Tutkimustiedonhallinnan peruskurssi

Hannu Toivonen, Marko Salmenkivi, Inkeri Verkamo
hannu.toivonen, marko.salmenkivi, inkeri.verkamo@cs.helsinki.fi

Helsingin yliopisto – Tietojenkäsittelytieteen laitos

Kurssin esittely

Perustiedot

- 581264-6 Tutkimustiedonhallinnan peruskurssi (3 ov)
- tietojenkäsittelytieteen cum lauden ja laudaturin valinnainen kurssi
- soveltuu myös muiden aineiden pääaineopiskelijoille, jotka joutuvat tekemisiin tutkimustietoaineistojen kanssa
- edellyttää tietojenkäsittelytieteen approbaturin tasoisia perustietoja ja -taitoja ohjelmoinnissa sekä tiedonhallinnassa
- tilastotieteen ja todennäköisyyslaskennan perusteita kerrataan nopeasti, pohjatiedoista on hyötyä

Mitä kurssilla käsitellään

- **tutkimusaineistojen käsittelyä ja hallintaa**
- datan visualisointi
- laskennallisia data-analyysimenetelmiä
- tutkimustiedonhallinnan erityistarpeet
- satunnaislukujen ja otosten generointi

Suhde muuhun opetukseen

- kurssin taustaa: johdatus sovellussuunnitteluun, tietokantojen perusteet, ohjelmoinnin perusteet
- kurssi sivuaa: tilastotiedettä, todennäköisyyslaskentaa
- kurssille jatkoa:
  - tiedonhallinta, tietokannan mallinnus, tietokantarakenteet ja algoritmit, tietovarastot, paikkatieto
  - tiedon louhinnan menetelmät, tiedon louhinnan erikoiskurssi

Kurssin rakenne ja aikataulu

- Ks. kurssin verkkosivut,
  http://www.cs.helsinki.fi/hannu.toivonen/teaching/tutihaK04/
- pakollinen harjoitustyö: ks. ohjeet verkossa,
  http://www.cs.helsinki.fi/u/laczak/tutiha/yleista.html

Kurssin suoritus

- luennot (ainoa vapaaehtoinen osa suoritusta)
- laskuharjoitukset (väh. 4 kertaa väh. puolet tehtävistä)
- harjoitustyö (6-20 p)
- tentti (13-40 p)
- yhteensä vähintään 30/60 p
- Huom:
  - harjoitustyön tekeminen on mahdollista vain kurssin yhteydessä
  - suoritus pelkällä loppukokeella ei ole mahdollista
  - tentin voi tarvittaessa uusia, jos harjoitustyö ja laskarit on ok

Kurssimateriaali

- **P.R. Cohen: Empirical Methods for Artificial Intelligence, luvut 2-5**
- **Z. Michalewicz (toim.): Statistical and Scientific Databases; Ellis Horwood Ltd, 1991, luvut 2, 4**
- **lisäksi erillisiä artikkeleita (ilm. myöhemmin)**
- **luentokalvot** (eivät yksin kata kurssin sisältöä)
- materiaalin saatavuus
  - kirjat kirjastossa, kurssilla käsiteltävät luvut kurssikansiossa
  - muu materiaali verkossa ja kansiossa
  - http://www.cs.helsinki.fi/hannu.toivonen/teaching/tutihaK04/

Tutkimustiedon(hallinnan) erityispiirteet

Erityispiirteet: osan sisältö

- Mistä tällä kurssilla on kysymys?
- Missä suhteissa tutkimustieto on erilaista kuin muu tieto?
- Millaisia erityisvaatimuksia tutkimustiedolla voi olla tiedon hallinnalle?
- Lähde: Z. Michalewicz (toim.): Statistical and Scientific Databases; luku 2

Mistä tutkimustieto on peräisin?

- tieteellinen tutkimus
  - havaintoja jostakin ilmiöstä (esim. säätila)
  - kokeellinen tutkimus (esim. hiukkaskiihdytys)
  - simulointikokeet (esim. ekologiset mallit)
- toiminnan sivutuotteena syntyneet tietokokoelmat
  - tieto kerätty alun perin muuhun tarkoitukseen
  - kerättävän tiedon rakenne riippuu yleensä vahvasti alkuperäisestä käyttötarkoituksesta
- tiedon tulva: analysointi on pullonkaula
- englanniksi "scientific and statistical databases"

Esimerkkejä sovellusalueista

- ympäristötutkimus
- meteorologia, seismologia
- avaruustutkimus
- arkeologia
- geenitutkimus, lääketiede
- kemia, fysiikka
- sosiologia, kielitiede
- ...

