Tilastotieteen johdantokurssi

Heikki Hyhkö

kevät 2013

1. Johdanto

Kurssin sisältö

- Johdanto
- Tilastollinen tutkimus
  - Tutkimuksen vaiheet, otanta, mittaaminen
- Yksiulotteisen empiirisen jakauman esittäminen
  - Taulukointi, graafiset esitykset, tunnusluvut
- Kaksiulotteinen jakauma
  - Ristiintaulukointi, korrelaatio, regressioanalyysi, graafiset esitykset
- Todennäköisyyslaskentaa
  - Todennäköisyys, satunnaismuuttuja
- Todennäköisyysjakaumat ja estimointi
  - Todennäköisyysjakaumat, piste-estimointi, luottamusvälit
- Tilastolliset testit
  - z-testi, t-testi ja χ^2-testi

Mitä on tilastotiede?

- Tilastotiede on menetelmätiede, jolla on sovelluksia kaikilla numeerista tietoa käsittelevillä tieteenaloilla.
  - biometria/biostatistiikka, demometria, psykometria, ekonometria, data mining, stokastiikka ...
- Tilastotiede on empiiristä tutkimusta.
  - Tilastollinen tutkimus kohdistuu aina mitattaviin ilmiöihin.
  - Ilmiöiden mittaaminen voi olla joskus hankalaa.
    - ihmisen mielipiteet vs. ihmisen pituus
- Tilastotiede on päättelyä epävarmuuden vallitessa.
  - Sattuman vaikutus otetaan huomioon.
- Tilastotieteessä ratkaistaan ongelmia havaintoaineistoon nojaten.
  - Havaintoaineistot ovat yleensä suuria ...
    tai niiden ainakin toivottaisiin olevan.

Mitä tilastotiede ei ole

- Tilastotiede ei ole oppia tilastoista ja niiden tekemisestä.
  - Tilastotoimi käsittelee ja laatii tilastoja.
- Statistiikka: valtiota käsittelevä tieto/aineisto
  - 1700-luvun saksalainen statistik on lähellä nykyistä valtio-oppia.
  - Englannissa kehiteltiin jo 1600-luvulla poliittista aritmetiikkaa, jolle loppujen lopuksi lainattiin saksalainen nimitys statistics.
- Matematiikka: mittateoria ja stokastiikka
  - Tilastollisen teorian kehitys on matematiikkaa, mutta empiiristen aineistojen ansiosta tilastotiede ei ole pelkkää matematiikkaa.
- Tietojenkäsittelytiede: ohjelmistot ja simulointi
  - Tilastotieteilijän työ on usein tilastollista ohjelmointia ja aineistojen käsittelyä ja täten varsin tietokonevetoista.
  - Laskentatehon kasvettua on simuloinnin merkitys huomattava.
- Olisiko parempi nimitys tietojenkäsittelytiede? (prof. Seppo Mustonen)

Tilastotieteen esihistoriaa

- Egyptissä ja Kiinassa tehtiin jo viisituhatta vuotta sitten selvityksiä väestömääristä, myöhemmin myös Intiassa, Kreikassa, Persiassa, Roomassa (Census: asekuntoisten miesten lista).
  - Nämä laskennat koskivat lähinnä asekuntoisten miesten lukumääriä ja väestön veronmaksukykyä.
  - Keskiajan Euroopassa ja muuallakin tämä tilastointi otti takapakkia.
- Varsinainen tilastotoimi sai alkunsa 1500-luvulla Keski-Euroopassa.
  - Kehittyvä valtionhallinto tarvitsi yhä enemmän numeerista tietoa.
  - Myös laajeneva kauppa- ja vakuutustoiminta vaati tarkempaa dataa.
- Todennäköisyyslaskennan kehityksen alkukimmokkeena olivat uhkapelien luomat laskentaongelmat 1600- ja 1700-luvulla Ranskassa.
- Samaan aikaan kehittyvä empiirinen luonnontiede kaipasi välineitä mittauksissa ilmenneiden satunnaisvirheiden hallintaan.
- Näistä tarpeista ja aatoksista muotoutui loppujen lopuksi tilastotiede.

Tilastotieteen historiaa

- Varhaisia innovaatioita (1600-1800 -luvulla)
  - Suurten lukujen laki (Jacob Bernoulli)
  - Keskeinen raja-arvolause (Abraham de Moivre)
  - (Käänteinen) ehdollinen todennäköisyys (Thomas Bayes)
  - Normaalijakauman ja satunnaisvirheiden yhteys (C.F. Gauss)
- 1800-luvun loppua pidetään modernin tilastotieteen alkuna.
  - Korrelaatiokäsite ja regressioanalyysin ajatus (Francis Galton)
  - Satunnaisluvut ja todennäköisyysjakaumat (Karl Pearson)
- Vasta 1900-luvulla mukaan astui varsinainen matemaattinen todennäköisyysteoria ja tieteellinen päättely.
  - t-jakauma (Student, eli William Gosset)
  - SU -menetelmä ja varianssianalyysi (R.A. Fisher)
  - Todennäköisyysteorian aksiomatisointi (Andrei Kolmogorov)

Mihin tilastotiedettä tarvitaan?

- Tilastotiede yrittää selittää ja ennustaa reaalimaailman ilmiöitä.
- Tilastotiede tavoittelee objektiivista ratkaisua ongelmiin.
- Tilastotieteen perusasiat kuuluvat yleissivistykseen.
  - tilastollinen/matemaattinen suhteellisuudentaju
  - poikkeavuus/muutos suhteessa satunnaisvaihteluun
- Tilastotieteen "väärinkäyttö"
  - Tahatonta, jos ei ymmärretä, mitä tehdään.
    - Tulkitaan kuvioita/testejä kaavamaisesti
    - Ei ymmärretä otannan merkitystä
  - Tahallista, jos pyritään vaikuttamaan esimerkiksi ihmisten asenteisiin.
    - Esitetään johdattelevia kysymyksiä
    - Valikoidaan vastaajia
    - Esitetään harhaanjohtavia grafiikoita

2. Havaintoaineiston hankinnasta

2.1 Tilastollisen tutkimuksen vaiheet

1. Tutkimusongelma
   - Suunnittelu
     - kysymyksenasettelu, ongelman rajaus, tulosten tarkkuus
2. Aineiston hankinta
   - otanta
   - kontrolloidut kokeet vs. havainnointi
   - valmiit aineistot
     - tilastolaitokset (Tilastokeskus), rekisterivirastot: (THL, Tulli...)
3. Mittaaminen
   - mittauksen kohde
   - mitattava ominaisuus
   - käytettävä mittari
   - mittarin hyvyys

Tilastollisen tutkimuksen vaiheet jatkoa

4. Havaintoaineiston käsittely
   - tietojen tallentaminen
   - aineistoon tutustuminen
   - aineiston muokkaaminen
5. Tilastolliset analyysit
   - aineiston kuvaaminen/esittäminen
     - taulukot/matriisit, graafiset esitykset
     - tilastollisten tunnuslukujen laskeminen
   - tilastollisten menetelmien käyttö
     - tilastollinen testaaminen
     - tilastollinen mallintaminen
6. Tulosten tulkinta, johtopäätökset ja tulosten esittäminen
   - tulosten luotettavuus
   - tulosten yleistettävyys

Tutkimustyypeistä

- Kokeellinen tutkimus
  - Koeyksiköille tehdään kontrolloituja kokeita eli käsittelyjä.
  - Koeyksiköiden olosuhteita muutetaan halutun määrän verran.
  - Muiden tekijöiden vaikutus pyritään minimoimaan.
  - Tutkitaan miten koeyksiköt reagoivat muutokseen.
  - Koe- ja vertailuryhmä valitaan satunnaisesti.
- Havainnoiva tutkimus
  - Tutkimuksen kohteista kerätään suoria havaintoja.
  - Kohteiden olosuhteisiin ei aktiivisesti vaikuteta.
  - Seurataan miten kohteet reagoivat erilaisiin olosuhteisiin.
  - Tutkimuksen kohteet tulee valita satunnaisesti perusjoukosta.

Kyselytutkimus vs. fysikaalinen mittaus

Kyselyyn liittyviä ongelmia:

- haastattelutapa
- otantamenetelmä
- kyselylomake

Lomakkeeseen liittyviä asioita

- Mitä, miten ja keneltä?
- Kyselyn kesto ja selkeys?
- Tarvitaanko esitutkimus?
- Mikä on kyselyn aikataulu?
- Kadon ennakointi.

Fysik. mittauksen ongelmia:

- mittaustapa
- otantamenetelmä
- mittalaite

Mittalaiteeseen liittyviä asioita

- Mitä, miten ja missä mitataan?
- Vaikuttavatko olosuhteet?
- Mikä on mittaustarkkuus?
- Mikä on sallittu vaihtelu?
- Toistettavuus?

Koejärjestelystä

- Klassinen koejärjestely
  1. Alkumittaus
  2. Ryhmiinjako
  3. Käsittely
  4. Loppumittaus
  5. Johtopäätös
- Sovitetut eli kaltaistetut parit
  - Pyritään valitsemaan taustatekijöiltään samanlaiset koehenkilöt
  - Esim. Kaksostutkimukset
- Kaksoissokkokoe
  - Sekä tutkittava että tutkija ovat tietämättömiä siitä, saako potilas plaseboa vai lääkettä.
  - Halutaan poistaa ns. plasebovaikutus.

2.2 Otanta

- Perusjoukko on joukko, josta tutkimuksessa halutaan tietoa.
  - Voidaan kutsua myös populaatioksi tai kohdeperusjoukoksi.
  - esim. suomalaiset, alle 5-vuotiaat lapset, diabeetikot...
- Kehikkoperusjoukko on joukko, joka on käytettävissä otantaan.
  - esim. Perusjoukko olkoon suomalaiset. Tehdään postikysely. $\rightarrow$ Kehikkoperusjoukko on ne suomalaiset, joiden osoitteet tiedetään.
- Tilastoyksikkö on se tilastollinen perusyksikkö, josta tietoa kerätään.
  - Voi olla joko havainto- tai koeyksikkö.
  - esim. henkilö, koululuokka, kunta, sairaala, kuulalaakeri...
- Perusjoukon alkio on yksi perusjoukon yksikkö.
- Otosyksikkö on otokseen poimittu alkio.

