä on s ja populaation keskihajonta on σ.
- Esimerkki: Jos luokan oppilaiden pituudet ovat 160, 165, 170, 175 ja 180 cm, niin keskihajonta on $\sqrt{62.5} \approx 7.91$ cm.

Kvartiilit

- Kvartiilit ovat lukuja, jotka jakavat aineiston neljään yhtä suureen osaan. Ne ovat mediaanin kaltaisia tunnuslukuja.
- Perusidea kvartiilien laskemisessa on: järjestetään aineiston arvot suuruusjärjestykseen ja valitaan neljännesvälien kohdalta arvot. Kvartiilien tarkkaa laskemista varten on useita tapoja (R:ssä on 9 eri tapaa) ja eri ohjelmistot käyttävät eri tapoja. Käsitellään laskemista myöhemmin.
- Kvartiileja merkitään Q_1, Q_2 ja Q_3.
- Esimerkki: Jos luokan oppilaiden pituudet ovat 160, 165, 170, 175 ja 180 cm, niin ensimmäinen kvartiili on laskutavasta riippuen 162.5 cm tai 165 cm, toinen kvartiili eli mediaani on 170 cm ja kolmas kvartiili on 177.5 cm tai 175 cm.

Kvartiiliväli

- Kvartiiliväli kertoo millä välillä aineiston keskimmäinen 50 prosenttia havainnoista sijaitsee.
  - Kvartiilivälistä voidaan laskea kvartiilivälin pituus, joka on kolmannen ja ensimmäisen kvartiilin erotus.
- Kvartiilivälin merkintä on (Q1,Q3) ja kvartiilivälin pituus on IQR (interquartile range).
- Esimerkki: Jos luokan oppilaiden pituudet ovat 160, 165, 170, 175 ja 180 cm, niin ensimmäinen kvartiili on 162.5 cm ja kolmas kvartiili on 177.5 cm. Kvartiiliväli on (162.5, 177.5) ja IQR $=$ 177.5-162.5 $=$ 15 cm.

Varianssin ja keskihajonnan laskeminen

Tulkintaa

- Varianssi on aina ei-negatiivinen eli $s^2 \geq 0$. Mitä suurempi varianssi on, sitä enemmän havainnoissa on hajontaa eli vaihtelua. Jos varianssi on nolla, kaikki luvut ovat yhtä suuria kuin keskiarvo.
- Keskihajonta kuvaa siis havaintojen vaihtelua keskiarvon ympärillä. Mitä pienempi keskihajonta on, sitä tiiviimmin luvut ovat ryhmittynyt keskiarvon ympärille. Mitä suurempi keskihajonta on, sitä laajemmin luvut ovat levittäytyneet keskiarvosta poispäin.

Esimerkki

- Esimerkiksi lukujen 2, 4, 6 ja 8 varianssin ja keskihajonnan laskemiseksi aloitetaan laskemalla niiden keskiarvo: $\overline{x} = \frac{1}{4}\sum_{i=1}^{4} x_i = \frac{1}{4}(2+4+6+8) = \frac{20}{4} = 5.$
- Sitten lasketaan jokaisen luvun ja keskiarvon erotuksen neliö: $(2-5)^2 = (-3)^2 = 9$, $(4-5)^2 = (-1)^2 = 1$, $(6-5)^2 = 1^2 = 1$, $(8-5)^2 = 3^2 = 9$
- Summataan neliöt yhteen ja jaetaan lukujen määrä vähennettynä yhdellä $n-1$:

$$\frac{9+1+1+9}{4-1} = \frac{20}{3} \approx 6.67$$

- Keskihajonta on neliöjuuri (R:ssä `sqrt(x)`) varianssista: $\sqrt{\frac{20}{3}} = \frac{2\sqrt{5}}{\sqrt{3}} \approx 2.58$

Varianssi ja keskihajonta R:ssä

- Lasketaan aineiston `hlotsim_dat` muuttujan pituus varianssi ja keskihajonta

```
var(hlotsim_dat$pituus)
```

```
## [1] 211.959
```

```
sd(hlotsim_dat$pituus)
```

```
## [1] 14.55881
```