Vrt. perinteinen tiedonhallinta

- tutkimustiedot ja niiden tarpeet vaihtelevat paljon

⇒ vaihtelevia ja monipuolisia tarpeita

- 1. tietomalli
- 2. operaatiovalikoima
- 3. rinnakkaisuus
- 4. tiedon eheys
- 5. toipuminen
- 6. fyysinen rakenne

1. Tietomalli

- yleisimmin käytetty tietomalli (ER):
  - alkiot esitetään tauluissa
  - taulujen alkioita liitetään toisiinsa vastinsarakkeiden perusteella
  - alkiot ovat joukkoja: ei järjestystä

- tutkimustiedon edellyttämä tietomalli:
  - merkittävä osa tutkimustiedosta on järjestettyä dataa:
    - aikasarjat
    - koesarjat
  - yksittäisten alkioiden rakenne voi olla hyvin monimutkainen:
    - koetuloksiin voi liittyä tutkittavan osan rakennekaavio
    - alkioiden väliset yhteydet

2. Operaatiovalikoima

- perinteisen tietokannan operaatiot ovat SQL-kyselyitä:
  - hae kaikki alkiot, jotka täyttävät annetun ehdon, ja tee niille haluttu operaatio
- tutkimustiedon vaatimat operaatiot:
  - sovellusaluekohtaisia operaatioita
  - operaatiot voivat sisältää monimutkaista laskentaa
  - useat operaatiot käsittelevät kaikkia tietueita
  - käsiteltävän datan valinta myös esim. visualisoinnista

- tällä kurssilla tarkasteltavia operaatioita
  1. aineistojen analysointi
     - eksploratiivinen analyysi
     - visualisointi
     - (ei-parametrinen) tilastollinen analyysi (Monte Carlo -menetelmät hypoteesien testaamiseen ja parametrien estimointiin)
  2. otokset ja niihin kohdistuvat operaatiot
     - tarvitaan suurten aineistojen analysointiin
     - miten otos otetaan tehokkaasti?
     - miten satunnaisotoksen käyttö vaikuttaa analyysien tuloksiin?

3. Rinnakkaisuus

- perinteisissä sovelluksissa: rinnakkaisuus = monta samanaikaista käyttäjää jotka voivat häiritä toisiaan erityisesti tietoa päivittäessään
- samanaikaisuus tutkimustiedon yhteydessä:
  - ei yleensä paljon samanaikaisia käyttäjiä (ja he käyttävät useimmiten eri osia tietokokoelmasta)
  - tutkimustietoa ei usein päivitetä lainkaan (tai vain poikkeustilanteissa)

  ⇒ tarvitaanko rinnakkaisuutta?
  - tietokokoelma voi olla hyvin laaja ja/tai hajautettu:

  ⇒ edellyttääkö tehokas käyttö rinnakkaisuutta?

4. Tiedon eheys

- tietokannan sisältämän tiedon eheyttä voidaan tarkistaa esim. tietotyypin perusteella:
  - arvot hyväksytyllä välillä
  - esitysmuoto

- tutkimustiedon eheys:
  - tarvitaan usein monimutkaisiakin eheystarkistuksia
  - tiedon puhdistamisen yhteydessä poimitaan eroon epätavalliset arvot (outliers):
    - mahdollisesti virheelliset tiedot
    - puuttuvat tiedot
    - mielenkiintoiset havainnot?
  - täydentäminen ja korjaaminen voi olla hankalaa tai mahdotonta (historia?)
  - poikkeavien tietojen käsittelytapa on päätettävä tapauskohtaisesti

5. Toipuminen

- perinteisiin tietokantoihin liittyvä varmistus- ja toipumismenettely:
  - vikatilanteessa palautetaan tietokanta viimeisen onnistuneen tapahtuman jälkeiseen tilanteeseen
- tutkimustiedon edellyttämä toipuminen:
  - tyypillisesti pitkiä tapahtumia: toipuminen ei saa merkitä sitä, että paljon tehtyä työtä hylätään

6. Fyysinen rakenne

- tietokannan tyypillinen fyysinen talletus:
  - tavoitteena tehokkaat kyselyt
  - taulut talletetaan riveittäin tiedostoiksi
  - lisäksi muodostetaan hakemistoja, joiden avulla halutut kyselyt voidaan toteuttaa tehokkaammin
- tutkimustiedolle sopiva fyysinen rakenne:
  - usein alun perin sellaisenaan tiedostoissa
  - onko paras talletusrakenne:
    - tiedosto?
    - puu tai jokin muu tietorakenne?
    - tietokanta?

