Data-analyysi: johdanto

Lähteet

- Data-analyysin osuus kurssilla perustuu lähinnä teokseen
  - P. Cohen: Empirical Methods for Artificial Intelligence
    - Eksploratiivinen data-analyysi, luku 2
    - Hypoteesin testaus ja parametrien estimointi, luvut 4-5

Johdantoa ja kertausta

- tutkimustieto **empiiristä**

= **havainnosta** (*observation*) tai **kokeesta** (*experiment*) peräisin olevaa

- siis ennen data-analyysiä datan kerääminen
- usein datan keräämiseen liittyy mittauksia
- havaitsemalla/mittaamalla saatu tieto sisältää (melkein) aina epätarkkuutta

⇒ ilmiöiden ymmärtämisessä ja ennustamisessa tarvitaan satunnaisuuden, satunnaisvaihtelun hallitsemista

⇒ tilastolliset menetelmät

Mittausten virhetyyppejä

- systemaattiset virheet
  - esim. mittarin virheellinen kalibrointi
  - mittarilla saadut tulokset voivat silti olla vertailukelpoisia keskenään
- satunnaiset virheet (**kohina** (*noise*))
  - esim. pyöristysvirheet
  - virhe mittaria luettaessa

  ⇒ suuressa aineistossa keskimäärin yhtä paljon liian pieniä ja suuria arvoja

  ⇒ hallittavissa hyvin tilastollisilla menetelmillä

Poikkeavat arvot (*outliers*)

- **huomattavasti** muista poikkeavat arvot ovat ongelmallisia
- esim. seuraava otos:

  `1 3 3 2 4 2 345 2`
- suhtautuminen poikkeaviin arvoihin on yksi data-analyysin keskeisiä kysymyksiä
- voi olla vaikea erottaa, milloin kyse on virheellisestä havainnosta, milloin taas ei

Poikkeavat arvot (2)

- poikkeava arvo vaikuttaa helposti analyysiin tuloksiin paljon
  - jos arvo poistetaan, mutta se olikin oikea – :-(
  - jos arvoa ei poisteta, mutta se olikin virheellinen – :-(
- käsityksemme *outlier*-arvojen määrästä täytyy vaikuttaa käyttämiimme analyysimenetelmiin, sillä
  - jotkut tunnusluvut ja menetelmät vähemmän sensitiivisiä poikkeaville arvoille kuin toiset
  - toiset turmeltuvat käyttökelvottomiksi helposti (esim. keskiarvo)
- aiheeseen palataan tunnuslukujen käsittelyn yhteydessä

Lukujen esitystarkkuus

- laskennalliselta kannalta mahdollinen virhelähde on myös lukujen rajallinen esitystarkkuus tietokoneessa
- voi joissakin tilanteissa aiheuttaa pahojakin vääristymiä
- ei yleensä …
- lähinnä pyöristysvirheiden kumuloituminen iteroitaessa
- joudutaan tällä kurssilla ohittamaan tällä maininnalla

Empiirinen tutkimus

1. eksploratiivinen vaihe (*exploratory data-analysis*)
2. evaluoiva, tarkentava vaihe (*confirmatory data-analysis*)

Eksploratiivinen data-analyysi

- dataan tutustumista
- yleiskuva, relevanttien kysymysten löytäminen
- konkreettisesti:
  graafiset esitykset ja **tiivistäminen tunnuslukujen avulla**
  - on hyvä jos tiedossa on täsmällisiä tutkimusongelmia
  - silti eksploratiivinen analyysi hyvä tehdä
  - liian tiukka kysymystenasettelu voi estää näkemästä muita seikkoja
  - annetaan datasta esille nousevien piirteiden vaikuttaa kysymyksiin

Eksploratiivinen data-analyysi (2)

- tutkimuksessa aina subjektiivinen puoli
- miksi valittiin tämä aihepiiri ja tämä data?
- miksi käytettiin näitä menetelmiä eikä joitakin muita?
- miksi valittiin tämä nollahypoteesi?
- miksi valittiin tämä malli?
- miksi tuloksia tulkittiin näin?

⇒ subjektiivisuutta ei pidä kieltää, mutta se pitää tiedostaa ja sitä pitää kontrolloida

Eksploratiivinen data-analyysi (3)

- konkreettisia asioita jotka liittyvät eksploratiiviseen vaiheeseen
  - datan esikäsittely
    - arvoalueet, rajat, järkevyys
    - virheellisten arvojen poistaminen
    - puuttuvat arvot; onko niitä ja miten suhtaudutaan?
  - säännönmukaisuuksien, toistuvien ilmiöiden etsiminen
  - muuttujien välinen riippuvuus, esim. korrelaatio
  - hypoteesien etsiminen (?)