Aito otos

- Kolme kriteeriä:
  1. Otokseen valittavat havaintoyksiköt on valittava siitä perusjoukosta, johon tutkimus kohdistuu.
  2. Havaintoyksiköt valitaan satunnaisotantaa käyttäen, niin ettei valitsija vaikuta tulokseen.
  3. Jokaisella perusjoukon alkeisyksiköllä on yhtä suuri mahdollisuus tulla valituksi otokseen (edustavuus).
- Jos nämä kriteerit eivät täyty, on kyseessä *näyte*.
- Aito otos $=$ edustava otos $\approx$ otos $\neq$ näyte.
  - Valitettavasti sanaa otos käytetään usein "tutkimusten" yhteydessä, vaikka kyse olisi valikoituneesta vastaajajoukosta.
  - Tällä kursilla otoksella tarkoitetaan aitoa otosta.

Perusjoukko vs. otos

- Perusjoukon tunnusluvut ovat tietoa.
  - Mitataan kaikki koripallojoukkueen pelaajat.
    - Tällöin tiedetään mikä on kyseisen joukkueen pituuden keskiarvo, keskihajonta, moodi ...
- Otoksen tunnusluvut ovat arvioita.
  - Valitaan satunnaisesti 6 koripallojoukkuetta, joista mitataan pelaajat
    - Tällöin voidaan arvioida koripalloliigan pelaajien pituuden keskiarvoa, keskihajontaa, moodia ...
- Perusjoukon tunnuslukujen epätarkkuus koostuu mittausvirheestä.
- Otoksen tunnusluvuissa on lisäksi otannasta johtuvaa satunnaisuutta.
  - Eli mittaus- ja otantavirheiden lisäksi normaali satunnaisvaihtelu on otettava huomioon, joka onkin tilastotieteen keskeinen tehtävä.

Otanta vs. kokonaistutkimus

- kokonaistutkimus
  - Tutkitaan jokainen perusjoukon alkio.
- otantatutkimus
  - Valitaan perusjoukosta otos eli satunnainen osa perusjoukosta.
  - Johtopäätökset perusjoukosta tehdään otoksen perusteella.
  - Jotta voidaan yleistää, niin otoksen on siis oltava edustava.
- otantatutkimuksen etuja:
  - edullisempi ja nopeampi
  - usein myös ainoa vaihtoehto
- Jos kyseessä ei ole aito otos (eli otanta oikein suoritettu), niin tilastollisilla tuloksilla ei ole mitään arvoa!
  - poikkeuksena tietysti kokonaistutkimukset

Otoskoko

- Otantasuhde $= \frac{\text{otoskoko}}{\text{perusjoukon koko}} = \frac{n}{N}$
- Ei ole yksiselitteistä tapaa määrätä tarvittava otoskoko.
  - Otoskoon täytyy kuitenkin olla "riittävä", jotta saadut tulokset voidaan yleistää koko perusjoukkoa koskeviksi.
- Otoskoko riippuu myös halutusta tarkkuusvaatimuksesta.
  - Tulosten tarkkuus ei kuitenkaan tietyn otoskoon jälkeen enää merkittävästi kasva.
- Mitä heterogeenisempi perusjoukko, sitä suurempi otos vaaditaan.
- Otantakustannukset asettavat myös omat rajansa.
- Luottamusvälien yhteydessä saamme kaavan otoskoon määrittämiseksi.

Otannan virhemahdollisuudet

- Kato
  - Eräkato
    - Otokseen valitulta alkeisyksiköltä ei saada jotain tietoa.
  - Yksikkökato
    - Otokseen valitusta alkeisyksiköstä ei saada mitään tietoja.
- Otantavirhe
  - vajaapeittävyys (tai alipeittävyys)
    - Otoksesta jäävät pois tietynlaiset alkeisyksiköt, jolloin otos ei edusta koko perusjoukkoa.
  - ylipeittävyys
    - Otoksessa on sinne kuulumattomia alkiota.
- Mittausvirhe (ei varsinaisesti otantavirhe)
  - vastausharha
    - Vastaukset ovat syystä tai toisesta virheellisiä.

Yksinkertainen satunnaisotanta (YSO)

- Kaikilla perusjoukon alkeisyksiköillä on yhtä suuri todennäköisyys tulla valituksi otokseen.
- Edut ja haitat
  - **+** yksinkertainen ja helppo
  - **+** ennakkotietoa perusjoukosta ei tarvita
  - **-** käytännössä vaikea toteuttaa
- YSO on useimpien otantamenetelmien lähtökohtana, vaikka sillä ei sellaisenaan olekaan kovin laajaa käyttöä.
- Eri otantamenetelmien (hajonta)estimaattien hyvyyttä testattaessa käytetään YSO'a vertailukohtana.
  - Ositettua otantaa käytetäänkin juuri estimoinnin parantamiseen.
- Systemaattista otantaa ja ryväsotantaa käytetään itse otannan suorittamisen helpottamiseen ja/tai kustannusten vähentämiseen.

Systemaattinen otanta (SO)

- Perusjoukon alkioiden täytyy olla järjestyneinä jonoon jonkin ominaisuuden perusteella.
- Järjestys voi liittyä tutkimuksen kohteena olevaan ominaisuuteen.
- Otoksen poiminta:
  1. Määrätään poimintaväli $k = \frac{N}{n}$.
  2. Valitaan satunnainen lähtöpiste väliltä $(1, k)$.
  3. Poimitaan lähtöpisteestä eteenpäin joka k:s alkio.
     - esim. poimitaan liukuhihnalta joka 10. tuote $(k = 10)$
- Edut ja haitat
  - **+** helppo ja nopea
  - **+** etukäteistiedon huomioonottaminen
  - **-** Vaarana se, että perusjoukossa ominaisuus vaihtelee jaksollisesti.
- Käytetään monissa rekisteripohjaisissa tutkimuksissa.

Ryväsotanta (RO)

- Alkeisyksiköt muodostavat ryhmiä eli rypäitä, joista poimitaan otos
- Ryväs muodostuu usein vierekkäisistä alkioista.
- Valitusta otoksesta tutkitaan kaikki alkeisyksiköt tai suoritetaan rypään sisällä uusi otanta.
- Edut ja haitat
  - **+** kustannustehokas
  - **-** suurempi otosvirhe
- Moniasteinen otanta: Perusjoukon hierarkkinen rakenne
  - esim. Suomen kunnat -> koulut -> luokat -> oppilaat
- Ryväsotanassa valittaisiin siis vain osa kunnista/kouluista/luokista, jolloin syntyisi sekä rahallista että ajallista säästöä.

Ositettu otanta (OO)

- Jaetaan perusjoukko ositteisiin tietyn ominaisuuden perusteella.
- Valitaan jokaisesta ositteesta alkeisyksiköt YSO:lla tai SO:lla.
- Päätetään ositteiden kiintiöt:
  - Tasainen kiintiöinti: Poimitaan jokaisesta ositteesta yhtä monta alkeisyksikköä.
  - Suhteellinen kiintiöinti: Poimitaan ositteista alkeisyksiköitä samassa suhteessa kuin niitä on perusjoukossa.
  - Optimikiintiöinti: Poimitaan suurempi otos ositteista, joissa on suuri vaihtelu.
- Edut ja haitat
  - **+** etukäteistiedon huomioonottaminen
  - **+** estimoinnin paraneminen
  - **-** käyttökelvoton mielekkään jakavan ominaisuuden puuttuessa

2.3 Mittaaminen

- Tilastotieteessä mittaamisella tarkoitetaan lukuarvon liittämistä tutkimuskohteeseen.
  - esim: sukupuoli, puolueen kannatus, henkilön mielipide, huoneen lämpötila, tavaran paino, väkiluku, maan vetovoimakiihtyvyys
- Mittaustarkkuus
  - Eri mitta-asteikoilla päästään erilaiseen mittaustarkkuuteen, joka vaikuttaa siihen mitä menetelmiä voidaan käyttää.
  - Mittaus kannattaa tehdä niin tarkasti kuin voidaan, mutta tulokset tulee esittää järkevällä tarkkuudella satunnaisvaihtelu huomioiden.
  - Riittävä mittaustarkkuus on hieman tarkempi, kuin toistomittauksessa samana pysyvä tarkkuus.
    - esim. Ihmisen pituuden mittaus: sentin tarkkuudella.
      - Aamulla ihminen voi olla sentin pidempi kuin illalla.

Mittauksen hyvyys

- Validiteetti (oikeellisuus)
  - Mittauksen täytyy olla validi eli mittaus kohdistuu siihen, mitä halutaan mitata.
  - Validi mittari esittää oikein mitattavaa ominaisuutta.
- Reliabiliteetti (luotettavuus)
  - Mittauksen täytyy olla luotettava ja tarkka eli reliaabeli.
  - Reliaabeli mittari antaa toistettaessa samanlaisen tuloksen.
- Harhattomuus (bias)
  - Mittaus on harhaton, jos se ei systemaattisesti yli- tai aliarvioi mitattavan ominaisuuden todellista arvoa.

Validi mittari on luotettava ja harhaton, mutta luotettava ja harhaton mittari ei välttämättä ole validi!