- (tutkimustiedolle sopiva fyysinen rakenne:)
  - tiedon rakenne:
    - mikä on sopivin järjestys?
    - tietojen ryhmittely = paikallisuus: mitä alkioita (tai alkioiden osia) käytetään usein yhdessä?
      - esim. spatiaalinen data
- valitun ratkaisun tehokkuus?
  - datan määrä?
  - datan kertymisnopeus?
  - harvat taulut: tilavaatimus?
  - tietojen käsittelyjärjestys: aikavaatimus?

Erityispiirteet: yhteenveto

- tarpeet poikkeavat yleisistä kaupallis-hallinnollisista tietokannoista, mm.
  - järjestetty data
  - sovelluskohtaiset operaatiot
  - puhdistaminen ja jalostaminen

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

- 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 huomioon ottamista

⇒ tilastolliset menetelmät

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

Joitain huomioitavia asioita

- Mittausvirheet
- Poikkeavat arvot ja niiden käsittely
- Lukujen esitystarkkuus

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

$\Rightarrow$ suuressa aineistossa keskimäärin yhtä paljon liian pieniä ja suuria arvoja

$\Rightarrow$ 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 voi vääristää analyysiin tuloksia
  - 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 herkempiä 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
- esim. pienet todennäköisyydet kannattaa esittää logaritmimuunnoksella (jolloin alkuperäisten lukujen kerto- ja jakolaskut suoritetaan muunnettujen lukujen yhteen- ja vähennyslaskuina)
- joudutaan tällä kurssilla ohittamaan tällä maininnalla

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

$\Rightarrow$ olisiko toisenlainen malli ollut hyödyllisempi

$\Rightarrow$ 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
- muuttuja NR on käsiteltävän rivin numero
- **awk '{print NR, $0}'** tulostaa rivinumeron ja rivin

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)

Yksinkertaisia unix-työkaluja

- `tr`: vaihtaa tietyn merkin esiintymät toiseksi (`-s` poistaa duplikaattimerkit)
- `sort`: lajittelee rivit (`-n`: numeerisesti, `-r`: käänteisesti, `-u`: duplikaatit poistaen)
- `uniq`: poistaa duplikaattirivit järjestetystä aineistosta (`-c`: tulostaa kunkin rivin lukumäärän)
- `head -n`: tulostaa n ensimmäistä riviä
- `tail -n`: tulostaa n viimeistä riviä
- `wc`: tulostaa merkkien (`-b`), sanojen (`-w`) ja rivien (`-l`) lukumäärät

Unix-työkaluja, jatkoa

- komentojen “putkittaminen” on usein kätevää
- esim:

```
tr -s ' ' '\n' < test.txt | sort | uniq -c \
| sort -nr | head -20
```

  tulostaa tiedoston test.txt 20 yleisintä sanaa
- komentotiedostot, ns. “shell scriptit”, ovat hyödyllisiä aineistojen käsittelyn uusimista, muuttamista ja dokumentointia varten
- ks. komentojen ohjeet, esim. `man tr`

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

⇒ **tunnuslukujen** (*statistic*) laskeminen

- yleisesti tunnusluku on mikä tahansa aineiston funktio $t = f(y_1, \ldots, y_n)$, jonka määräämiseksi on tunnettava kaikki havainnot $y_i, 1 \leq i \leq n$
- jatkossa merkinnällä $y_{(i)}$ tarkoitetaan i:nneksi pienintä havaintoa (siis kaikilla i: $y_{(i)} \leq y_{(i+1)}$)
- esimerkiksi havaintojen minimi- ja maksimiarvo $y_{(1)}$ ja $y_{(n)}$ ovat tunnuslukuja
- tunnusluku voi olla myös vektori

Helsingin yliopisto Tietojenkäsittelytieteen laitos

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))

Helsingin yliopisto Tietojenkäsittelytieteen laitos

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

Helsingin yliopisto Tietojenkäsittelytieteen laitos

Datan arvotyypit ja asteikot

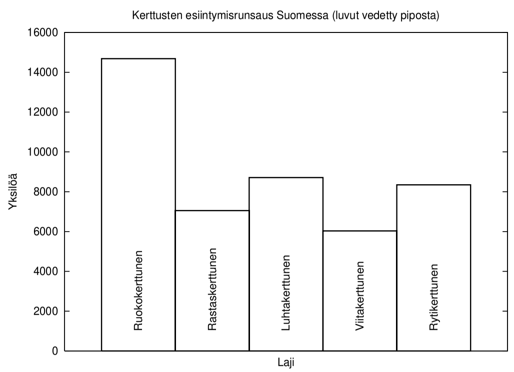
- aineiston muuttujilla (attribuuteilla) voi olla erityyppisiä arvoja
- muuttujan tyyppi vaikuttaa siihen
  - mitä tunnuslukuja voidaan laskea
  - mitä visualisointimenetelmiä kannattaa käyttää
  - mitä analyysimenetelmiä voidaan soveltaa