Evaluoiva, tarkentava vaihe

- tutkimusongelmien tarkempi määrittäminen
- hypoteesien testaaminen
- ilmiön yksityiskohtaisempi ja kompleksisempi **mallintaminen**

⇒ mallin parametrien **estimointi**

⇒ ilmiön **ennustaminen** ja/tai ymmärtäminen

Tulkinta

- data-analyysiin liittyy myös **tulosten tulkinta**
  - ovatko esim. löydetyt riippuvuudet tutkimusalueen kannalta oleellisia
  - onko niistä löydettävissä syy-seuraus -suhteita
  - auttavatko ne ennustamaan ilmiön käyttäytymistä
  - ⇒ olisiko toisenlainen malli ollut hyödyllisempi
  - ⇒ mahdolliset uudet kokeet ja uuden datan keruu

Työkaluja data-analyysiin, osa I (Awk)

Sorvin ääreen

- data-analyysi vaatii sopivat työkalut
  - raakadatan esikäsittely
  - tietokannat
  - tilastollinen analyysi
  - visualisointi
- jatkossa opitaan alkeita eräiden työkalujen käytössä

Sorvin ääressä

- usein tutkimusdata ei ole tietokannassa eikä sitä aina sellaiseen kannata viedäkään

⇒ tiedostoissa olevan datan “kääntely” ja “pyörittely”

- taulukkomuotoisessa datassa
  - sarakkeiden ja rivien poistaminen
  - annetun ehdon täyttävien rivien valitseminen
  - yksinkertaiset laskuoperaatiot

Awk

- **awk** on yksinkertainen ja vanha UNIX-työkalu
- ohjelmointikieli, jolla helppo käsitellä rivi+sarake-muodossa olevia tekstitiedostoja
- awk-ohjelma annetaan joko komentorivillä tai tiedostosta
- tulkittava kieli (ei kääntämistä)
- awk-kielen syntaksi muistuttaa monessa kohdin C-kieltä
- tässä yhteydessä vain hyvin lyhyt esittely
  - tavoite: oman datan käsittely siten, että jokainen osaa tehdä pikkuoperaatioita datalleen ilman laajempia ohjelmointikieliä

Awk (2)

- awk-ohjelma käy läpi kaikki syötetiedoston rivit ja suorittaa kullekin riville ohjelmassa annetut käskyt
- syötetiedostoja voi olla myös useampia
- awk-ohjelmassa ei tarvitse määritellä muuttujia
- tyyppimuunnokset ovat automaattisia
- ohjelmassa voi viitata syötetiedoston rivin i:nteen kenttään merkinnällä $i
- merkintää $0 käytetään viittaamaan koko riviin

Awk, yksinkertaisia esimerkkejä

- **awk '{print $1,$3}' input.txt** tulostaa tiedoston input.txt kunkin rivin ensimmäisen ja kolmannen sarakkeen arvon
- **awk '{print $0}' input.txt** tulostaa tiedoston input.txt jokaisen rivin sellaisenaan (eli koko tiedoston sellaisenaan)
- awk-ohjelman sisäinen muuttuja NF sisältää tiedon käsiteltävän rivin sarakkeiden lukumäärästä
- **awk '{print $NF,$(NF-1)}' input.txt** tulostaa viimeisen ja viimeistä edellisen kentän arvot

Awk, kontrollikäskyt

- awk-ohjelmissa ovat käytettävissä mm. seuraavat kontrollikäskyt:
  - **if** (*condition*) *statement* [**else** *statement*]
  - **while** (*condition*) *statement*
  - **do** *statement* **while** (*condition*)
  - **for** (*expr1; expr2; expr3*) *statement*

Awk, lisää esimerkkejä

- **awk '{if ($1==1) print $1,$3; else print $2,$4}' input.txt**
- jos ensimmäisen sarakkeen arvo 1, tulostetaan ensimmäinen ja kolmas, muuten toinen ja neljäs sarake
- awk-ohjelmalle voi antaa parametreja komentoriviltä optiolla -v
- **awk -v field=3 '{print $(field)}' input.txt**
- tulostaa annetun sarakkeen (parametri *field*) arvot
- jos parametreja useita, optio -v tarvitaan jokaisen parametrin edessä

Awk-ohjelman osat

- edellä awk-ohjelmissa on ollut vain yksi osa: varsinainen suoritusosa, joka on pakollinen osa
- yleisesti awk-ohjelma voi sisältää kolme osaa:
  1. alustusosa (alkaa sanalla BEGIN)
  2. varsinainen suoritusosa
  3. lopetusosa (alkaa sanalla END)

Awk-ohjelman osat

- alustusosan käskyt suoritetaan *ennen* syötetiedoston lukemista
  - esim. muuttujien alustaminen voidaan näin suorittaa
- varsinaisen suoritusosan käskyt suoritetaan siis kerran jokaiselle syötetiedoston riville
- lopetusosan käskyt suoritetaan syötetiedon lukemisen *jälkeen*
  - tyypillisesti tulostetaan lopuksi joidenkin muuttujien arvoja

Awk-ohjelma, esimerkki

- olkoon seuraava ohjelma tiedostossa esimerkki.awk

```
BEGIN{}
{
sum=sum+$1;
count++
}
END{print sum, sum/count}
```

- komento awk -f esimerkki.awk input.txt tulostaa tiedoston input.txt rivien ensimmäisten sarakkeiden arvojen summan ja keskiarvon