Kvantiilit

Kvantiilit

- Kvantiilit (eli fraktiilit) ovat hajontalukuja, jotka jakavat aineiston tiettyyn määrään yhtä suuria osia. Ne ovat yleistys mediaanille ja kvartiileille, jotka ovat itsessään myös kvantiileja.
- Kvantiilit lasketaan järjestämällä aineiston arvot suuruusjärjestykseen ja valitsemalla tiettyjen osuuksien kohdalta arvot. Jos osuus ei osu tarkalleen yhteen arvoon, voidaan käyttää interpolaatiota eli laskennallista väliarvoa.
  - Kvantiilien määrittämiseksi on useita eri menetelmiä (samat kuin kvartiileille). Esimerkiksi R-kielessä kvantiilit voidaan laskea yhdeksällä eri tavalla.
- Kvantiilien merkintöjä ovat Q_k, missä k on osuus desimaalimuodossa.
- Esimerkki: Jos luokan oppilaiden pituudet ovat 160 cm, 165 cm, 170 cm, 175 cm ja 180 cm, niin 0.25-kvantiili eli ensimmäinen kvartiili on 162.5 cm ja 0.75-kvantiili eli kolmas kvartiili on 177.5 cm.

Kvantiilien laskeminen käsin: merkintöjä

Olkoon meillä suuruusjärjestykseen asetettu aineisto (n=20), jossa pienin arvo 1 ja suurin arvo 19.

```
## 1.  2.  3.  4.  5.  6.  7.  8.  9. 10. 11. 12. 13. 14. 15. 16. 17. 18. 19. 20.
##  1   4   4   7   8   9  10  10  11  11  13  14  14  14  14  15  15  18  18  19
```

Käytetään seuraavilla kalvoille merkintöjä

- $x_{(j)}$ = “j:nneksi pienin arvo”
- $\lfloor m \rfloor$: luku m pyöristettynä alaspäin lähimpään kokonaislukuun

Kvantiilien laskeminen (teoria)

p-kvantiili voidaan laskea funktiolla

$$Q_i(p) = (1-\gamma)x_{(j)} + \gamma x_{(j+1)}$$

missä

- γ määrittää miten paljon painotetaan lukuja $x_{(j)}$ ja $x_{(j+1)}$
- i vastaa R-funktion `quantile(x,probs,type)` `type`-argumenttia. Oletuksena R:ssä `type=7`.

p-kvantiili on siis painotettu summa kahdesta luvusta, jotka määräytyvät luvun $j = \lfloor np + m \rfloor$ perusteella. Luvun j laskemiseksi tarvittava luku m riippuu valitusta arvosta i (`type`). Oletusarvolla `type`$=$7 jolloin luku m saadaan kaavalla $m = 1 - p$.

Kvantiilien laskeminen käsin (type=7)

- Halutaan laskea p-kvantiili. Esim. jos lasketaan alakvartiili, niin $p = 0.25$. Olkoon n aineiston havaintojen lukumäärä, tässä $n = 20$.
- Lasketaan apuparametri γ. Jos halutaan laskea R:n käyttämää oletusparametrisointia käyttäen (type=7), niin $m = 1 - p = 1 - 0.25 = 0.75$.
- Koska $n = 20$ ja $p = 0.25$ niin nyt $j = \lfloor 20 \cdot 0.25 + 0.75 \rfloor = \lfloor 5.75 \rfloor = 5$.
- Nyt γ saadaan siten että lasketaan ensin $g = np + m - j$, mistä saadaan $g = 20 \cdot 0.25 + 0.75 - 5 = 0.75$. Jos type = 4,5,6,7,8 tai 9, niin $\gamma = g$.