Kategoria-asteikko

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

Kategoria-asteikko, esimerkki

- attribuutin ’laji’ moodi on ’ruokokerttunen’



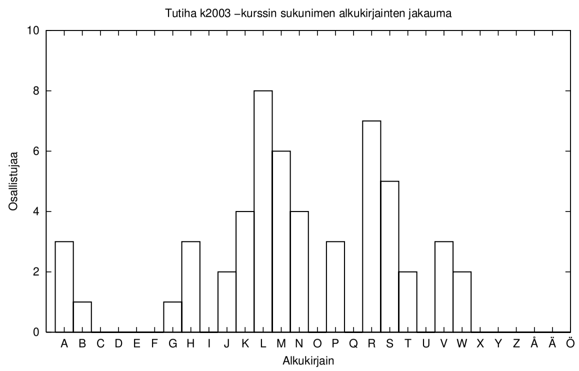
Ordinaaliasteikko

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



Numeeriset asteikot

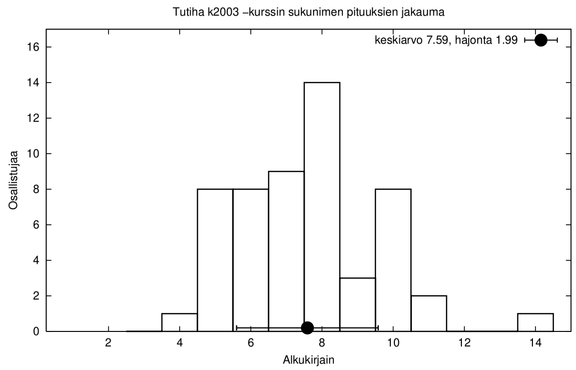
- seuraavilla kolmella asteikolla arvot numeerisia
- arvoilla voi mielekkäästi laskea
- arvot voivat olla diskreettejä tai jatkuvia

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

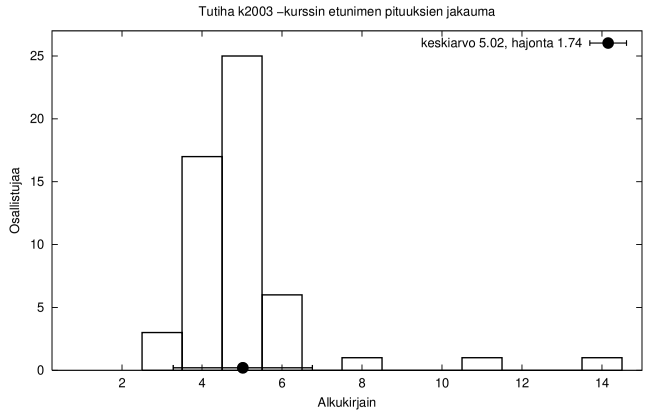
Esimerkki, absoluuttinen asteikko

- 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*

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

  - sopii numeerisille asteikoille
  - heikkous: yksikin huomattavasti poikkeava arvo (outlier) vaikuttaa paljon; vinot ja monihuippuiset jakaumat ongelmallisia

  ⇒ keskiarvo voi siirtyä kauaksi sieltä missä suurin osa havainnoista on

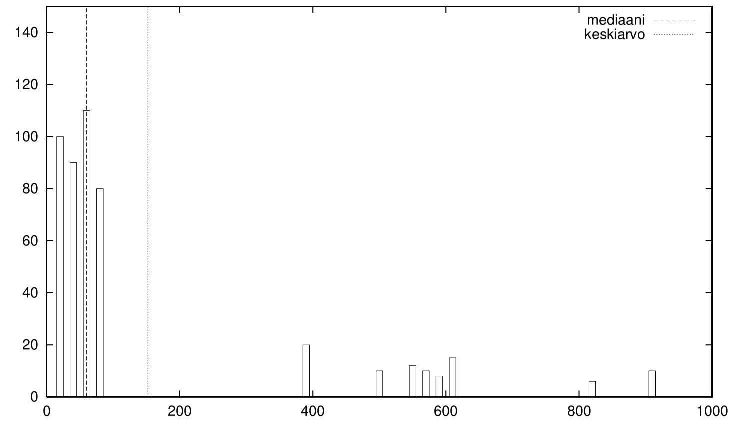
Arvojen keskittyminen (2)

- *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 ⇒ 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 ⇒ 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
  - vrt. mäkihypyn pistelasku, jossa viiden tuomarin tyylipisteistä pienin ja suurin jätetään huomiotta

Arvojen vaihtelevuus

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

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

  - *varianssi* on keskihajonnan neliö
  - keskihajonnan intuitiivinen merkitys: havainnon keskimääräinen etäisyys keskiarvosta