Awk-ohjelma, taulukot

- 1-ulotteisen taulukon (t) i:nteen alkioon viitataan t[i]
- 2-ulotteisen taulukon (t) alkioon (i,j) viitataan t[i,j]
- taulukon indeksin ei tarvitse olla kokonaisluku vaan se voi olla mikä tahansa merkkijono
- seuraava ohjelma laskee ensimmäisen sarakkeen arvojen jakauman ja tulostaa sen

```
{sum[$1]++}
END{for (i in sum) print i,sum[i]}'
```


Awk, taulukot

- lopetusosassa käytetään taulukoihin liittyvää kontrollikäskyä, joka käy läpi kaikki taulukon indeksit: **for** (*val* **in** *array*) *statement*
- hyödyllinen kun taulukon indeksit ovat merkkijonoja
- kun indeksit ovat lukuja, tulostus ei (yleensä) tapahdu numerojärjestyksessä
- enemmän awk:ista: Unix Manual Pages (KDE Help Center tai komentoriviltä man awk)

Eksploratiivinen data-analyysi

Visualisointi

- datajoukon eri piirteiden esittäminen graafisesti
- eksploratiivisen data-analyysin keskeinen osa
- tutkija myös esittää saamansa tulokset paljolti graafisin esityksin!
- laadukas visualisointi ei ole helppoa:
  erilaiset aineistot vaativat erilaisia graafisia esityksiä

Visualisointi, työkaluja

- tilasto-ohjelmistot (esim. SAS, Splot, SPSS, Survo) sisältävät monipuoliset mahdollisuudet graafisiin esityksiin
- tällä kurssilla opitaan alkeita visualisoinnista käyttäen Gnuplot-ohjelmistoa
- kurssin lopulla tieteellinen visualisointi (CSC:n tutkijan luento)

Aineiston tiivistäminen

- aineiston ominaisuuksien tiivistäminen

$\Rightarrow$ **tunnuslukujen** (*statistic*) laskeminen

- yleisesti tunnusluku on mikä tahansa aineiston funktio $t = f(y_1, \ldots, y_n)$, jonka määräämiseksi on tunnettava kaikki havainnot $y_i, 1 \leq i \leq n$
- esimerkiksi havaintojen minimi- ja maksimiarvo ovat tunnuslukuja
- tunnusluku voi olla myös vektori

Tunnusluvuista

- käytännössä tunnusluvuilla pyritään ilmaisemaan tiettyjä havaintojen ominaisuuksia, kuten
  - arvojen keskittymistä (moodi, keskiarvo, mediaani)
  - arvojen leveyttä, laajuutta, hajoamista (keskihajonta, varianssi, kvartiiliväli, minimi, maksimi)
  - arvojen jakauman tasaisuutta (vinous, huipukkuus)
  - useampien muuttujien riippuvuuksia (korrelaatio, kovarianssi, χ^2 (khiin neliö))

Yksittäisen muuttujan tarkastelu

- eksploratiivisen data-analyysin käsitteleminen aloitetaan seuraavassa **yksittäisten muuttujien** tarkastelemisesta
  - muuttujien (attribuuttien) eri tyypit
  - yksittäisen muuttujan arvojen jakaumien tunnusluvut
  - arvojen jakaumien esittäminen graafisesti

Datan arvotyypit ja asteikot

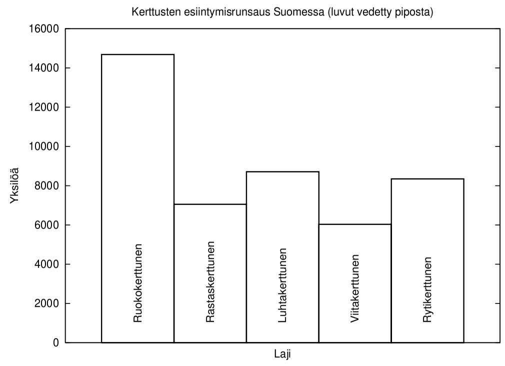
- aineiston muuttujilla (attribuuteilla) voi olla erityyppisiä arvoja
- muuttujan tyyppi vaikuttaa siihen
  - mitä tunnuslukuja voidaan laskea
  - mitä visualisointimenetelmiä kannattaa käyttää
  - mitä analyysimenetelmiä voidaan soveltaa

Kategoria-asteikko

- joukko erillisiä (=diskreettejä) arvoja, joilla on vain nimet, mutta ei järjestystä (esim. sukupuoli, silmien väri, lintulaji)
- tunnuslukuja:
  moodi (l. tyyppiarvo) = aineiston yleisin arvo
- visualisointi: histogrammi

Kategoria-asteikko, esimerkki

- attribuutin 'laji' moodi on 'ruokokerttunen'