Mitta-asteikot

- Mitta-asteikot järjestettynä hienoimmasta karkeimpaan:
  - (absoluuttinen asteikko)
  4. relatiivi- eli suhdeasteikko
  3. intervalli- eli välimatka-asteikko
  2. ordinaali- eli järjestysasteikko
  1. nominaali- eli laatueroasteikko
  - (dikotominen asteikko)
- Mitta-asteikot ovat sisäkkäisiä, siten että muuttujat voidaan aina uudelleen luokitella karkeamilla asteikoilla.
  - esim. kengän sovitus "mittaus"-tilanteena:
    1. Vain yksi kengän koko: kenkä sopii, kenkä ei sovi
    2. Kolme kokoa: small, medium, large
    3. Perinteiset kengännumerot
    4. Mittatilauskengät

Erityisasteikot

- Dikotominen asteikko
  - Havainnot on luokiteltavissa kahteen luokkaan.
  - Yleensä luokat koodataan numeroilla 0 ja 1, jolloin saadaan ns. binäärinen muuttuja.
  - Esim. Dummy-muuttujat ovat yleensä dikotomisia.
  - Karkein mitta-asteikoista
- Absoluuttinen asteikko
  - Mitta-asteikko on yksikäsitteinen ja sitä ei voi muuntaa.
  - Lukumäärille ja suhteellisille osuuksille tarkoitettu asteikko.
  - Nollakohta on aina lukittu (0% tai 0 kappaletta).
    - Suhteellisissa osuuksissa ylärajakin on lukittu (100%).
  - Lukumäärissä yläraja on avoin, mutta määritelty (numeroituva ääretön).
  - Muuttujat ovat suhdeasteikollisia.

1. Laatuero- eli nominaaliasteikko

- Alkiot luokitellaan tiettyihin ennalta määrättyihin luokkiin ominaisuuksiensa perusteella.
- Kahdesta eri alkeisyksiköstä voidaan todeta vain kuuluvatko ne samaan luokkaan vai eivät (a=b).
- Kutsutaan myös luokitteluasteikoksi.
- esim. sukupuoli, ammatti, diagnoosi, . . .
- Invariantti luokituserot säilyttäville muutoksille.
  - Luokkatunnukset voi melko vapaasti vaihtaa.

2. Järjestys- eli ordinaaliasteikko

- Luokat voidaan järjestää ominaisuutensa mukaan paremmuusjärjestykseen (a<b).
- Luokkien välimatkoja ei voida määritellä.
- Huom! Keskiarvon laskeminen on siis kielletty.
- esim. Likertin asteikko, koulutustaso, . . .
- Invariantti järjestyksen säilyttäville muutoksille.
  - Luokkien järjestyksen säilyttävät muutokset ovat sallittuja.

3. Välimatka- eli intervalliasteikko

- Voidaan ilmaista havaintoyksiköiden mitattavan ominaisuuden välisiä eroja tai paremmuutta tiettyinä mittayksikköinä (a-b).
- Kutsutaan myös väliasteikoksi.
- Nollapiste ja mittayksikkö ovat vapaasti valittavissa.
- esim. lämpötila celsius-asteissa, vuosiluvut, . . .
- Invariantti etäisyyden säilyttäville muutoksille.
  - Yhteen- ja vähennyslasku ovat sallittuja.

4. Suhde- eli relatiiviasteikko

- Voidaan laskea mitattavan ominaisuuden arvojen välisiä suhteita (a/b).
- Asteikolla on absoluuttinen nollakohta, jossa mitattavaa ominaisuutta ei ole yhtään.
- Suhdeasteikollinen muuttuja saa vain ei-negatiivisia arvoja.
- esim. pituus, paino, väkiluku, . . .
- Invariantti etäisyyssuhteet säilyttäville muutoksille.
  - Kerto- ja jakolasku ovat sallittuja.

Diskreetti vs. jatkuva

- Jatkuva muuttuja voi saada minkä tahansa arvon joltain väliltä.
  - mittaustarkkuus määrittää tarkan arvon
- Diskreetti muuttuja voi saada vain tiettyjä arvoja.
  - kaikki laatuero- ja järjestysasteikolliset muuttujat
  - lisäksi myös lukumäärät ja niihin vertautuvat muuttujat
- Ongelmana jatkuvat muuttujat, joiden mittari/mittaustarkkuus määritellään diskreetiksi.
  - esim. palkka euroina, ikä vuosina.
  - Mieti onko väliin jäävillä arvoilla mielekäs tulkinta.
    - tuntipalkka 15.746 euroa, mielekäs $\rightarrow$ jatkuva
    - ikä 27.462 vuotta, epämielekäs $\rightarrow$ diskreetti
  - Tulkinta ei välttämättä ole yksikäsitteinen, sillä tilastografiikan kannalta pitäisin molempia muuttujia jatkuvina.

Muuttuja vs. ominaisuus

- Ominaisuuksista saadaan mittaamalla muuttujia!
  - esim. Ihmisellä on ominaisuuksia: pituus, silmien väri, sukupuoli...
    - Näistä ominaisuuksista saadaan vastaavia muuttujia mittaamalla.
  - Huom. Nummenmaan tulkinta muuttujasta poikkeaa yllä mainitusta määritelmästä, josta seuraa ero muuttujien jatkuvuuden määrittelylle.
- On olemassa jatkuvia ominaisuuksia, joista mitattaessa tulee diskreettejä muuttujia.
  - Mielipiteen vahvuus on mitä ilmeisimmin tyypiltään jatkuvaa. Tästä ei kuitenkaan seuraa, että mielipide muuttujana olisi jatkuva.
  - esim. Pidän enemmän suklaasta kuin salmiakista?
    - Vastaus 2.478, ei mielekästä tulkintaa $\rightarrow$ diskreetti.
- Tällä kurssilla ei olla kiinnostuneita ominaisuuksien jatkuvuudesta, vaan siitä, onko ominaisuudesta mitattu muuttuja jatkuva vai diskreetti.

Kvanti vs. kvali

- On olemassa sekä kvantitatiivista (määrällistä) tutkimusta että kvalitatiivista (laadullista) tutkimusta.
  - Kaikki tilastollinen tutkimus on kvantitatiivista.
- Tilastotieteessä on kuitenkin kvantitatiivisia ja kvalitatiivisia muuttujia!
- Kvantitatiivinen muuttuja saa automaattisesti numeroarvon.
  - Suhde- ja välimatka-asteikolliset muuttujat ovat kvantitatiivisia.
- Kvalitatiiviselle muuttujalle annetaan (koodataan) numeroarvo.
  - Laatueroasteikolliset muuttujat ovat kvalitatiivisia.
  - Järjestysasteikollisia muuttujia pidetään yleensä kvalitatiivisina, vaikka niillä onkin kvantitatiivisia ominaisuuksia.
- Kvantitatiivisiä muuttujia kutsutaan myös numeerisiksi muuttujiksi.
- Kvalitatiivisiä muuttujia kustutaan myös kategorisiksi muuttujiksi.

Muuttujien ominaisuuksia

- Jatkuvien muuttujien kuvaajana käytetään histogrammia.
- Diskreettien muuttujien kuvaajana käytetään pylväsdiagrammia.
- Kvantitatiivisia muuttujia testataan parametrisilla testeillä.
  - Jolloin täytyy tehdä jakaumaoletuksia.
- Kvalitatiivisia muuttujia testataan epäparametrisilla testeillä.
  - Jolloin ei tarvitse tehdä jakaumaoletuksia.
- Muuttujatermistöä
  - vaste = selitettävä muuttuja = dependent variable = (riippuva muuttuja)
  - selittäjä = selittävä muuttuja = independent variable = (riippumaton muuttuja)
  - satunnaismuuttuja = random variable = stochastic variable
  - vakio(muuttuja)= constant variable

3. Yksiulotteisen empiirisen jakauman esittäminen

3.1 Havaintomatriisi

- Havaintoaineisto on mittausten tuloksena saatu aineisto eli data, jossa havaintoyksikköjen ominaisuuksiin on liitetty arvot.
- Havaintoaineistossa laatuero- ja järjestysasteikolliset muuttujat koodataan numeroiksi.
  - Huom! Näillä luvuilla ei saa suorittaa laskutoimenpiteitä!
- Välimatka- ja suhdeasteikolliset muuttujathan saavat arvonsa automaattisesti.
- Havaintoaineisto esitetään yleensä taulukkona (matriisina) tai graafisena esityksenä.
- Havaintomatriisiin yhteyteen liitetään tiedot aineistosta ja muuttujista.
  - Tätä aineistoa kuvaavaa informaatiota kutsutaan metadataksi.

Havaintomatriisin peruskäsitteitä

- tilastoyksikkö
  - yksittäinen tutkimuksen kohde
- muuttuja
  - tilastoyksiköstä mitattu ominaisuus
- jakauma (jakaumavektori)
  - kaikki muuttujan saamat arvot
  - Muuttujan jakauma muodostaa havaintomatriisin sarakkeen.
- profiili (havaintovektori)
  - tilastoyksikön kaikkien muuttujien arvot
  - Tilastoyksikön profiili muodostaa havaintomatriisin rivin.
- solu
  - matriisin yksittäinen ruutu
  - yksittäinen muuttujan saama havaintoarvo

Havaintomatriisi

- Esimerkkinä toimikoon harjoitusaineistomme ylänurkka:

havaintonro	sp	ikä	siv	pituus	$muuttuja$
1	0	49	2	167	
2	1	49	2	170	
3	1	45	$solu$	175	
4	1	50	2	180	$\leftarrow profiili$
5	0	25	2	165	
$tilastoyksikko$		$jakauma \uparrow$			

3.2 Frekvenssijakauma

- Yhden muuttujan jakaumaa kuvataan frekvenssijakaumalla eli suoralla jakaumalla.
- luokan i frekvenssi $(f_i) =$ havaintojen lukumäärä muuttujan luokassa i.
- frekvenssien summa $=$ kokonaisfrekvenssi
- Esimerkkinä: Läänien kuntalukumäärät v. 2003

Lääni	Frekvenssi (f_i)
Etelä-Suomi	88
Länsi-Suomi	204
Itä-Suomi	66
Oulu	50
Lappi	22
Ahvenanmaa	16
Yhteensä	446

Esimerkki

20 opiskelijaa saivat tentissä seuraavat pistemäärät:

$$0, 0, 0, 0, 0, 0, 1, 1, 1, 2, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6.$$

Muodostetaan pistemäärien frekvenssijakauma:

Pistemäärä	0	1	2	3	4	5	6	Yhteensä
Frekvenssi (f_i)	6	3	1	0	0	5	5	20

Summafrekvenssit (F_i)