Eli tässä tapauksessa

$$Q_7(0.25) = (1 - 0.75)x_{(5)} + 0.75x_{(6)} = 0.25 \cdot 8 + 0.75 \cdot 9 = 8.75$$

Kvantiilien laskeminen R-kielellä

```r
# halutut luvut vektorissa (ei tarvitse olla järjestetty)
luvut <- c(1,4,4,7,8,9,10,10,11,11,13,14,14,14,14,15,15,18,18,19)
# 0.25-kvantiili eli alakvartiili
quantile(luvut,probs=0.25) #oletus type=7
```

```
##  25%
## 8.75
```

```r
quantile(luvut,probs=0.25,type=2)
```

```
## 25%
## 8.5
```

```r
quantile(luvut,probs=0.25,type=6) # SPSS
```

```
##  25%
## 8.25
```

```r
# useita kvantiileja samassa komennossa
quantile(luvut,probs=c(0,0.25,0.50,0.75,1))
```

```
##    0%   25%   50%   75%  100%
##  1.00  8.75 12.00 14.25 19.00
```

Viiden numeron yhteenveto

- Viiden numeron yhteenveto: minimi, alakvartiili, mediaani, yläkvartiili ja maksimi
  - Tarvitaan laatikkokuvion piirtämiseen (tähän palataan myöhemmin)
- Voidaan laskea nopeasti R-funktiolla `summary` (tulostaa myös keskiarvon) tai `fivenum`
  - Lasketaan muuttujalle pituus

```
summary(hlotsim_dat$pituus)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   142.0   166.5   172.0   174.3   185.2   203.0
```

```
fivenum(hlotsim_dat$pituus)
```

```
## [1] 142.0 166.0 172.0 185.5 203.0
```

Lineaarimuunnos

Lineaarimuunnos

Olkoon a ja b reaalilukuja ja X on muuttuja.

Lineaarimuunnos havainnoille $x_1, x_2, \ldots, x_n$ on

$$y_i = a + bx_i$$

jossa i käy läpi arvot $1, \ldots, n$.

- Esim. jos muuttuja X saa lukuarvot 160, 165, 170, 175, 180 ja on valittu $a = -70$ ja $b = 0.50$ niin Y-muuttujan arvot ovat: 10, 12.5, 15, 17.5, 20
- a-parametri siirtää havaintojen sijaintia: negatiivinen siirtää vasemmalle, positiivinen oikealle.
- b-parametri skaalaa havaintojen etäisyyttä origosta: $b < 1$ siirtää lähemmäs nollaa, $b > 1$ kauemmas

Lineaarimuunnoksen vaikutus keskiarvoon

Lineaarimuunnos vaikuttaa keskiarvoon samalla tapaa kuin yksittäisiin arvoihin:

$$\bar{y} = a + b\bar{x}$$

Ts. jos tiedämme että X-muuttujan keskiarvo on 170 niin voimme laskea että Y-muuttujan keskiarvo on $-70 + 0.50 * 170 = 15$

- Kokeillaan R:n avulla

```
x <- c(160, 165, 170, 175, 180)
mean(x)
```

```
## [1] 170
```

```
y <- -70+0.50*x
mean(y)
```

```
## [1] 15
```

Lineaarimuunnoksen vaikutus varianssiin

Olkoon s_x^2 X-muuttujan varianssi ja Y-muuttuja on saatu lineaarimuutoksella kuten edellisellä kalvolla. Nyt

$$s_y^2 = b^2 s_x^2$$

Ts. b-kerroin vaikuttaa varianssiin siten että alkuperäisen muuttujan varianssi kerrotaan b^2:lla, jolloin saadaan uuden muuttujan varianssi. Kokeillaan R:n avulla

```
x <- c(160, 165, 170, 175, 180)
var(x)
```

```
## [1] 62.5
```

```
y <- -70+0.50*x
var(y)
```

```
## [1] 15.625
```

```
0.50^2 * var(x)
```

```
## [1] 15.625
```

Lineaarimuunnoksen vaikutus keskihajontaan

Olkoon s_x X-muuttujan keskihajonta ja Y-muuttuja on saatu lineaarimuutoksella kuten edellisellä kalvolla. Nyt

$$s_y = |b| s_x$$

Ts. uusi keskihajonta saadaan kertomalla alkuperäisen muutujan keskihajontaa b:n itseisarvolla. Kokeillaan R:n avulla

```
x <- c(160, 165, 170, 175, 180)
sd(x)
```

```
## [1] 7.905694
```

```
y <- -70+0.50*x
sd(y)
```

```
## [1] 3.952847
```

```
0.50 * sd(x)
```

```
## [1] 3.952847
```