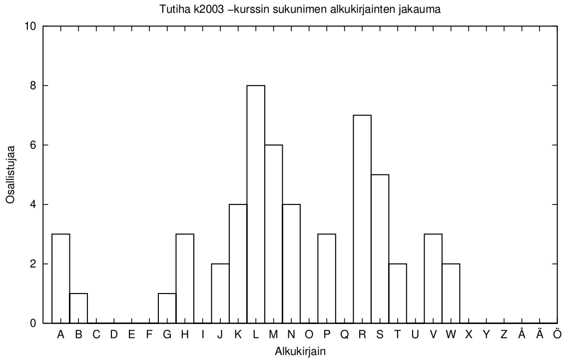
Ordinaaliasteikko

- arvot diskreettejä ja niille on määritelty järjestys
- tunnuslukuja (moodin lisäksi):
  - minimi - ja maksimiarvot
  - *mediaani* = aineiston keskimmäinen arvo

$$m(y_1, \ldots, y_n) = \begin{cases} y_{(\frac{n}{2}+\frac{1}{2})} & \text{kun } n \text{ pariton,} \\ y_{(\frac{n}{2})}, \text{tai } y_{(\frac{n}{2}+1)} & \text{kun } n \text{ parillinen} \end{cases}$$

Ordinaaliasteikko, esimerkki

- muuttuja: tälle kurssille ilmoittautuneiden sukunimien ensimmäinen kirjain
- aineistossa moodi on ’L’ ja mediaani ’M’



Numeeriset asteikot

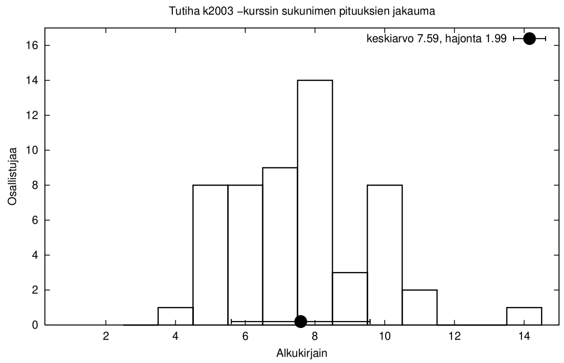
- seuraavilla kolmella asteikolla arvot numeerisia
- arvoilla voi mielekkäästi laskea
- arvot voivat olla diskreettejä tai jatkuvia

Numeeriset asteikot (2)

- **intervalliasteikko**
  - arvojen *erotuksilla* mielekäs tulkinta
  - esim. lämpötilan mittaukset Celsius-asteina
- **suhdeasteikko**
  - arvojen *suhteilla* mielekäs tulkinta
  - asteikko voidaan vapaasti valita
  - esim. lämpotilat Kelvin-asteina
- **absoluuttinen asteikko**
  - kuten edellinen, mutta myös mitta-asteikko on absoluuttinen
  - esim. osuudet maapallon väestöstä

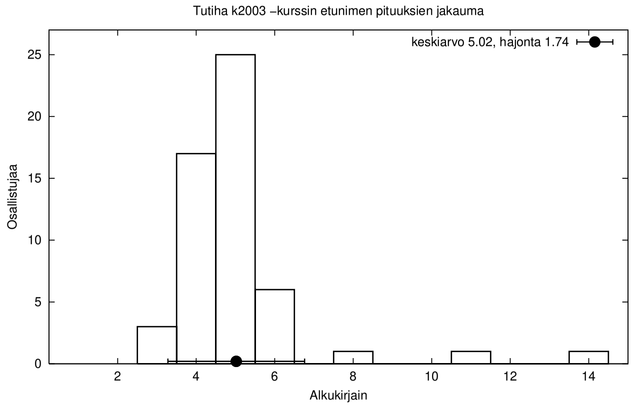
Esimerkki, absoluuttinen asteikko

- jatkoa esimerkille nimien alkukirjaimesta
- nyt muuttujana sukunimen pituus



Esimerkki, absoluuttinen asteikko

- etunimien pituuksien jakauma



Arvojen keskittyminen

- seuraavat tunnusluvut ovat tavallisimpia otoksen yksittäisen muuttujan keskittymisen kuvaajia
- (aritmeettinen) keskiarvo

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

  - heikkous: yksikin huomattavasti poikkeava arvo (outlier) vaikuttaa paljon

  $\Rightarrow$ ei sovi käytettäväksi, jos paljon kohinaa
  - myös hyvin vinot jakaumat ongelmallisia

  $\Rightarrow$ keskiarvo voi siirtyä kauaksi sieltä missä suurin osa havainnoista on

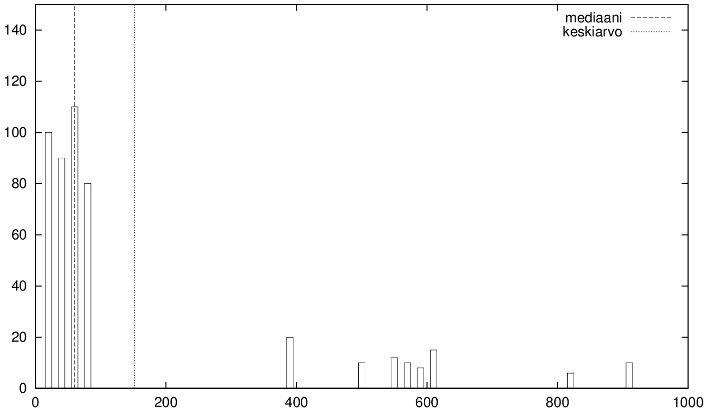
Arvojen keskittyminen (2)