- Jos muuttuja on vähintään *järjestysasteikollinen*, voidaan laskea summa- eli kumulatiiviset frekvenssit.
- Summafrekvenssillä tarkoitetaan luokan ylärajaan mennessä kertyneiden havaintojen määrää:

$$F_i = \text{luokan i frekvenssi} + \text{edellisten luokkien frekvenssien summa}$$

- Kaikkien luokkien yhteenlaskettu summafrekvenssi on havaintojen lukumäärä:

$$F_k = n$$

  - $k =$ luokkien lukumäärä, $n =$ havaintojen lukumäärä

Esimerkki

Pistemäärä	Frekvenssi (f_i)	Summafrekvenssi (F_i)
0	6	6
1	3	$6 + 3 = 9$
2	1	$6 + 3 + 1 = 10$
3	0	$6 + 3 + 1 + 0 = 10$
4	0	$6 + 3 + 1 + 0 + 0 = 10$
5	5	$6 + 3 + 1 + 0 + 0 + 5 = 15$
6	5	$6 + 3 + 1 + 0 + 0 + 5 + 5 = 20$
Yhteensä	20	

Prosenttifrekvenssit (p_i)

- Lasketaan jokaisen luokan frekvenssin prosenttiosuus kokonaisfrekvenssistä.

$$p_i = 100(\frac{f_i}{n})$$

- Prosenttifrekvenssit ilmoitetaan joskus myös desimaalilukuina (suhteelliset frekvenssit).
  - Onkin yleensä aivan samantekevää ilmoitetaanko suhteelliset osuudet murtolukuina, desimaalilukuina vai prosentteina.

- Huomatkaa ero prosentin ja prosenttiyksikön välillä!
  - Maksimipisteiden suhteellinen osuus kasvaa 20 prosenttia: $25\% + 20/100 * 25\% = 25\% + 5\% = 30\%$
  - Maksimipisteiden suhteellinen osuus kasvaa 20 prosenttiyksikköä: $25\% + 20\% = 45\%$

Esimerkki

Pistemäärä	Frekvenssi (f_i)	Prosenttifrekvenssi (p_i)
0	6	$100(\frac{6}{20}) = 30\%$
1	3	$100(\frac{3}{20}) = 15\%$
2	1	$100(\frac{1}{20}) = 5\%$
3	0	$100(\frac{0}{20}) = 0\%$
4	0	$100(\frac{0}{20}) = 0\%$
5	5	$100(\frac{5}{20}) = 25\%$
6	5	$100(\frac{5}{20}) = 25\%$
Yhteensä	20	100%

Summaprosenttifrekvenssit (P_i)

- Luokan ylärajaan asti kertyneiden prosenttifrekvenssien summa.

$$P_i = 100(\frac{F_i}{n})$$

Pistemäärä	f_i	p_i	Summaprosenttifrekvenssi (P_i)
0	6	30%	30%
1	3	15%	30% + 15% = 45%
2	1	5%	45% + 5% = 50%
3	0	0%	50% + 0% = 50%
4	0	0%	50% + 0% = 50%
5	5	25%	50% + 25% = 75%
6	5	25%	75% + 25% = 100%
Yhteensä	20		

Muuttujan luokittelusta

- Jos muuttuja on luonteeltaan *jatkuva*, on taulukointia varten suoritettava muuttujan luokittelu.
- Esim. On mitattu 20 opiskelijan pituudet ja saatu seuraavat tulokset:

$$152, 155, 155, 157, 160, 162, 162, 169, 171, 171,$$
$$172, 174, 176, 179, 181, 185, 187, 192, 195, 202$$

- Luokitellussa ei ole yksikäsitteisiä sääntöjä, paitsi
  - Havainto saa kuulua vain yhteen luokkaan.
  - Jos havaintoja jätetään pois, se tulee syineen kertoa.
- esim. Jaetaan aineisto 10 sentin levyisiin luokkiin:

150-159, 160-169, 170-179, 180-189, 190-199, 200-209

  - Tässä luokkien väliin on jäävinään yksi sentti, mikä seuraa mittaustarkkuudesta.

Esimerkki

Pyöristetyt luokkarajat	Varsinaiset luokkarajat	Luokkakeskukset
$150 - 159$	$149.5 - 159.5$	154.5
$160 - 169$	$159.5 - 169.5$	164.5
$170 - 179$	$169.5 - 179.5$	174.5
$180 - 189$	$179.5 - 189.5$	184.5
$190 - 199$	$189.5 - 199.5$	194.5
$200 - 209$	$199.5 - 209.5$	204.5

- Varsinaiset luokkarajat valitaan aina tarkemmiksi kuin mittaustarkkuus.
- Vaihtoehtoinen merkintätapa on:
  150-160, 160-170, 170-180, 180-190, 190-200, 200-210
  - Tällöin yläraja ei kuulu luokkaan.
  - Varsinaiset luokkarajat ovat kuitenkin samat.

Luokkien lukumäärä (k)

- Tavoitteena on saada mahdollisimman havainnollinen kuva jakaumasta.
  - Jos luokkia on liikaa, niin aineistosta ei saa hyvää yleiskuvaa.
  - Jos luokkia on liian vähän, niin menetetään liikaa yksityiskohtia.
- Luokkien lukumäärä riippuu:
  - havaintojen lukumäärästä
  - jakauman muodosta
- Pyritään välttämään tyhjiä luokkia (nollaluokka).
- Ensimmäinen ja/tai viimeinen luokka ei saa olla tyhjä luokka ilman erityisen painavaa syytä.
  - esim. luokitus aiemmasta tutkimuksesta
- Pienissä aineistoissa luokkien lukumäärä on välillä $\sqrt[3]{n}$ ja $\sqrt{n}$.
- Yleensä siis n. 5-10 luokkaa riippuen otoskoosta.

Luokan pituus (c)

- Muuttuja on luokiteltava niin, että luokat peittävät koko vaihteluvälin (min, max) ja ne eivät saa peittää toisiaan.
- Luokan pituus (c) on vaihteluvälin pituus $(R = max - min)$ jaettuna luokkien lukumäärällä (k):

$$c = \frac{R}{k}$$

- Esimerkissä $c = \frac{R}{k} = \frac{202-152}{6} = \frac{50}{6} = 8.333\ldots \approx 10$
- Tasavälinen luokittelu
  - Luokan pituus on jokaisessa luokassa sama.
- Epätasavälinen luokittelu
  - Luokan pituus vaihtelee.

Luokkarajat ja luokkakeskus

- Luokat esitetään joko pyöristettyjen tai varsinaisten luokkarajojen avulla.
- Suositus on, että käytetään luokkarajoina lukuja, jotka ovat samaa mittaustarkkuutta kuin alkuperäiset havainnot.
- Varsinaiset luokkarajat aikaansaadaan siten, että mittaustarkuuden perään lisätään 5.
  - Tällöin mikään havainto ei voi ikinä kuulua kahteen luokkaan!
  - esim. Ihmisten pituuden mittaustarkkuus on senttimetri, joten varsinaiset luokkarajat asetetaan puolen sentin tarkkuudella.
- Avoimessa luokassa toinen luokkaraja on kiinnittämätön.
  - Avoimia luokkia tulisi välttää.
- Luokkakeskusta laskettaessa käytetään varsinaisia luokkarajoja:

$$\text{Luokkakeskus} = \frac{(\text{luokan yläraja} + \text{luokan alaraja})}{2}.$$

3.3 Tilastollinen grafiikka

- Graafisia esityksiä muuttujan jakaumasta kutsutaan myös kuvioiksi.
- Kuvio valitaan muuttujan tyypin mukaan (jatkuva vai diskreetti).
- Laatuero- ja järjestysasteikon muuttujat ovat diskreettejä muuttujia.
- Välimatka- ja suhdeasteikon muuttujat ovat yleensä jatkuvia muuttujia,
  - mutta poikkeuksiakin on, kuten lukumäärät.
- Hyvän kuvion ominaisuuksia:
  - Otsikko, joka kertoo mitä, missä ja milloin
  - Akselien nimet ja mittayksiköt selvästi näkyvissä
  - Tarpeelliset akselien katkaisut on tehtävä ja merkittävä

Diskreetti muuttuja

Pylväsdiagrammi

- Koostuu suorakulmioista, joiden korkeuksina ovat luokkien frekvenssit tai prosenttifrekvenssit.
- Jos muuttuja on vähintään järjestysasteikollinen, niin luokkien järjestystä ei saa sekoittaa.
- Piirretään tavallisesti niin, että muuttujan arvot tulevat vaaka-akselille ja frekvenssit pystyakselille.
- Pylväät ovat aina saman levyisiä ja irti toisistaan!
  - Pylväät ovat irti toisistaan, koska muuttuja voi saada arvoja vain tietyissä pisteissä.
- Varjostukset ovat turhia ja yleensä haitallisia.

Jatkuva muuttuja

Histogrammi

- Piirretään luokitellusta aineistosta varsinaisten luokkarajojen mukaan.
  - Tilasto-ohjelmat eivät välttämättä toimi näin.
- Kuin pylväsdiagrammi, mutta pylväät ovat kiinni toisissaan.
  - Pylväät ovat kiinni toisissaan, koska muuttuja voi teoriassa saada minkä tahansa arvon vaihteluväliltä.
- Koostuu suorakulmioista, joiden kantoina ovat luokkarajat ja korkeuksina frekvenssit.
- Verrataan pylväiden pinta-aloja keskenään!
- Epätasavälisessä luokittelussa pylväät ovat eri levyisiä.

Kertymäfunktion kuvaaja

- Empiirinen kertymäfunktio on porrasfunktio, jonka kuvaaja nousee jokaisen havainnon kohdalla ja pysyy havaintojen välillä ennallaan.
- Kertymäfunktio voidaan piirtää joko suhteellisien osuuksien tai lukumäärien mukaan:
  - Suhteelliset osuudet (todennäköisyydet):
    - Funktion kuvaaja lähtee nollasta ja saavuttaa viimeisen havainnon kohdalla ykkösen.
  - Lukumäärät (frekvenssit):
    - Funktion kuvaaja lähtee nollasta ja saavuttaa viimeisen havainnon kohdalla havaintojen kokonaislukumäärän (n).
- Jatkuvan muuttujan kertymäfunktion kuvaajaa kutsutaan myös summakäyräksi.