Fraktiilit

- otoksen *p-fraktiili* on pienin sellainen otoksen arvo, jota pienempiä tai yhtä suuria on osuus p otoksen arvoista
- = suurin sellainen otoksen arvo, jota suurempia tai yhtä suuria on osuus $1 - p$ otoksen arvoista
- mediaani on 0.5-fraktiili (eli 50%-fraktiili)
- 0.25-fraktiili ja 0.75-fraktiili ovat *kvartiileja*
- vähintään ordinaaliasteikon muuttujille
- fraktiilit eivät ole herkkiä poikkeaville arvoille

Fraktiilivälit

- *kvartiiliväli* on 2-ulotteinen tunnusluku (0.25-fraktiili, 0.75-fraktiili)
  - keskimmäiset 50% havainnoista ovat kvartiilivälillä
  - kvartiilivälin pituus 0.75-fraktiili – 0.25-fraktiili on aineiston hajontaa kuvaava tunnusluku, johon poikkeavat arvot eivät vaikuta
- *p-fraktiilivälillä* on osuus p otoksen arvoista
  - välin ala- ja yläpuolella likimain osuudet $(1-p)/2$ arvoista
- hypoteesin testauksessa tarkastellaan usein 0.95-, 0.99- ja 0.999-fraktiilivälejä

Väliyhteenveto

- yksittäisten muuttujien tarkastelu
- erilaiset mitta-asteikot
  - kategoria-asteikko
  - ordinaaliasteikko
  - numeeriset asteikot: intervalli-, suhde-, absoluuttinen asteikko
- tunnuslukuja
  - keskiarvo, mediaani, moodi
  - otoskeskihajonta
  - fraktiilit, fraktiilivälit

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 (.../tutihaK04/gnuplot_short.pdf) ja Gnuplot Central (http://www.gnuplot.info/)

Gnuplot (2)

- kukin syötetiedoston rivi vastaa yhtä (data)pistettä (kuvan piste tai histogrammin pylväs)
- seuraavassa esimerkissä histogrammia varten valmiiksi kurssin osallistujien nimistä laskettu tiedosto nimi.dat
  1. sarake: nimen pituus
  2. tämän pituisten etunimien lukumäärä
  3. tämän pituisten sukunimien lukumäärä

Datatiedosto nimi.dat

```
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

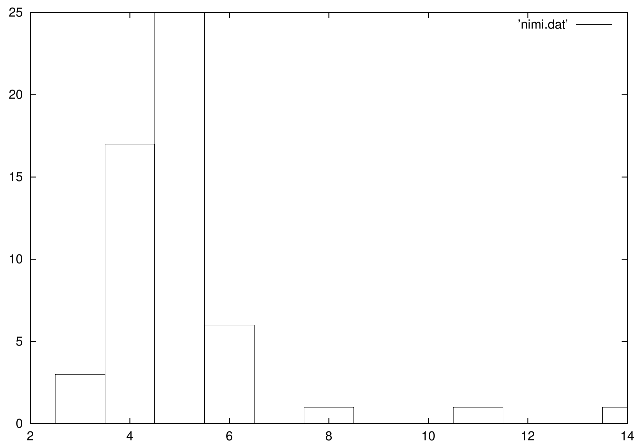
- histogrammin piirto

```
gnuplot> plot 'nimi.data' with boxes
```

- "with boxes" = piirrä datasta histogrammi
- oletusarvoisesti käytetään kahta ensimmäistä saraketta (arvo, lkm)-parina pylvään piirämiseen
  - (gnuplotin dokumentaation mukaan histogrammien kohdalla ei ole näin, mutta käytännössä näyttää olevan)
- gnuplot asettaa oletusarvoisesti x-akselin ja y-akselin skaalan automaattisesti (autoscale)

Histogrammit ja Gnuplot (2)

- `gnuplot> plot 'nimi.data' with boxes`



Histogrammit ja Gnuplot (3)

- akselien automaattinen skaalaus toimii huonosti histogrammien tapauksessa

$\Rightarrow$ aseta manuaalisesti: set xrange, set yrange

- 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> plot 'nimi.data'
```

tai

```
gnuplot> plot [0:15][0:30] 'nimi.data' with boxes
```


Plot-komento

- plot-komennolle voi antaa paljon lisämääreitä:

```
plot {<ranges>}
     {<function> | "<datafile>" {using <fields>}}
     {title "<title-text>"}
     {with <style>}
```

- title: käyrän tms. nimi (koko kuvan otsikko asetetaan komennolla *set title "<title>"*)
- fields: mitä syötteen sarakkeita käytetään graafisessa esityksessä (oletus: 1, 2, ...), esim.
  - `gnuplot> plot 'nimi.data' using 1:3 with boxes`
  - erilaiset esitystyypit vaativat eri määrän sarakkeita