- mediaani
  - ei yhtä herkkä yksittäisille virhearvoille kuin keskiarvo
  - intervalli-, suhde- ja absoluuttinen asteikko: mediaani voidaan määrittää yksikäsitteisesti myös silloin kun n on parillinen
  - n parillinen $\Rightarrow$ keskimmäisten arvojen keskiarvo

$$m(y_1, \ldots, y_n) = (y_{(\frac{n}{2})} + y_{(\frac{n}{2}+1)})/2, \text{ kun } n \text{ parillinen}$$

Esimerkki: keskiarvo ja mediaani

- hyvin vino jakauma $\Rightarrow$ mediaani lähempänä havaintojen valtaosaa



Arvojen keskittyminen (3)

- kohinaisille aineistoille aritmeettista keskiarvoa sopivampi on usein
  tasoitettu keskiarvo (*trimmed mean*)
  - arvojen ääripäistä jätetään osa huomioimatta keskiarvoa laskettaessa

Arvojen vaihtelevuus

- tavallisimpia arvojen vaihtelevuutta kuvaavia tunnuslukuja
  - (otos)keskihajonta (*standard deviation*, *SD*)

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

  - otosvarianssi on keskihajonnan neliö
  - keskihajonnan intuitiivinen merkitys: havainnon keskimääräinen etäisyys keskiarvosta

Keskihajonta ja otoskeskihajonta

- jos keskihajonta on keskiarvo , miksi nimittäjässä ei esiinny havaintojen lukumäärä n vaan $n-1$?
  - oletus: havainnot otos jostakin (tuntemattomasta) jakaumasta
  - otoskeskihajonnan odotusarvo on tuon jakauman keskihajonta
  - kun (otoksen perusteella) arvioidaan taustalla olevaa jakaumaa $n-1$:llä jakaminen toimii tässä mielessä paremmin kuin n:llä
  - suurten otosten kyseessä ollessa tällä erolla ei ole merkitystä

Fraktiilit ja fraktiilivälit

- otoksen p-fraktiili on se otoksen arvo, jota
  - pienempiä tai yhtä suuria on 100 x p % otoksen arvoista
  - yhtä suuria tai suurempia on 100 x (1-p) % otoksen arvoista
- mediaani on p-fraktiili, jossa $p = 0.5$
- kvartiili on p-fraktiili, jossa $p = 0.25$
- kvartiiliväli on (2-ulotteinen) tunnusluku (0.25-fraktiili, 0.75-fraktiili), vastaavasti muut fraktiilivälit
- hypoteesin testauksen yhteydessä tarkastellaan usein 0.95-, 0.99- ja 0.999-fraktiilivälejä

Työkaluja data-analyysiin, osa II (Gnuplot)

Visualisointi (yksi muuttuja)

- Gnuplot-ohjelmisto (alkujaan UNIX, Linux)
- käynnistyy komennolla gnuplot
- ohjelmasta poistuminen

```
gnuplot>quit
```

- apua saa käskyllä help tai lisäksi tarkentamalla käskyn tai osan siitä

```
gnuplot>help
gnuplot>help set
gnuplot>help set xrange
```

- ks. myös pikaohjeet verkossa ja Gnuplot Central

Gnuplot (2)

- ohjelman asetusten tallentaminen seuraavia työskentelykertoja varten

```
gnuplot>save set 'tiedosto.set'
```

- komentojen antaminen tiedoston kautta

```
gnuplot>load 'tiedosto.set'
```

- kuvien tulostaminen tiedostoon (PostScript)

```
gnuplot>set term postscript
gnuplot>set output ``kuvatiedosto.ps''
gnuplot>replot
```


Gnuplot (3)

- kukin syötetiedoston rivi vastaa
  - yhtä pistettä (esitystyypit: pisteet, viivat)
  - yhtä pylvästä (histogrammit)
- käyttäjä määrittää mitä syötteen sarakkeita käytetään graafisessa esityksessä
- erilaiset esitystyypit vaativat eri määrän sarakkeita
- seuraavassa esimerkissä etu- ja sukunimien lukumäärät sisältävä datatiedosto histogrammin piirtämistä varten (1. sarake: nimen pituus, 2. sarake: etunimien lukumäärä, 3. sarake: sukunimien lukumäärä)

Datatiedosto, esimerkki

```
3    3    0
4    17   1
5    25   8
6    6    8
7    0    9
8    1    14
9    0    3
10   0    8
11   1    2
12   0    0
13   0    0
14   1    1
```


Histogrammit ja Gnuplot

- gnuplot asettaa oletusarvoisesti x-akselin ja y-akselin skaalan automaattisesti (autoscale)
- toimii huonosti histogrammien tapauksessa

⇒ aseta manuaalisesti: set xrange, set yrange, set boxwidth

- y-akselilla korkein arvo ei saa saavuttaa asteikon ylälaitaa (matalin arvo nolla!)

```
gnuplot> set data style boxes
gnuplot> set xrange[0:15]
gnuplot> set yrange[0:30]
gnuplot> set boxwidth 1
gnuplot> plot 'nimi.data' using 1:2
```