Piirakkakuvio

- Kuvaa muuttujan arvojen jakautumista ympyrän sektoreiden pinta-alojen mukaan.
  - Kuvaa yhtä lailla sekä suhteellisia osuuksia että frekvenssejä.
  - Toimii hyvin, jos karkea arvio suhteellisista osuuksista on riittävä tai erot ovat suuria.
  - Kuvioon kannattaa sisällyttää prosenttiosuudet myös lukuina.
  - Useissa tapauksissa kuitenkin pylväiden korkeuksia/pinta-aloja on helpompi verrata kuin sektoreiden pinta-aloja.
  - Piirrettäessä useita piirakoita, voidaan piiraiden koot suhteuttaa luokkien kokoihin.
  - Piirakkadiagrammia ei saa koskaan kallistaa, sillä se vääristää pinta-alasuhteita!

Muita kuvioita

- Janadiagrammi
  - Pylväsdiagrammi, jossa pylväät korvataan janoilla.
  - Käytetään lähinnä pistetodennäköisyyksien kuvaamiseen.
- Ikäpyramidi
  - Kahden vastakkain piirretyn histogrammin yhdistelmä.
- Frekvenssimurtoviiva/frekvenssimonikulmio
  - Jatkuvan muuttujan esitystapa, jossa kuvitteellisen histogrammin luokkakeskukset yhdistetään toisiinsa viivoilla.
- Box-plot eli jana-laatikko -diagrammi
  - Kuvaa samalla sekä muuttujan keskittymistä että hajaantumista.
  - Tätä käsitellään tarkemmin viisilukuisen -yhteenvedon yhteydessä, jonka perusteella jana-laatikko -diagrammi piirretään.

3.4 Tunnuslukuja

- Tunnuslukujen avulla on tarkoitus tiivistää aineistosta saatua informaatiota.
- Tunnusluvut kuvaavat muuttujan arvojen jakauman perusominaisuuksia.
- *Keskiluvut* kuvaavat muuttujan keskimääräisten, tyypillisten tai yleisten arvojen paikkaa aineistossa.
- *Hajontaluvut* kuvaavat sitä, kuinka voimakkaasti havainnot ovat keskittyneet vastaavan keskiluvun ympäristöön.
- Muut tunnusluvut kuvaavat mm. jakauman vinoutta ja huipukkuutta.
- Tunnusluvut eivät aina ole yhtä informatiivisia kuin kuviot.

3.4.1 Keskilukuja

- Suhdeasteikko
  - Geometrinen keskiarvo
  - Harmoninen keskiarvo
- Välimatka-asteikko
  - Aritmeettinen keskiarvo
- Järjestysasteikko
  - Mediaani
- Laatueroasteikko
  - Moodi

Moodi eli tyyppiarvo (mo)

- Sen luokan mittaluku tai symboli, jolla on suurin frekvenssi.
- Moodi voidaan määrittää kaikilla mitta-asteikolla.
- Jatkuvalla (luokitellulla) muuttujalla moodi on sen luokan luokkakeskus, jolla on suurin frekvenssi.
- Moodeja/moodiluokkia voi aineistossa olla useampia kuin yksi.
- Moodi on robusti, eli vakaa.
  - ts. poikkeavat havainnot eivät vedä sitä puoleensa.
- esimerkkejä:
  1. $1, 1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 4, 5, 6$: mo$= 4$
  2. S,S,S,M,L,L,XL,XL : mo=S
  3. $1, 1, 2, 2, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 6$: mo$= 2, 4$

Moodi luokitellusta aineistosta

$luokkakeskus$	154.5	164.5	174.5	184.5	194.5	204.5
$frekvenssi$	4	4	6	3	2	1

- Moodiluokka on $170 - 179$ ja moodi tämän luokan luokkakeskus 174.5.
- Jatkuvan muuttujan tapauksessa moodi voidaan määritellä tarkemminkin kuin moodiluokan luokkakeskuksena:

$$mo = L_{mo} + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot c$$

$L_{mo} =$ moodiluokan varsinainen alaraja
$\Delta_1 =$ moodiluokan ja edeltävän luokan frekvenssien erotus
$\Delta_2 =$ moodiluokan ja seuraavan luokan frekvenssien erotus
$c =$ luokkavälin pituus.

- $mo = 169.5 + \frac{(6-4)}{(6-4)+(6-3)} \cdot 10 = 173.5$

Mediaani (md/me)

- Suuruusjärjestykseen järjestetyn aineiston keskimmäinen havainto, mikäli havaintoja on pariton määrä.
- Jos havaintoja on parillinen määrä, on mediaani jompi kumpi keskimmäisistä havainnoista ...
  - tai kahden keskimmäisen havainnon keskiarvo, jos sen saa laskea.
- Mediaanin voi määrittää vähintään järjestysasteikolliselle muuttujalle.
- Jatkuvalla muuttujalla mediaani voidaan määritellä myös mediaaniluokan luokkakeskuksena.
- Mediaani on robusti, eli vakaa.
- Esim. pituudet:

$$152, 155, 155, 157, 160, 162, 162, 169, 171, 171,$$
$$172, 174, 176, 179, 181, 185, 187, 192, 195, 202.$$

  - Keskimmäiset havainnot ovat $10.$ ja $11.$ havainto eli 171 ja 172.
  - Siis $md = \frac{171+172}{2} = 171.5$.

Mediaani luokitellusta aineistosta

$luokkakeskus$	154.5	164.5	174.5	184.5	194.5	204.5	$\sum$
$frekvenssi$	4	4	6	3	2	1	20

- Mediaaniluokka on $170 - 179$ ja mediaani luokan luokkakeskus 174.5.
- Jatkuvan muuttujan tapauksessa mediaani voidaan määritellä tarkemminkin kuin mediaaniluokan luokkakeskuksena:

$$md = L_{md} + \frac{n/2 - \Sigma f_i}{f_{md}} \cdot c$$

$L_{md} =$ mediaaniluokan varsinainen alaraja
$f_{md} =$ mediaaniluokan frekvenssi
$\Sigma f_i =$ mediaaniluokkaa edeltävien luokkien summafrekvenssi
$c =$ luokkavälin pituus.

- $md = 169.5 + \frac{(20/2-(4+4))}{6} \cdot 10 = 172.833$

Summasymboli $\sum$ ja sen ominaisuuksia

- Lukujonoa, jossa on n kappaletta numeroita, merkitään: $x_1, x_2, \ldots, x_n$
- Näiden lukujen summa on: $\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$.
- Summasymbolin laskennallisia ominaisuuksia:
  - $\sum_{i=1}^{k} f_i x_i = f_1 x_1 + f_2 x_2 + \ldots + f_k x_k$
  - $\sum_{i=1}^{n} a = na$
  - $\sum_{i=1}^{n} a x_i = a \sum_{i=1}^{n} x_i$
  - $\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$

Aritmeettinen keskiarvo ($\overline{x}$)

- Kun ihmiset normaalisti puhuvat keskiarvosta, he tarkoittavat tällä aritmeettista keskiarvoa!
- Keskiarvon laskemiseen vaaditaan vähintään välimatka-asteikollinen muuttuja.
- Keskiarvoa laskettaessa lasketaan havaintoarvot yhteen ja jaetaan saatu summa havaintoarvojen lukumäärällä.

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

- Esim. pituudet:

$$152, 155, 155, 157, 160, 162, 162, 169, 171, 171,$$
$$172, 174, 176, 179, 181, 185, 187, 192, 195, 202$$

$$\overline{x} = \frac{152+155+155+157+160+\ldots+192+195+202}{20} = 172.85$$

Keskiarvo luokitellusta aineistosta

- Jos muuttujan arvot on luokiteltu, voidaan keskiarvo laskea frekvenssien ja luokkakeskusten avulla.

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i x_i}{n} = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_k x_k}{n}$$

$luokkakeskus$	154.5	164.5	174.5	184.5	194.5	204.5	$\sum$
$frekvenssi$	4	4	6	3	2	1	20

$$\overline{x} = \frac{4 \cdot 154.5 + 4 \cdot 164.5 + 6 \cdot 174.5 + 3 \cdot 184.5 + 2 \cdot 194.5 + 1 \cdot 204.5}{20} = 173.5$$

Keskiarvon ominaisuuksia

- Välimatka- ja suhdeasteikollisten muuttujien keskiluku.
- Keskiarvo ei ole robusti, sillä poikkeavat havainnot vetävät sitä puoleensa.
- Keskiarvopiste sijaitsee muuttujan arvojen painopisteessä.
- Kun kaikkiin havaintoihin lisätään vakio, niin keskiarvokin kasvaa saman vakion verran.
- Kun kaikki havainnot kerrotaan vakiolla, niin keskiarvokin kertautuu samalla vakiolla.
- Keskiarvo on ensimmäinen origomomentti.

Moodi, mediaani ja keskiarvo

Näitä kolmea lukua vertailemalla voidaan saada selville jotain (yksihuippuisen) jakauman vinoudesta.

- $mo < md < \overline{x}$
  - jakauma on vino oikealle
- $mo \approx md \approx \overline{x}$
  - jakauma on symmetrinen
- $\overline{x} < md < mo$
  - jakauma on vino vasemmalle

Vinon aineiston keskiluku

- Toisinaan heikomman mitta-asteikon tunnusluvut antavat paremman kuvan havaintojen keskittymisestä.
- Suhdeasteikolliselle vinolle muuttujalle kannattaa keskilukuna käyttää mediaania, jos halutaan saada aito käsitys muuttujan keskittymisestä.
- Esim. Palkkajakaumat ovat vinoja:
  - Sillä on paljon pieniä ja keskisuuria palkkoja, mutta vain vähän palkkoja, jotka ovat todella suuria.
  - Koska keskiarvo ei ole robusti, nämä suuret palkat nostavat keskiansiota huomattavasti korkeammaksi, kuin se aidosti on.
  - Joten palkka-aineistoja laskettaessa, kannattaa laskea mediaani aritmeettisen keskiarvon sijaan.