Riippuvuuden tarkastelu

Riippuvuuden tarkastelu

Kahden muuttujan välistä riippuvutta voidaan tarkastella riippuen muuttujien luonteesta:

- Määrälliset muuttujat
  - Hajontakuvion avulla
  - Pearsonin tai Spearmanin korrelaatiokertoimen avulla
- Järjestysasteikolliset muuttujat
  - Spearmanin korrelaatiokertoimen avulla
- Luokitteluasteikolliset muuttujat
  - Ristiintaulukoinnin avulla (testaaminen jää peruskurssille)

Pearsonin tulomomenttikorrelaatiokerroin

Pearsonin tulomomenttikorrelaatiokerroin

- Soveltuu jatkuville ja väh. välimatka-asteikollisille muuttujille
- Mittaa lineaarista riippuvuutta eli miten vahvasti hajontakuviossa pisteparvi on keskittynyt suoralle
- Arvot välillä $-1 \leq r_{xy} \leq 1$
- Määritellään tätä varten otoskovarianssi, joka on otosvarianssin läheinen sukulainen

Otoskovarianssi

Olkoon muuttujista X ja Y havaintoja $x_1, x_2, \ldots, x_n$ ja $y_1, y_2, \ldots, y_n$ s.e. (x_i, y_i) on mitattu samalta tilastoyksiköltä. Otoskovarianssi s_{xy} on

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

Jos lasketaan muuttujan X kovarianssi itsensä kanssa, niin saadaan X:n varianssi s_x^2.

Pearsonin korrelaatiokerroin r_{xy}

Pearsonin (tulomomentti)korrelaatiokerroin saadaan jakamalla otoskovarianssi X:n ja Y:n otoskeskihajonnoilla:

$$
\begin{aligned}
r_{xy} &= \frac{s_{xy}}{s_x \cdot s_y} && (1)\\
&= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x \cdot s_y} && (2)\\
&= \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i - \bar{x})}{s_x}\frac{(y_i - \bar{y})}{s_y} && (3)
\end{aligned}
$$

Tulkinta

- $r_{xy} = 0$: ei lainkaan lineaarista riippuvuutta
- $r_{xy} = 1$: täydellinen positiivinen lineaarinen riippuvuus (arvot nousevalla suoralla)
- $r_{xy} = -1$: täydellinen negatiivinen lineaarinen riippuvuus (arvot laskevalla suoralla)

Muuta huomioitavaa

- korrelaatiolla ei ole mittayksikköä (ei ole esim. kg)
- on herkkä oudokeille kuten keskiarvo ja keskihajonta

Pearsonin korrelaatio käsin laskien

- Lasketaan ensin keskiarvot ja keskihajonnat
  - $\bar{x} = 8.175$ ja $s_x = 1.5107945$,
  - $\bar{y} = 4.25$ ja $s_y = 4.5$

x	y	$\bar{x}$	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})/s_x$	$(y - \bar{y})/s_y$	$\frac{(x-\bar{x})}{s_x}\frac{(y-\bar{y})}{s_y}$
6.3	-2						
8.2	4						
8.2	8						
10.0	7						

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i - \bar{x})}{s_x}\frac{(y_i - \bar{y})}{s_y} =$$

Lineaarimuunnoksen vaikutus korrelaatioon

Olkoon, lineaarimuunnokset havainnoille $(x_1, y_1), \dots, (x_n, y_n)$

$$x_i' = a + bx_i$$

ja

$$y_i' = c + dy_i$$

jossa i käy läpi arvot $1, \dots, n$ ja a, b, c, d ovat reaalilukuja. Nyt

$$r_{x'y'} = r_{xy}$$

kun b ja d ovat etumerkeiltään samat, ja

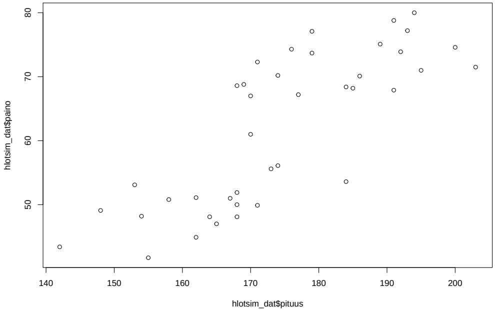
$$r_{x'y'} = -r_{xy}$$

kun b ja d ovat etumerkeiltään erit.