Plot-komento

- plot-komennon yhteydessä voi lisämääreillä muuttaa viivojen ja pisteiden tyylejä ja leveyksiä/kokoa
- title-määreellä kerrotaan mitä kuvassa halutaan lukevan (ei koko kuvan otsikko, se asetetaan komennolla *set title*)
- oletusarvoinen viivan leveys on yleensä liian pieni

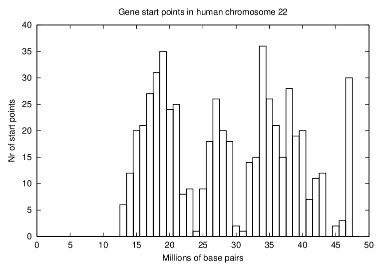
```
gnuplot> plot 'nimi.data' using 1:2
title 'etunimet' linewidth 3
```

- komentoja voi lyhentää

```
gnuplot> plot 'nimi.data' u 1:2 t 'etunimet' lw 3
```


Histogrammit

- kuvaan, jonka histogrammi antaa aineistosta vaikuttavat
  - alaraja
  - luokkavälin pituus (=pylvään paksuus)



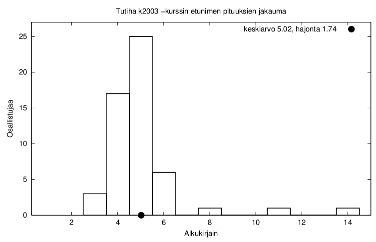
- pylvään leveyden valinta voi vaikuttaa oleellisesti

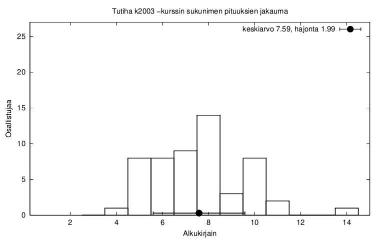
Histogrammit

- valintaan vaikuttavia asioita:
  - havaintojen lukumäärä
  - muuttujan arvojen vaihteluväli
  - mahdolliset luonnolliset luokkarajat
  - vierekkäisten luokkien frekvenssien vaihtelun määrä
    - perussääntö: kaikkea aineistossa esiintyvää vaihtelua ei saisi luokkajaossa kadottaa
  - käytännössä: kannattaa kokeilla erilaisia arvoja
  - aloita mieluummin tiheämmästä ja harvenna siitä!

Histogrammien vertailu

- jos haluaa verrata useita histogrammeja, on syytä käyttää samaa asteikkoa





Replot-komento

- pelkkä replot-komento toistaa edellisen plot-käskyn

```
gnuplot> replot
```

- hyödyllistä kun muuttaa jotakin asetusta ja haluaa sitten plotata samat arvot uudelleen
- replot-komentoa voi myös käyttää, jos haluaa useiden muuttujien arvoja samaan kuvaan
- lisätään seuraavaksi sukunimien pituudet kuvaavaan histogrammiin keskiarvo pisteenä x-akselilla

Pisteiden piirtäminen

- olkoon tiedostossa avg.data yksi rivi

```
5.02 1.74 0
```

- 1. sarake keskiarvo, 2. sarake otoskeskihajonta
- käytetään 1. saraketta (ja kolmatta)

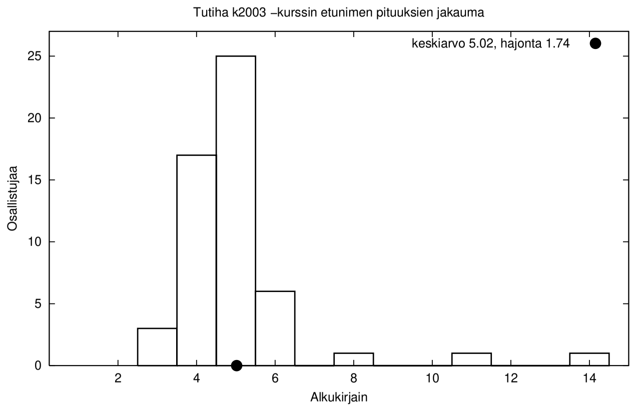
```
gnuplot> set data style points
gnuplot> replot 'avg.data' using 1:3 title 'keskiarvo'
```

Pisteiden piirtäminen (2)

- parhaat pistetyypit (henk.koht. mielipide)
  - pistetyypit 6 ja 7 (avoin ja väritetty pyöreä piste)
  - pistetyypit 4 ja 5 (avoin ja väritetty neliö)
  - pistetyyppi 8 (kolmio)

```
gnuplot> replot 'avg.data' using 1:2 title 'keskiarvo'
pointtype 7 pointsize 2
```


Etunimet



Hajontapylväät (*errorbars*)

- hajonnan ja luottamusvälien esittäminen pisteen ympärillä
- jos halutaan kuvata y-akselin suuntaista variaatiota valitaan datatyyli errorbars (= yerrorbars)
- jos halutaan kuvata x-akselin suuntaista variaatiota valitaan datatyyli xerrorbars
- palataan etunimien pituuksia kuvaavaan histogrammiin
- tilanne ennen keskiarvoa kuvaavan pisteen piirtämistä
- kuvataan nyt pelkän keskiarvopisteen lisäksi myös keskihajonnan suuruinen hajontapylväs