Harmoninen keskiarvo (H)

- Käytetään ”käänteislukujen” keskiarvojen laskemiseen.

$$H = \frac{n}{\sum_{i=1}^{n}\left(\frac{1}{x_i}\right)}$$

- Esim. Henkilö ajaa samat matkat nopeuksilla: 60, 120 ja 40 km/h.

Keskinopeus on: $H = \frac{3}{\left(\frac{1}{60}+\frac{1}{120}+\frac{1}{40}\right)} = 60$ (km/h)

- Laajennus eri pituisille matkoille:

$$H = \frac{m}{\sum_{i=1}^{n}\left(\frac{f_i}{x_i}\right)}, \text{ jossa } m = \sum_{i=1}^{n} f_i$$

- Esim. 20 km 60 km/h, 40 km 120 km/h, 10 km 40 km/h:

$$H = \frac{70}{\left(\frac{20}{60}+\frac{40}{120}+\frac{10}{40}\right)} = 76.37 \text{ (km/h), jossa } m = 20+40+10 = 70$$

Geometrinen keskiarvo (G)

- Geometrista keskiarvoa käytetään suhdelukujen keskiarvojen laskemiseen.

$$G = \sqrt[n]{\prod_{i=1}^{n} x_i}, \text{ jossa } \prod_{i=1}^{n} x_i = x_1 \cdot x_2 \cdot \ldots \cdot x_n$$

- Esim. Lainan korot ovat: 1.v. 5 %, 2.v. 10 %, 3.v. 15 %, 4.v. 5 %.
  - $G = \sqrt[4]{1.05 \cdot 1.10 \cdot 1.15 \cdot 1.05} = 1.0867$
  - Keskikorko on siis 8.67%
- Geometrista keskiarvoa kutsutaan myös keskiverroksi.
- Geometrinen ja harmoninen keskiarvo voidaan laskea vain suhdeasteikollisille muuttujille.
- Geometrinen ja harmoninen keskiarvo eivät ole robusteja.

3.4.2 Hajontalukuja

- Suhdeasteikko
  - Variaatiokerroin
- Välimatka-asteikko
  - Keskihajonta
  - Varianssi
  - Keskiarvon keskivirhe
  - Kvartiilipoikkeama
- Järjestysasteikko
  - Vaihteluväli
  - Kvartiiliväli
- Laatueroasteikko
  - Entropia

Entropia (H)

- Entropia, eli satunnaisuusaste on laatueroasteikollisten muuttujien ainoa ja suhteellisen harvoin käytetty hajontaluku.
- Entropia kuvaa kuinka tasaisesti havainnot ovat jakautuneet luokkiin.
- Mitä pienempi entropia on, sitä enemmän havainnot ovat keskittyneet muutamiin luokkiin.
  - Entropia vaihtelee 0'n ja $\log_2(k)$'n välillä. (k= luokkien lkm.)

$$H = -\sum_{i=1}^{k} p_i \cdot \log_2(p_i)$$, jossa p_i on luokan i suhteellinen frekvenssi.

- Hieman edellistä parempi hajontamitta on suhteellinen entropia.

$$H_s = \frac{-\sum_{i=1}^{k} p_i \cdot \log_2(p_i)}{\log_2(k)}$$

  - Suhteellinen entropia vaihtelee välillä 0 ja 1.

Ala- ja yläkvartiilit (Q_1 ja Q_3)

- Kvartiilit jakavat suuruusjärjestykseen järjestetyn aineiston osiin.
- kvartiilit: jakavat aineiston neljään osaan
  - alakvartiili Q_1: 25% havainnoista pienempiä
  - keskikvartiili Q_2: 50% havainnoista pienempiä (= mediaani)
  - yläkvartiili Q_3: 75% havainnoista pienempiä
- tertiilit: jakavat aineiston kolmeen osaan
- kvintiilit: jakavat aineiston viiteen osaan
- desiilit: jakavat aineiston kymmeneen osaan
  - 4. desiili: 40% havainnoista pienempiä
- prosenttipisteet: jakavat aineiston sataan osaan
  - 15. prosenttipiste: 15% havainnoista pienempiä

Vaihtelu- ja kvartiiliväli

- Vaihtelu- ja kvartiilivälin voi laskea vähintään järjestysasteikolliselle muuttujalle.
- Vaihteluväli on havaintopari, joka koostuu pienimmästä ja suurimmasta havainnosta. $W = (min, max)$
- Kvartiiliväli on havaintopari, joka koostuu ala- ja yläkvartiilista. (Q_1, Q_3)
  - Kvartiiliväli on siis väli, joka peittää puolet havainnoista.
- Edellä mainitut arvot eivät välttämättä ole lukuja, vaan ne voivat myös olla luokkien arvoja.
  - esim. W=(täysin eri mieltä, täysin samaa mieltä)
- Vähintään välimatka-asteikollisille muuttujille voidaan laskea myös ...
  - Vaihteluvälin pituus: $w = max - min$
  - Vastaavasti voidaan laskea kvartiilivälin pituus: $IQR = Q_3 - Q_1$
  - Kvartiilivälin pituuden puolikasta kutsutaan kvartiilipoikkeamaksi: $Q = \frac{Q_3 - Q_1}{2}$

5-lukuinen yhteenveto

- Viisilukuinen yhteenveto vähintään järjestysasteikolliselle muuttujalle on seuraava:

$$(min, Q_1, Me, Q_3, max)$$

- Kun (vähintään välimatka-asteikollisesta) luokitellusta aineistosta lasketaan viisilukuinen yhteenveto, niin minimi on alimman luokan alaraja ja maksimi on ylimmän luokan yläraja.
- Vastaavasti ala- ja yläkvartiileja sekä mediaania laskettaessa voidaan käyttää kyseisten luokkien luokkakeskuksia.
- Ala-, keski- ja yläkvartiilien sijainnit:
  - Q_1: $\frac{n}{4} + \frac{1}{2}$, Q_2: $\frac{n}{2} + \frac{1}{2}$, Q_3: $\frac{3 \cdot n}{4} + \frac{1}{2}$
- Välimatka-asteikolliselle muuttujalle voi 5-lukuisen yhteenvedon perusteella piirtää ns. jana-laatikko -diagrammin (boxplot-kuvio) .

Keskihajonta (σ)

- Keskihajonta kertoo, kuinka paljon havaintoarvot keskimäärin poikkeavat keskiarvosta.

Oppilas	Matematiikka	Englanti	$\overline{x}$	σ
Matti	4	10	7	3
Maija	9	5	7	2
Kaisa	6	8	7	1
Kalle	7	7	7	0

- Huom. Kouluarvosanat ovat järjestysasteikollisia muuttujia, joista kuitenkin säännöllisesti lasketaan keskiarvoja!
- Keskihajonnan (kuten keskiarvonkin) saa oikeasti laskea vain välimatka- ja suhdeasteikollisille muuttujille, joten tämä yleinen käytäntö on väärä!

Keskihajonta (σ) vs. otoskeskihajonta (s)

- Perusjoukon keskihajontaa merkitään σ'lla ja otoskeskihajontaa s'llä.
- Perusjoukon keskihajontaa laskettaessa jakajana on n, kun taas otoskeskihajontaa laskettaessa jakajana on $(n-1)$.
- Tällä kurssilla (ja varsin usein muutenkin) termillä keskihajonta viitataan otoskeskihajontaan.
- Varianssi (σ^2) on keskihajonnan (σ) toinen potenssi.
- Otosvarianssi (s^2) on otoskeskihajonnan (s) toinen potenssi.
- Varianssi on toinen keskusmomentti, johon tällä kurssilla törmätään lähinnä normaalijakauman toisena parametrina.
- Keskihajonta ja varianssi eivät voi saada negatiivisia arvoja!

Keskihajonta (s)

- Olkoon meillä jälleen lukujono, jossa on n lukua:

$$x_1, x_2, \ldots, x_n$$

- Keskihajonnan teoreettisen määritelmän mukainen kaava:

$$s = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

- Eli jokaisesta havainnoista vähennetään keskiarvo ja tämä erotus neliöidään. Neliöidyt erotukset lasketaan yhteen, jonka jälkeen saatu luku jaetaan vapausasteillaan ja lopuksi otetaan vielä neliöjuuri.

Esimerkki

- Aineisto muodostuu seuraavista havainnoista: $2, 5, 7, 8, 13$
- Aineiston keskiarvo on $\overline{x} = \frac{2+5+7+8+13}{5} = 7$
- Muodostetaan taulukko:

x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
2	-5	25
5	-2	4
7	0	0
8	1	1
13	6	36
$\sum$	0	66

- $s = \sqrt{\frac{1}{5-1} \cdot 66} = \sqrt{16.5} \approx 4.06$

Keskihajonta (s)

- Käytetään samaa aineistoa, jossa on n lukua:

$$x_1, x_2, \ldots, x_n$$

- Parempi kaava otoskeskihajonnan käsinlaskentaan:

$$s = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2\right]}$$

- Kaava on suora johdos aikaisemmasta ja tuottaa riittävällä laskutarkkuudella samat tulokset.
- Erityisesti tämän kaavan hyvyys tulee ilmi myöhemmin korrelaatiokerrointa laskettaessa.