Pearsonin korrelaatio R-kielellä

- Pearsonin korrelaatio R:ssä lasketaan funktiolla `cor(x,y)` vektoreille `x` ja `y`

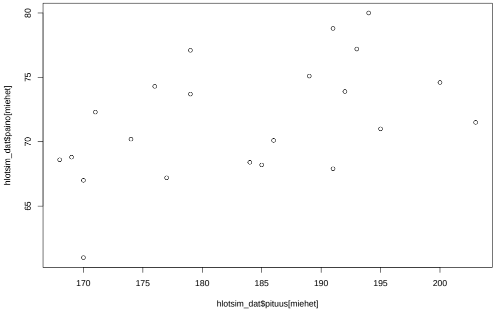
```
library(datas4uef)
data("hlotsim_dat")
plot(hlotsim_dat$pituus,hlotsim_dat$paino)
```




```
cor(hlotsim_dat$pituus,hlotsim_dat$paino)
```

```
## [1] 0.7930337
```

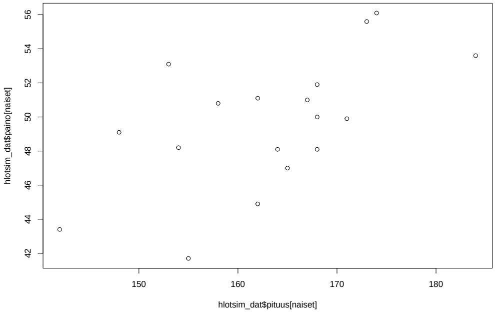
```r
# rajoitutaan miehiin
miehet <- hlotsim_dat$sukupuoli=="mies"
plot(hlotsim_dat$pituus[miehet],hlotsim_dat$paino[miehet])
```




```r
cor(hlotsim_dat$pituus[miehet],hlotsim_dat$paino[miehet])
```

```
## [1] 0.4974266
```

```r
# rajoitutaan naisiin
naiset <- hlotsim_dat$sukupuoli=="nainen"
plot(hlotsim_dat$pituus[naiset],hlotsim_dat$paino[naiset])
```




```r
cor(hlotsim_dat$pituus[naiset],hlotsim_dat$paino[naiset])
```

```
## [1] 0.5958413
```

Spearmanin järjestyskorrelaatiokerroin

Spearmanin järjestyskorrelaatiokerroin

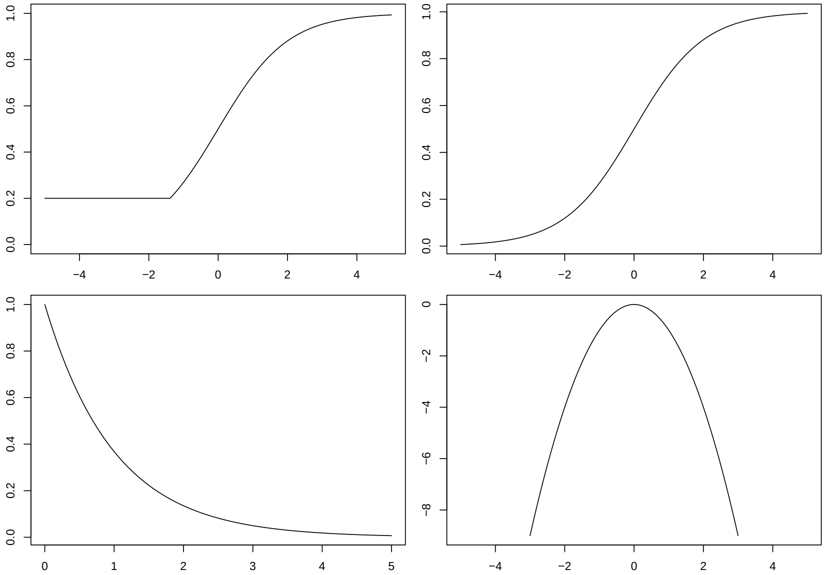
- Soveltuu jatkuville ja väh. järjestysasteikollisille muuttujille
- Mittaa monotonista riippuvuutta, eli miten yhteensopivia kahden muuttujan havaintoarvojen suuruusjärjestykset ovat.
  - Tyypillisin käyttötilanne on silti välimatka-asteikollisille muuttujille, kun lineaarinen yhteys ei ole uskottava, mutta monotoninen on.
- Arvot välillä $-1 \leq r_{xy} \leq 1$
- Määritellään tätä varten monotonisesti kasvava (ja vähenevä) funktio

Monotoninen funktio

Olkoon f funktio s.e. $y = f(x)$.