Plot-komento (2)

- with: kaaviotyyli + lisämääreitä
  - oletusarvoinen viivan leveys on yleensä liian pieni

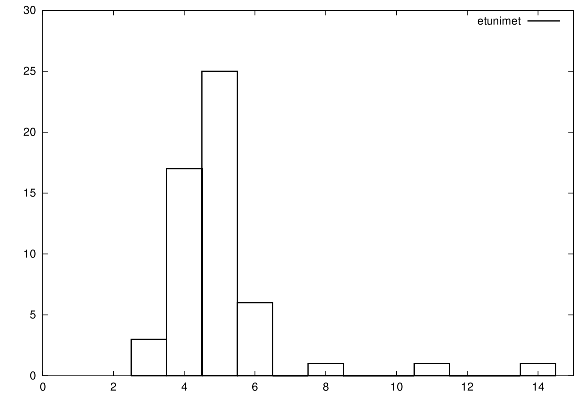
```
gnuplot> plot 'nimi.data' title 'etunimet' \
with boxes linewidth 3
```

- komentoja voi lyhentää

```
gnuplot> plot 'nimi.data' t 'etunimet' w boxes lw 3
```

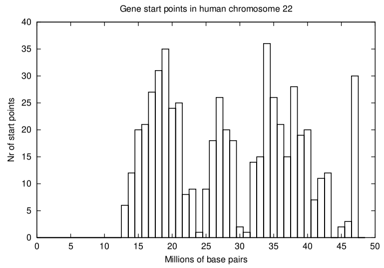

Histogrammit ja Gnuplot (2)

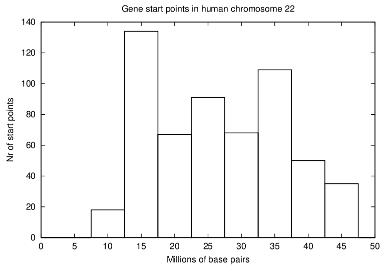
-
```
gnuplot> plot [0:15][0:30] 'nimi.data' \
title 'etunimet' with boxes linewidth 3
```



Histogrammit

- histogrammin antamaan vaikutelmaan vaikuttavat
  - y-akselin alaraja (0 on paras valinta)
  - luokkavälin pituus (=pylvään paksuus)





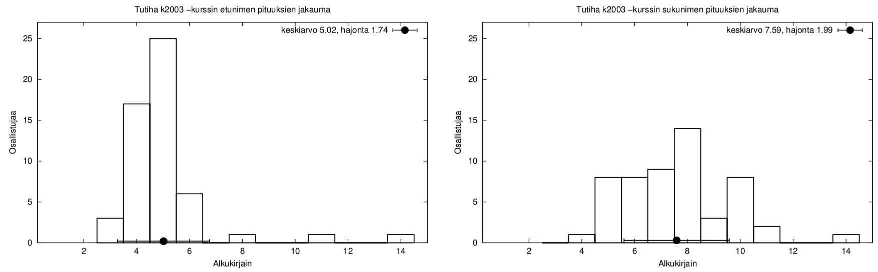
- pylvään leveyden valinta voi vaikuttaa oleellisesti

Histogrammit

- pylvään leveyden valintaan vaikuttavat:
  - 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 samoja asteikkoja



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

```
gnuplot> plot ...
gnuplot> replot ...
```

- lisätään seuraavaksi etunimien pituudet kuvaavaan histogrammiin keskiarvo pisteenä x-akselilla

Pisteiden piirtäminen

- olkoon tiedostossa avg.data yksi rivi

```
5.02 1.74 0.5
```

- 1. sarake keskiarvo, 2. sarake otoskeskihajonta, 3. sarake y:n arvo
- käytetään 1. ja 3. saraketta

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


Pisteiden piirtäminen (2)

- selkeitä pistetyyppejä (postscriptissa)
  - pistetyypit 6 ja 7 (avoin ja väritetty pyöreä piste)
  - pistetyypit 4 ja 5 (avoin ja väritetty neliö)
  - pistetyyppi 8 (kolmio)

```
gnuplot> replot 'avg.data' u 1:3 title 'keskiarvo' \
pointtype 7 pointsize 2
```

- piste- ja viivatyypit ovat erilaisia näytöllä ja postscriptissa

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)