Esimerkki

- Aineisto muodostuu seuraavista havainnoista: $2, 5, 7, 8, 13$
- Muodostetaan taulukko:

	x_i	x_i^2
	2	4
	5	25
	7	49
	8	64
	13	169
$\sum$	35	311

- $s = \sqrt{\frac{1}{5-1}\left[311 - \frac{35^2}{5}\right]} = \sqrt{16.5} \approx 4.06$

Keskihajonta luokitellusta aineistosta

- Keskihajonnan laskeminen luokitellusta aineistosta:
- $s = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n} f_i x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} f_i x_i)^2\right]}$

$luokkakeskus$	154.5	164.5	174.5	184.5	194.5	204.5
$luokkakeskus^2$	23870.25	27060.25	30450.25	34040.25	37830.25	41820.25
$frekvenssi$	4	4	6	3	2	1

- $s = \sqrt{\frac{1}{20-1}\left[(606025) - \frac{1}{20}(3470)^2\right]} \approx 14.473$
- Luokittelemattomasta aineistosta laskettu keskihajonta:
  - $s = \sqrt{\frac{1}{20-1}\left[(601459) - \frac{1}{20}(3457)^2\right]} \approx 14.357$
  - Hieman tarkuutta menetettiin, mutta ei kovin merkittävästi.

Keskihajonnan ominaisuuksia

- $s = 0$, jos kaikki havainnot ovat samoja.
- Mitä suurempia havaintojen poikkeamat keskiarvosta ovat, sitä suurempi on s.
- Kun kaikki havainnot kerrotaan samalla luvulla, niin keskihajontakin kertautuu samalla luvulla.
- Keskihajonta on invariantti siirron suhteen, eli kun jokaiseen havaintoon lisätään sama luku, niin s pysyy ennallaan!
- Keskihajonnan neliötä kutsutaan *varianssiksi* s^2.
- Keskihajonta ei ole robusti, sillä poikkeavat havainnot kasvattavat keskihajontaa.
- Keskihajonta saa edelleenkin vain positiivisia arvoja.

3.4.3 Muita tunnuslukuja

- Keskiarvon keskivirhe (standard error) kuvaa sitä, kuinka paljon otoksesta laskettuun keskiarvoon sisältyy virhettä.

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

- Variaatiokertoimella V voidaan vertailla eri suuruusluokkaa olevien suhdeasteikollisten muuttujien hajontaa.

$$V = \frac{s}{\overline{x}}$$

- (Absoluuttinen) keskipoikkeama on harvinaisempi hajontaluku.

$$MAD = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}$$

- Huom. MAD = Mean absolute deviation, mutta MAD on myös Median absolute deviation, jolloin kaavassa keskiarvo korvataan mediaanilla.

Vinous ja huipukkuus

- Vinous (k_3 = kolmas keskusmomentti) kuvaa jakauman symmetrisyyttä.
  - Jakauma on vino oikealle, kun $\gamma_1 > 0$ ja vino vasemmalle, kun $\gamma_1 < 0$.
  - Kun $\gamma_1 = 0$, niin jakauma on symmetrinen.

$$\gamma_1 = \frac{k_3}{s^3}, \text{ missä } k_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3$$

- Huipukkuus (k_4 = neljäs keskusmomentti) kuvaa jakauman huipun leveyttä.
  - Huipun leveyttä verrataan normaalijakaumaan, jonka $\gamma_2 = 0$.
  - Jakauma on huipukas, kun $\gamma_2 > 0$ ja lattea, kun $\gamma_2 < 0$.

$$\gamma_2 = \frac{k_4}{s^4} - 3, \text{ missä } k_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4$$

- Yksinkertaisempi vinousmitta on Pearsonin P.
  - Jakauma on vino oikealle, kun $P > 0$ ja vino vasemmalle, kun $P < 0$.

$$P = \frac{\overline{x} - Mo}{s}$$

Tunnusluvut mitta-asteikoittain

	Laatueroasteikko	Järjestysasteikko
Keskiluvut	mo	mo, md
Hajontaluvut	H_s	$H_s, W, (Q_1, Q_3)$

•

	Välimatka-asteikko	Suhdeasteikko
Keskiluvut	$mo, md, \overline{x}$	$mo, md, \overline{x}, H, G$
Hajontaluvut	w, W, Q, s, se	w, W, Q, s, se, V

4. Kaksiulotteinen jakauma

Riippuvuus

- Jokaisesta tilastoyksiköstä tarkastellaan nyt kahta muuttujaa.
- Tutkitaan kahden muuttujan välistä yhteyttä eli riippuvuutta.
  - graafisesti (korrelaatiodiagrammi/hajontakuvio)
  - numeerisesti (korrelaatiokertoimet)
- Muuttujan x arvot tuntemalla yritetään ennustaa muuttujan y arvoja.
  - Miten alkoholijuomien kokonaiskulutus riippuu niiden hintatasosta?
  - Miten keuhkosyöpään sairastumisen todennäköisyys riippuu tupakoinnin kestosta?
- Tilastollinen tutkimus ei kerro kumpi muuttujista selittää kumpaa.
  - Painolla voi ennustaa pituutta.
- Riippuvuuden löytäminen ei kuitenkaan vielä todista syy-seuraus -suhteen olemassaoloa.

Moniulotteisuus

- Tavallisesti tutkittavaan ilmiöön liittyy useampia osatekijöitä ja sattuma.
  - Miten keuhkosyöpään sairastumisen todennäköisyys riippuu tupakoinnin kestosta, määrästä ja aloitusiästä?
  - Mitkä ovat perintötekijöiden ja elämäntapojen vaikutusosuudet?
- Kuinka hyvin moniulotteisia asioita pystytään mittaamaan/mallintamaan?
  - Kuinka hyvin bruttokansantuote mittaa hyvinvointia?
  - Kuinka hyvin työttömyysaste mittaa hyvinvointia?
  - Miten työttömyysaste riippuu bruttokansantuotteen kasvuvauhdista?

Syy-seuraus -suhde

- Jos muuttujien välillä ei ole todellista syy-yhteyttä, vaikka tilastolliset tarkastelut viittaavat yhteyksien olemassaoloon, yhteyksiä on tapana kutsua *näennäisyhteyksiksi*.
- Taustalla on yleensä kolmas tekijä, joka aiheuttaa riippuvuuden.
  - Mitä enemmän kuukauden aikana syödään jäätelöä, sitä enemmän ko. kuussa hukkuu ihmisiä. (kuukausi sekoittava tekijä)
  - Mitä enemmän paloautoja tulee palopaikalle, sitä suuremmat ovat palovahingot. (palon suuruus sekoittava tekijä)
  - Amerikkalaiset pörssianalyytikot ovat väittäneet, että auringonpilkkujen määrillä ja pörssikursseilla on voitu havaita tilastollista riippuvuutta. (sattuma)

4.1 Taulukot

- Kahden muuttujan riippuvuutta voidaan tarkastella ristiintaulukon avulla.

- Jatkuvan muuttujan tapauksessa taulukointi edellyttää luokittelua.
- Riippuvuutta voidaan tarkastella χ^2-tunnusluvun avulla (palataan myöhemmin).

	Syöpä	Ei syöpää	$\sum$
Tupakoi	37	2963	3000
Ei tupakoi	13	6987	7000
$\sum$	50	9950	10000

4.2 Grafiikka

Korrelaatiodiagrammi/pistediagrammi

- Kahden jatkuvan muuttujan välisen riippuvuuden graafinen esitystapa.
- Havaintoarvot esitetään koordinaatistossa.
  - pystyakselille selitettävä muuttuja
  - vaaka-akselille selittävä muuttuja
- "Pisteparven" muodosta tehdään johtopäätökset.
  - Jos pisteparven ympäri piirretty viiva muodostaa suunnilleen ympyrän, niin muuttujat ovat riippumattomia.
  - Mitä soikeampi pisteparven ympäri piirretyn viivan muodostama kuvio on, niin sitä enemmän muuttujien välillä on riippuvuutta.
  - Muuttujien skaalaus tulee tietysti ottaa huomioon.

Laatuero- ja järjestysasteikollisille muuttujille ei ole täysin soveliasta riippuvuuskuvaajaa.

Esimerkki

- Aineistossa kahdeksan henkilöä, joilta on mitattu ikä ja verenpaine.
- Jokainen piste vastaa henkilön (ikä, verenpaine) -paria.

Henkilö	1	2	3	4	5	6	7	8
Ikä	25	27	33	36	42	45	56	60
Verenpaine/mmHg	115	110	126	136	128	143	159	147

- Selittävä muuttuja on mitä ilmeisimmin Ikä, joten sen arvot tulevat x-akselille.
- Selitettävä muuttuja on näin ollen Verenpaine, joten sen arvot tulevat y-akselille.

Lineaarinen riippuvuus

- Lineaarisen riippuvuuden kuvaamiseen hajontakuviossa käytetään suoraa.
- Korrelaatio on lineaarisen riippuvuuden mitta.
- Riippuvuus voi olla myös muunlaista kuin lineaarista.
  - esim. eksponentiaalista
- Jos muuttujat ovat riippumattomia, ne ovat myös korreloimattomia.
- Mutta vaikka muuttujat ovat korreloimattomia ($r \approx 0$), ne voivat silti riippua toisistaan jopa matemaattisesti:
  - Muuttujien välillä voi olla funktionaalinen riippuvuus:
    - Havainnot voivat olla ympyrän kehällä, muodostaa paraabelin tms.

4.3 Tunnuslukuja

- Välimatka- ja suhdeasteikko
  - Tulomomenttikorrelaatiokerroin
    - Pearsonin r
- Järjestysasteikko
  - Järjestyskorrelaatiokertoimet
    - Spearmanin rho
    - Kendallin tau
- Laatueroasteikko
  - Köyhän miehen korrelaatiokertoimet
    - Kontingenssikerroin
    - Cramerin V

4.3.1 χ^2-tunnusluku

- Ristiintaulukosta laskettu tunnusluku, joka kuvaa muuttujien riippuvuuden määrää.
- Lasketaan ristiintaulukosta odotetut frekvenssit (jos muuttujat olisivat riippumattomia) ja verrataan niitä havaittuihin frekvensseihin.
- Odotetut frekvenssit:

$$e_{ij} = \frac{f_{i*} \cdot f_{*j}}{n}$$

- χ^2-tunnusluku:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

  - k ja l ovat sarakkeiden ja rivien lukumäärät.