- Funktio f on monotonisesti kasvava, jos kaikille x-arvoille pätee:
  - kun x:n arvoa kasvatetaan, niin y:n arvo kasvaa tai vähintään pysyy samana
  - eli matemaattisemmin ilmaistuna jos $x_1 < x_2$ niin $y_1 = f(x_1) \leq f(x_2) = y_2$
- Vastaavasti funktio f on monotonisesti vähenevä, jos kaikille x-arvoille pätee:
  - kun x:n arvoa kasvatetaan, niin y:n arvo vähenee tai korkeintaan pysyy samana
  - eli matemaattisemmin ilmaistuna jos $x_1 < x_2$ niin $y_1 = f(x_1) \geq f(x_2) = y_2$

Mitkä seuraavista ovat monotonisia funktioita?



Spearmanin järjestyskorrelaatiokerroin r_s

- Oletetaan jälleen, että muuttujista X ja Y on havaintoja $x_1, x_2, \ldots, x_n$ ja $y_1, y_2, \ldots, y_n$ s.e. (x_i, y_i) on mitattu samalta tilastoyksiköltä, ja että X ja Y ovat vähintään järjestysasteikollisia.
- Muodostetaan X:n ja Y:n havaintoja vastaavat järjestysluvut kummallekin erikseen s.e. pienin arvo saa järjestysluvuksi 1, seuraavaksi pienin 2 jne. Mikäli jotain arvoa on enemmän kuin yksi, eli tulee "tasapeli", niin tällaisille annetaan kyseisten järjestyslukujen keskiarvo. Esim. jos havainnot ovat 6.3, 8.2, 8.2 ja 10.0, niin järjestysluvuiksi tulee 1, 2.5, 2.5, 4, missä $2.5 = \frac{2+3}{2}$.
  - R-kielessä `rank(x)` määrittää x-vektorin järjestysluvut.
- Jos muuttujat ovat valmiiksi järjestysasteikollisia, niin järjestysluvut ovat valmiina, eikä niitä tarvitse erikseen määrittää.

Spearmanin järjestyskorrelaatiokerroin r_s

Yleisesti pätee seuraava:

$$r_s = \frac{s_{R(X),R(Y)}}{s_{R(X)} \cdot s_{R(Y)}},$$

missä $R(X)$ ja $R(Y)$ ovat X- ja Y-havaintojen järjestysluvut.

Ts. kun X:n ja Y:n havainnot korvataan järjestysluvuillaan, niin niiden Spearmanin korrelaatio voidaan laskea järjestysluvuista samalla kaavalla kuin Pearsonin korrelaatio.

Erikoistapaus, kun ei tasapelejä

Mikäli järjestysluvut ovat kokonaislukuja (ei tasapelejä) niin voidaan käyttää seuraavaa kaavaa:

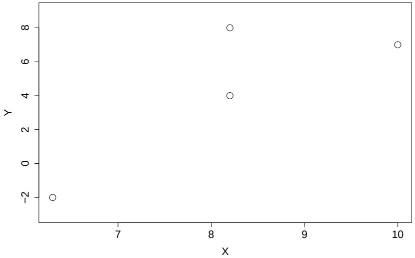
$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

missä

$d_i = R(x_i) - R(y_i)$ on järjestyslukujen erotus havaintoparille (x_i, y_i).
$n =$ havaintoparien (rivien) määrä.