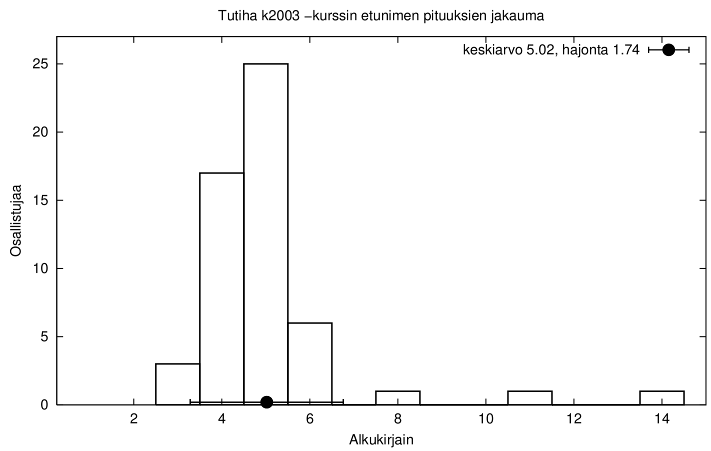
- datarivillä tarvitaan nyt kolmea saraketta:
  1. x-koordinaatti
  2. hajonnan leveys x-akselin (tai y-akselin) suunnassa pisteen (x,y) ympärillä
  3. y-koordinaatti

```
gnuplot> set data style xerrorbars
gnuplot> replot 'avg.data' using 1:3:2 \
title 'keskiarvo 5.02, keskihajonta 1.74'
```


Hajontapylväät (3)

```
gnuplot> set title "Tutiha k2003 -kurssin ..."
gnuplot> set xlabel "Alkukirjain"
gnuplot> set ylabel "Osallistujia"
gnuplot> plot [0:15][0:30] \
'nimi.data' notitle w boxes linewidth 3, \
'avg.data' using 1:3:2 \
  title 'keskiarvo 5.02, keskihajonta 1.74' \
  lt 2 lw 4 pt 7 ps 3
```


Koko kuva



Kuvien tallentaminen

- kuvien tulostaminen tiedostoon (PostScript)

```
gnuplot> set term postscript
gnuplot> set output "kuvatiedosto.ps"
gnuplot> plot ...
```

- nykyisen kuvan tallentaminen (eps-tiedostoon)

```
gnuplot> set term postscript eps
gnuplot> set output "kuvatiedosto.eps"
gnuplot> replot
```

- kuvien piirto taas ruudulle

```
gnuplot> set term X11
```


Asetusten tallentaminen

- ohjelman asetusten (ei kuvien) tallentaminen

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

- komentojen antaminen tiedoston kautta

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

- gnuplotin ajaminen unixin komentoriviltä

```
$ gnuplot tiedosto.set plotcommands.gpl
```


Eksploratiivinen data-analyysi, kahden muuttujan tarkastelu

Kahden muuttujan riippuvuus

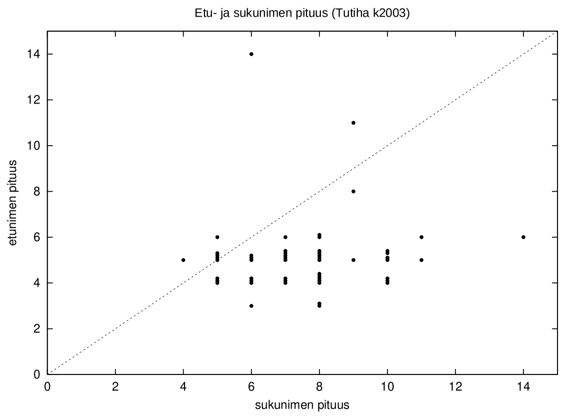
- kahden (satunnais)muuttujan X ja Y riippuvuus toisistaan
  - funktionaalinen riippuvuus
    = X:n arvo määrää yksikäsitteisesti Y:n arvon, tai päinvastoin
  - tilastollinen (stokastinen) riippuvuus
    = X:n arvo vaikuttaa Y:n arvon jakaumaan, tai päinvastoin
- funktionaalinen riippuvuus on luonnollisesti 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)
- `gnuplot> plot 'tiedosto.dat' with points \`
  `pointtype 5 pointsize 3`

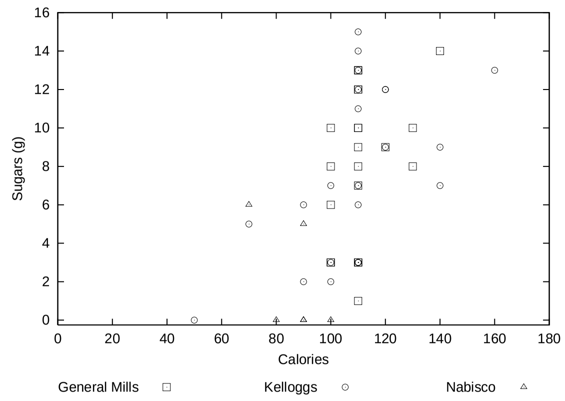
Sirontakuvio (*scatter plot*)