Hajontapylväät (2)

- datarivillä tarvitaan nyt kolmea saraketta:
  1. x-koordinaatti
  2. y-koordinaatti
  3. hajonnan leveys x-akselin (tai y-akselin) suunnassa pisteen (x,y) ympärillä

```
gnuplot> set data style xerrorbars
gnuplot> replot 'avg.data' using 1:3:2
title 'keskihajonta'
```

Hajontapylväät (3)

- oletusarvojen käytön sijasta myös hajontapylvään viiva- ja pistetyypit ja -koot voi antaa eksplisiittisesti

```
gnuplot> plot 'avg.data' using 1:3:2 title 'keskiarvo j
linetype 2 linewidth 2 pointtype 6 pointsize 1.5
```

- tai lyhentäen

```
gnuplot> plot 'avg.data' u 1:3:2
t 'keskiarvo ja -hajonta' lt 2 lw 4 pt 7 ps 3
```


Eksploratiivinen data-analyysi, kahden muuttujan tarkastelu

Kahden muuttujan riippuvuus

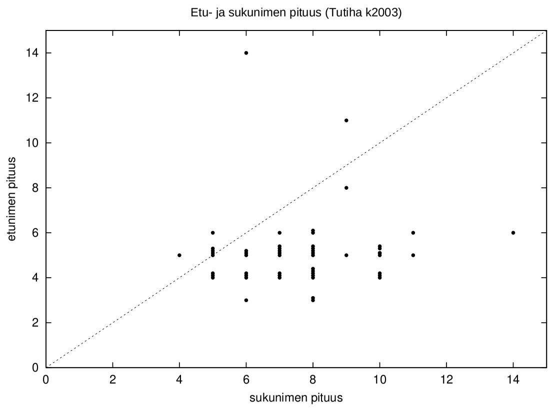
- kahden (satunnais)muuttujan X ja Y riippuvuus toisistaan
  - funktionaalinen riippuvuus
    = X:n arvo määrää yksikäsitteisesti Y:n arvon, tai päinvastoin
  - tilastollinen (stokastinen) riippuvuus
    = X:n arvo vaikuttaa Y:n arvon jakaumaan, tai päinvastoin
- funktionaalinen riippuvuus on luonnollisesti tilastollisen riippuvuuden erikoistapaus

Kahden muuttujan riippuvuus (2)

- tilastollista riippuvuutta arvioitaessa on tarkasteltava muuttujien X ja Y yhteisjakaumaa
- visualisointi: vähintään intervalliasteikon muuttujien kyseessä ollessa voi piirtää havaintojen sirontakuvion (*scatter plot*)
- jokaista havaintoa kohti yksi piste (x_i, y_i)

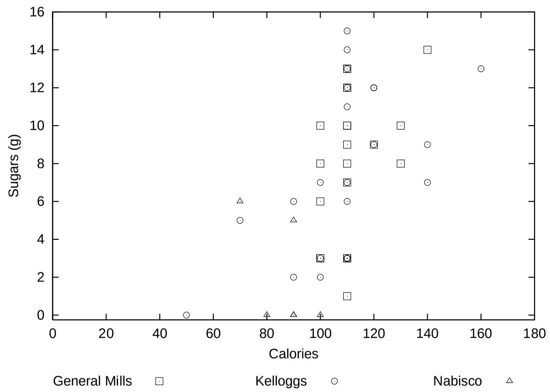
Sirontakuvio (*scatter plot*)

- jokaista kurssin osallistujaa kohti yksi piste
- x-akseli: sukunimen pituus, y-akseli: etunimen pituus



Kahden muuttujan riippuvuus (3)

- kolmas muuttuja (kategoria-asteikko) voidaan erotella erilaisin kuvioin/värein)
- esimerkki, eri yritysten murot:



Kovarianssi

- otoskovarianssi on tunnusluku, joka kuvaa vähintään intervalliasteikon muuttujien välisen assosiaation suuruutta

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- keskistys
- mittaa *lineaarista riippuvuutta*
- huono puoli: otoskovarianssin arvot voivat olla mielivaltaisen suuria tai pieniä, riippuen muuttujien hajonnasta

Korrelaatiokerroin

- *(Pearsonin) korrelaatiokerroin* ρ saadaan otoskovarianssista **standardoimalla** se muuttujien keskihajontojen tulolla

$$\rho(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

$\Rightarrow -1 \leq \rho(x,y) \leq 1$

- $\rho = 1$, kun pisteet (x_i, y_i) ovat samalla nousevalla suoralla
- $\rho = -1$, kun pisteet (x_i, y_i) ovat samalla laskevalla suoralla

Keskistys ja standardointi

- keskistys = siirretään mitta-asteikon nollakohta havaintojen keskiarvojen kohdalle
  - keskistys tehdään vähentämällä jokaisesta havainnosta havaintojen keskiarvo
- standardointi = muunnetaan mitta-asteikkoa siten, että keskihajonnan suuruisesta poikkeamasta tulee suuruudeltaan 1
  - standardointi tehdään jakamalla jokainen keskistetty havainto havaintojen keskihajonnalla