Esimerkki

X	Y	R(X)	R(Y)
6.3	-2		
8.2	4		
8.2	8		
10.0	7		



R-kielellä:

```r
x <- c(6.3, 8.2, 8.2, 10.0)
y <- c(-2, 4, 8, 7)
cor(x,y,method="spearman")
```

```
## [1] 0.6324555
```

Ristiintaulukko eli kontingenssitaulu

- Ristiintaulukoinnin ja khi toiseen -riippumattomuustestin (Tilastotieteen peruskurssilla) avulla voidaan selvittää riippuvatko kaksi luokitteluasteikollista muuttujaa toisistaan tilastollisesti.
  - Perusidea on, että jos muuttujat eivät riipu toisistaan, niin toisen arvon tietäminen ei anna mitään tietoa (todennäköisyysmielessä) toisen arvosta.
  - Jos riippuvuutta on, niin ristiintaulukolla voidaan tarkastella millaista riippuvuus on: mitkä havaintoarvoparit ovat yleisempiä kuin, jos muuttujat olisivat riippumattomia.
- Oppikirjoissa näkee usein esimerkkejä, joissa muuttujat on johdettu välimatka-asteikollisista muuttujista luokittelemalla ne. Tämä ei ole välttämättä hyvä idea tutkimuskäyttöön, sillä ko. tilanteeseen on parempia menetelmiä.

Esimerkki

Kuopion sosiaali- ja terveystointen yhdistämisen vaikutuksista tehdyssä tutkimuksessa kyselylomakkeessa on kysytty henkilön sukupuoli ja takaavatko vastaanottotilat yksityisyyden.

Taulukko 1: Vastaajien kokemus vastaanottotilojen yksityisyyden suojasta Kuopiossa.

Mielipide	Mies	Nainen	Yht.
Huonosti	7	144	151
Hyvin	40	292	332
Yht.	47	436	483

Taulukko 2: Vastaajien kokemus vastaanottotilojen yksityisyyden suojasta Kuopiossa (ehdolliset prosenttijakaumat).

Mielipide	Mies	Nainen	Yht.
Huonosti	14.9 %	33.0 %	31.3 %
Hyvin	85.1 %	67.0 %	68.7 %
Yht.	100.0 %	100.0 %	100.0 %

Ristiintaulukon muodostaminen aineistosta

Sukupuoli	Tiedekunta
M	FILO
N	LUMET
N	LUMET
M	TT
N	TT

```
##
##      FILO TT LUMET
##    M    1  1     0
##    N    0  1     2
```

Taulukon indeksointi ja merkinnät

- Ristiintaulukko (ja kaikki muut taulukot matematiikassa) indeksoidaan s.e. indeksi $i = 1, 2, \ldots, r$ määrittää millä rivillä taulukossa ollaan. Indeksi $j = 1, 2, \ldots, c$ määrittää mikä sarake on kyseessä.
- Koska ristiintaulukko sisältää lukumääriä eli frekvenssejä, niin merkitään niitä f_{ij} rivillä i sarakkeessa j.

f_{11}	f_{12}	f_{13}
f_{21}	f_{22}	f_{23}

- Oletetaan jatkossa, että muuttuja X:n arvot ovat riveillä ja muuttujan Y arvot sarakkaissa.

Rivi- ja sarakesummat

- Usein on tarpeen laskea frekvenssien summia riveittäin tai sarakkeittain
  - Rivisumma riville i on

    $$f_{i\cdot} = \sum_{j=1}^{c} f_{ij},$$

    missä i on valitun rivin indeksi (luku).
  - Sarakesumma sarakkeelle j on

    $$f_{\cdot j} = \sum_{i=1}^{c} f_{ij},$$

    missä j on valitun sarakkeen indeksi (luku).

Marginaalijakauma eli reunajakauma

- X:n marginaalijakauma saadaan laskemalla kaikki rivisummat $f_{i\cdot}$ ja raportoimalla ne yhdessä niihin liittyvien X:n arvojen kanssa.
- Vastaavasti Y:n marginaalijakauma saadaan laskemalla kaikki sarakesummat $f_{\cdot j}$ ja raportoimalla ne yhdessä niihin liittyvien Y:n arvojen kanssa.