- 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

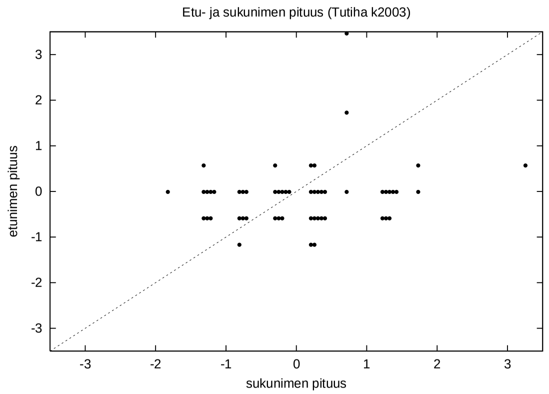
$$x_i' = \frac{(x_i - \bar{x})}{\sigma_x}, y_i' = \frac{(y_i - \bar{y})}{\sigma_y}$$

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)

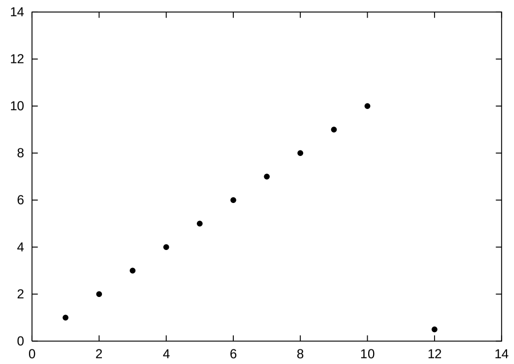


- $\rho(sukunimi, etunimi) \approx 0.07$

$\Rightarrow$ ei (lineaarista) riippuvuutta

Korrelaatiokerroin

- poikkeavan havainnon vaikutus



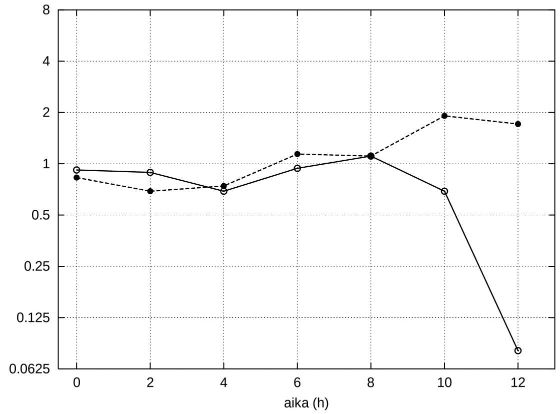
- yksi poikkeava havainto $(12, 0.5)$ riittää pudottamaan korrelaation arvoon 0.47

Mitä korrelaatio ei kerro?

- muuttujat voivat olla epälineaarisesti riippuvia, vaikka niiden välinen korrelaatio 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
  - 0.5: ainetta A on kaksi kertaa enemmän kuin ainetta B
  - 2: ainetta B on kaksi kertaa enemmän
- mm. tällaisissa tilanteissa on luontevaa esittää arvot käyttäen logaritmistä asteikkoa
- gnuplotissa

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

Gnuplot, viivadiagrammit

- selkeät pisteet, jotka yhdistetty viivalla

```
gnuplot> plot 'aikasarja' with linespoints
gnuplot> plot 'aikasarja' not w lp \
lt 3 lw 2 pt 7 ps 0.8
```

- pisteet, joiden kautta viiva kulkee (pisteitä ei piirretä)

```
gnuplot> plot 'aikasarja' with lines
gnuplot> plot 'aikasarja' not with l lt 3 lw 2
```


Komentotiedostot

- kokeiluja on helppo tehdä käsin komentoja kirjoittamalla
- analyysin toistaminen taas on helpompaa jos komennot on tallennettu
  - datan prosessointi päivityksen jälkeen
  - kuvan yksityiskohdan muuttaminen
  - datan käsittelyn tarkistaminen jälkikäteen
- usein koko prosessointiketju kannattaa automatisoida sitä mukaa kun tietää mitä haluaa tehdä
- awk, gnuplot, ...

Komentotiedostot (2)

- gnuplotin komentotiedostot unix-komentoriviltä:

```
$ gnuplot plotcommands.gpl
```

- awk-tiedostot suoraan unix-komentoriviltä:

```
$ awk -f awkcode.awk data.dat
```

- unix komentotiedostot:

```
#!/bin/bsh
awk -f awkcode.awk data.dat > plotdata.dat
gnuplot plotcommands.gpl
```


Ekspl. data-analyysi: yhteenveto

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