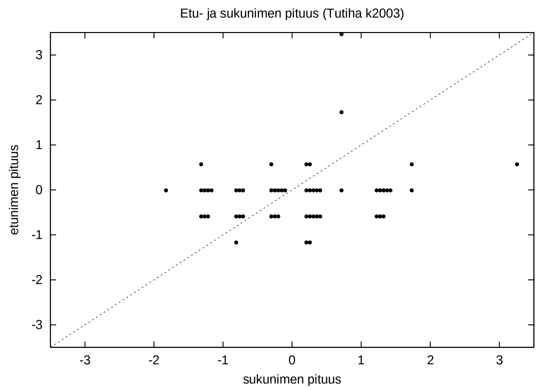
$$x_i' = \frac{(x_i - \bar{x})}{\sigma_x}, y_i' = \frac{(y_i - \bar{y})}{\sigma_y}$$

Keskistys ja standardointi (2)

- kurssilaisten etunimille $\bar{x} = 5.02$, $\sigma_x = 1.74$
- sukunimille $\bar{y} = 7.59$, $\sigma_y = 1.99$
- $x_1 = 6, y_1 = 5$

$$\Rightarrow x_1' = \frac{6 - 5.02}{1.74} \approx 0.56, y_1' = \frac{5 - 7.59}{1.99} \approx -1.30$$

Keskistys ja standardointi (3)

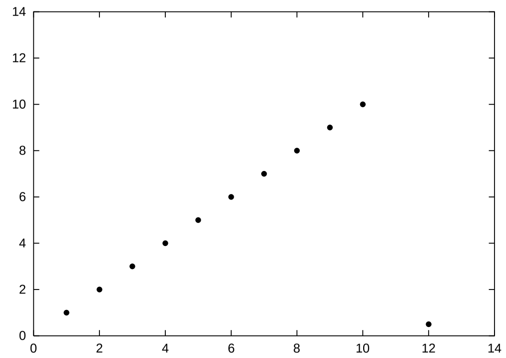


- $\rho(sukunimi, etunimi) \approx 0.07$

$\Rightarrow$ ei (lineaarista) riippuvuutta

Korrelaatiokerroin

- poikkeavan havainnon vaikutus



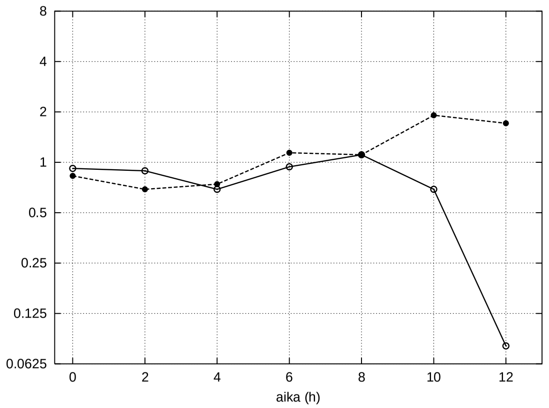
- yksi poikkeava havainto riittää pudottamaan korrelaation arvoon 0.47

Mitä korrelaatio ei kerro?

- muuttujat voivat olla epälineaarisesti riippuvia, vaikka niiden välinen korrelaatio on on pieni (siis lähellä nollaa)
- korrelaatio ja kovarianssi eivät myöskään kerro mitään siitä, onko muuttujien välillä syy-seuraus -suhdetta (eli *kausaalista* riippuvuutta)

Aikasarja

- mittausarvot ajan funktiona



Aikasarja (2)

- esimerkin arvot geenien aktiivisuutta ja passiivisuutta indikoivien aineiden pitoisuuksien suhteita
  - arvo 0.5 kertoo, että ainetta A on kaksi kertaa enemmän kuin ainetta B
  - arvo 2 kertoo, että ainetta B on kaksi kertaa enemmän kuin ainetta A
  - mm. tällaisissa tilanteissa on luontevaa esittää arvot käyttäen logaritmistä asteikkoa
  - gnuplotissa asetetaan log-asteikko/palautetaan tasavälinen asteikko komennoilla

```
gnuplot>set logscale y 2
gnuplot> set nologscale y
```


Gnuplot, viivadiagrammit

- pisteet, jotka yhdistetty viivalla (pisteet ja viivat niiden välille piirretään)

```
gnuplot> set data style linespoints
gnuplot> plot 'aikasarja' u 1:2
gnuplot> plot 'aikasarja' u 1:2 notitle lt 3 lw 2
pt 7 ps 0.8
```

- pisteet, jotka yhdistetty viivalla (vain viivat piirretään)

```
gnuplot> set data style lines
gnuplot> plot 'aikasarja' u 1:2
gnuplot> plot 'aikasarja' u 1:2 notitle lt 3 lw 2
```

Yhteenveto

- eksploratiivinen data-analyysi
  - dataan tutustumista
  - visualisointi, tiivistäminen tunnusluvuin
  - yhden muuttujan tarkastelu
    - histogrammit
    - keskiarvo, mediaani, otoskeskihajonta, kvartiiliväli
  - kahden muuttujan tarkastelu
    - sirontakuviot
    - aikasarjat
    - korrelaatiokerroin