Ehdollinen (empiirinen) jakauma

- Kiinnittämällä muuttujan X arvo, ja tarkastelemalla muuttujan Y jakaumaa saadaan Y:n ehdollinen jakauma ehdolla X.
  - Vastaavasti vaihtamalla X:n ja Y:n roolit keskenään saadaan X:n ehdollinen jakauma Y:n suhteen (eli ehdolla Y).
- Taulukon rivit ovat sarakemuuttujan Y ehdolliset jakaumat. Samoin sarakkeet ovat rivimuuttujan X ehdolliset jakaumat.
- Ehdollisista (frekvenssi)jakaumista voidaan edelleen määrittää ehdolliset prosenttijakaumat.
  - Riippuvuuden tarkastelu empiirisesti tapahtuu ehdollisia prosenttijakaumia vertailemalla. Suuret erot ehdollisissa prosenttijakaumissa antavat viitteitä riippuvuudesta - varsinkin, jos aineisto ei ole kovin pieni!

R:n käyttäminen

- Ristiintaulukon muodostaminen
- Marginaalijakaumien laskeminen vaiheuttain + funktiolla add.margins
- Ehdollisten jakaumien laskeminen
- Ehdollisten prosenttijakaumien laskeminen

Ristiintaulukon muodostaminen

- Havaintomatriisista ristiintaulukko saadaan tehtyä table-funktiolla

```
data("hlotsim_dat")
rt <- table(hlotsim_dat$sukupuoli,hlotsim_dat$pääaine)
rt
```

```
##
##          TK1K TILTK TTRA2 YM3 TTBiomed2
##   nainen    2     7     4   4         1
##   mies      5     2     3   4         8
```

- Jos aineistoa ei ole, mutta lukumäärät ovat tiedossa, niin voidaan luoda taulukko matrix-komennolla (aineisto, joka oli aiemmin)

```
rt2 <- matrix(c(7, 144,
                40,292),nrow=2,ncol=2,byrow=T)
dimnames(rt2) <- list("Mielipide"=c("Huonosti","Hyvin"),
                      "Sukupuoli"=c("Mies","Nainen"))
rt2
```

```
##            Sukupuoli
## Mielipide  Mies Nainen
##   Huonosti    7    144
##   Hyvin      40    292
```

Marginaalijakaumien laskeminen vaiheittain

- Marginaalijakaumat eli reunajakaumat

```
rowSums(rt2)
```

```
## Huonosti    Hyvin
##      151      332
```

```
colSums(rt2)
```

```
##   Mies Nainen
##     47    436
```

```
sum(rt2)
```

```
## [1] 483
```

- marginaaliprosenttijakaumat

```
prop.table(rowSums(rt2))
```

```
##  Huonosti     Hyvin
## 0.3126294 0.6873706
```

```
prop.table(colSums(rt2))
```

Marginaalijakaumien laskeminen funktiolla addmargins

- Tai addmargins -funktiolla
  - **Huom!** addmargins-funktio antaa taulukon, josta ei tule enää laskea rivi-/sarakesummia eikä osuuksia R:ssä! Sum-sarake ja rivi eivät ole erityisasemassa.

```
addmargins(rt2)
```

```
##            Sukupuoli
## Mielipide  Mies Nainen Sum
##   Huonosti    7    144 151
##   Hyvin      40    292 332
##   Sum        47    436 483
```

Ehdollisten jakaumien laskeminen

```
prop.table(rt2,margin=1) #riveittäin, eli rivien osuuksien summa on 1
```

```
##            Sukupuoli
## Mielipide         Mies    Nainen
##   Huonosti 0.04635762 0.9536424
##   Hyvin    0.12048193 0.8795181
```

```
prop.table(rt2,margin=2) #sarakkeittain
```

```
##            Sukupuoli
## Mielipide        Mies    Nainen
##   Huonosti 0.1489362 0.3302752
##   Hyvin    0.8510638 0.6697248
```

Ehdollisten prosenttijakaumien laskeminen

```
100*prop.table(rt2,margin=1) #riveittäin
```

```
##            Sukupuoli
## Mielipide        Mies   Nainen
##   Huonosti   4.635762 95.36424
##   Hyvin     12.048193 87.95181
```

```
100*prop.table(rt2,margin=2) #sarakkeittain
```

```
##            Sukupuoli
## Mielipide       Mies   Nainen
##   Huonosti 14.89362 33.02752
##   Hyvin    85.10638 66.97248
```