Esimerkki

	Syöpä	Ei syöpää	$\sum$
Tupakoi	37	2963	3000
Ei tupakoi	13	6987	7000
$\sum$	50	9950	10000

	Syöpä	Ei syöpää	$\sum$
Tupakoi	$e_{11} = \frac{3000 \cdot 50}{10000} = 15$	$e_{12} = \frac{3000 \cdot 9950}{10000} = 2985$	3000
Ei tupakoi	$e_{21} = \frac{7000 \cdot 50}{10000} = 35$	$e_{22} = \frac{7000 \cdot 9950}{10000} = 6965$	7000
$\sum$	50	9950	10000

Esimerkki

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$
$$= \frac{(37-15)^2}{15} + \frac{(2963-2985)^2}{2985} + \frac{(13-35)^2}{35} + \frac{(6987-6965)^2}{6965} \approx 46.32$$

- Mitä suurempi χ^2-tunnusluku on, sitä enemmän muuttujien välillä on riippuvuutta.
- χ^2-tunnusluvut riippuvat kuitenkin suuresti otoskoosta ja luokkien lukumäärästä, joten eri aineistoista saatuja lukuja ei voi suoraan vertailla keskenään!

χ^2-johdannaiset

- Kontingenssikerroin C
  - Lasketaan χ^2-testisuureen avulla seuraavan kaavan mukaisesti $C = \sqrt{\frac{\chi^2}{n+\chi^2}}$, jossa n on havaintojen lukumäärä.
  - Kontingenssikerroin huomioi otoskoon, muttei luokkien lukumäärää, joten sitä kannattaa käyttää vasta kun matriisi on kokoluokkaa $5 * 5$.
- Cramerin V
  - Lasketaan χ^2-testisuureen avulla seuraavan kaavan mukaisesti $V = \sqrt{\frac{\chi^2/n}{m-1}}$, jossa m on rivien ja sarakkeiden lukumääristä pienempi.
  - Huomioi sekä otoskoon että luokkien lukumäärän.
- Köyhän miehen korrelaatiokertoimia, koska eivät kerro riippuvuuden suuntaa, vain sen määrän, eli vaihtelevat nollan ja ykkösen välillä, toisin kuin aidot korrelaatiokertoimet, jotka vaihtelevat välillä -1 ja 1.

Esimerkki

- Esimerkissämme: $\chi^2 \approx 46.32$
  - Kontingenssikerroin
    - $C = \sqrt{\frac{46.32}{10000+46.32}} \approx 0.0679$
  - Cramerin V
    - $V = \sqrt{\frac{46.32/10000}{2-1}} \approx 0.0681$
- Mitä lähempänä yhtä arvot ovat sitä suurempi on riippuvuus muuttujien välillä ja vastaavasti arvo 0 tarkoittaa tilastollista riippumattomuutta.
- Huom. Sekä kontingenssikerroin että Kramerin V saattavat osoittaa suurilla epätasaisesti jakautuneilla aineistoilla heikkoa riippuvuutta, vaikka muuttujien välillä olisikin selkeä riippuvuus.

4.3.2 Järjestyskorrelaatioista

Järjestyskorrelaatiokertoimet

- Spearmanin rho (ρ)
- Kendallin tau (τ)
- Goodmanin ja Kruskalin gamma (γ)

Järjestyskorrelaatiokertoimien edellytyksiä ja ominaisuuksia

- Muuttujien pitää olla vähintään järjestysasteikollisia.
- Vertaavat muuttujien järjestysnumeroita toisiinsa.
- Järjestyskorrelaatiokertoimet ovat aitoja korrelaatiokertoimia, joten ne vaihtelevat välillä -1 ja 1.
  - -1 tarkoittaa täydellistä negatiivista riippuvuutta.
  - 0 osoittaa tilastollista riippumattomuutta.
  - 1 tarkoittaa täydellistä positiivista riippuvuutta.

Spearmanin rho (ρ)

Spearmanin järjestyskorrelaatiokerroin (ρ)

- Vertaa muuttujien arvojen järjestysnumeroita toisiinsa.
- $\rho = 1 - \left(\frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n^3 - n} \right)$, jossa d_i on järjestyslukujen erotus i:nnessä havaintoyksikössä.
- Spearmanin korrelaatiokertoimelle pätee: $-1 \leq \rho \leq 1$.
- Spearmanin ρ on järjestysluvuille laskettu Pearsonin r.
  - Jos aineistossa on paljon samoja tuloksia (tasapelejä), niin tällöin kannattaa laskea järjestyslukujen Pearsonin korrelaatiokerroin.

Esimerkki

Kaksi tuomaria antoi 10 kilpailijalle pisteitä, joskin eri skaalalla. Mikä on tuomarien pisteiden korrelaatio?

Kilpailija	1	2	3	4	5	6	7	8	9	10	$\sum$
Tuomari x	73	62	54	85	90	77	70	92	58	80	
Tuomari y	22	19	27	34	48	36	37	29	26	42	
$R(x)$	5	3	1	8	9	6	4	10	2	7	
$R(y)$	2	1	4	6	10	7	8	5	3	9	
d_i	3	2	-3	2	-1	-1	-4	5	-1	-2	
d_i^2	9	4	9	4	1	1	16	25	1	4	74

$$\rho = 1 - \left(\frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n^3 - n} \right) = 1 - \frac{6 \cdot 74}{10^3 - 10} \approx 0.55$$

Kendallin tau (τ)

Kendallin järjestyskorrelaatiokertoimet τ_a, τ_b ja τ_c

- Vertaillaan järjestetyn aineiston järjestysnumeroita toisiinsa.
- Tällä kurssilla käsittelemme perusteellisemmin vain τ_a'ta.
- $\tau_a = \frac{2 \cdot P}{(n^2 - n)/2} - 1 = \frac{S}{(n^2 - n)/2}$
  - P on oikeiden järjestysten lukumäärä järjestetyssä aineistoissa.
  - S on oikeiden ja väärien järjestysten erotus järjestetyssä aineistossa.
  - Tasapelien tapauksessa: P'tä laskettaessa lisätään 0.5 ja S'ää 0.
- Jos muuttujien sisällä on paljon tasapelejä, niin käytetään kaavaa:
  - $\tau_a = \frac{S}{\sqrt{([\binom{n}{2} - n_x][\binom{n}{2} - n_y])}}$
    - Huom. $\binom{n}{2} = (n^2 - n)/2 = \frac{n \cdot (n-1)}{2}$
- Kendallin korrelaatiokertoimille pätee: $-1 \leq \tau \leq 1$.

Esimerkki

Samat tuomarit ja samat pisteet, kuin aiemminkin, mutta nyt järjestettynä tuomarin x antamien pisteiden mukaiseen suuruusjärjestykseen.

Kilpailija	3	9	2	7	1	6	10	4	5	8	$\sum$
Tuomari x	54	58	62	70	73	77	80	85	90	92	
Tuomari y	27	26	19	37	22	36	42	34	48	29	
$R(x)$	1	2	3	4	5	6	7	8	9	10	
$R(y)$	4	3	1	8	2	7	9	6	10	5	
P	6	6	7	2	5	2	1	1	0		30

$$\tau_a = \frac{2 \cdot P}{(n^2 - n)/2} - 1 = \frac{2 \cdot 30}{(10^2 - 10)/2} - 1 = \frac{4 \cdot 30}{(10^2 - 10)} - 1 = \frac{120}{90} - 1 \approx 0.333$$

- P on siis x-muuttujan mukaan järjestyneiden y-muuttujan järjestyslukujen oikeiden järjestysten lukumäärä.

Muita järjetyskorrelaatiokertoimia

- Kendallin τ_b (huomioi tasapelit)
  - $\tau_b = \frac{P - Q}{\sqrt{P + Q + X_0} \cdot \sqrt{P + Q + Y_0}}$
- Kendallin ja Stuartin τ_c (suurille aineistoille)
  - $\tau_c = \frac{S \cdot 2 \cdot m}{n^2 (m - 1)}$
- Goodmanin ja Kruskalin γ (ei huomioi tasapelejä)
  - $\gamma = \frac{P - Q}{P + Q}$
- Merkkien selitykset
  - S on oikeiden ja väärien järjestysten erotus järjestetyssä aineistossa.
  - P on oikeiden järjestysten lukumäärä järjestetyssä aineistoissa.
  - Q on väärien järjestysten lukumäärä järjestetyssä aineistoissa.
  - X_0 on X-muuttujan suuntaisten tasapelien lukumäärä.
  - Y_0 on Y-muuttujan suuntaisten tasapelien lukumäärä.
  - m on rivien ja sarakkeiden lukumääristä pienempi.

4.3.3 Korrelaatiokerroin

- Kahden numeerisen muuttujan lineaarisen riippuvuuden mitta.
  - Muuttujat ovat siis vähintään välimatka-asteikollisia.
- Korrelaatiokertoimelle pätee: $-1 \leq r \leq 1$.
- Kun $r = 1$, on muuttujien välillä täydellinen positiivinen lineaarinen riippuvuus.
- Kun $r = -1$, on muuttujien välillä täydellinen negatiivinen lineaarinen riippuvuus.
- Kun $r \approx 0$, muuttujat ovat korreloimattomia.
- Korrelaatiokerroin on herkkä poikkeaville havainnoille.
  - Eli korrelaatiokerroin ei ole robusti.
- Korrelaatiokerroin on invariantti sekä siirron että skaalauksen suhteen.
  - Eli kun kaikki havainnot kerrotaan jollain vakiolla ja/tai niihin lisätään jokin vakio, niin korrelaatiokerroin ei muutu.

Korrelaatiokerroin ja kovarianssi

- Puhuttaessa pelkästä korrelaatiokertoimesta tarkoitetaan Pearsonin tulomomenttikorrelaatiokerrointa r.
- Korrelaatiokerroin on kahden muuttujan välinen kovarianssi jaettuna muuttujien varianssien tulon neliöjuurella.
- Kovarianssi s_{xy}
  - Kahden muuttujan yhteisvaihtelua mittaava tunnusluku.
  - $s_{xy} = Cov(X, Y)$ on muuttujien X ja Y välinen kovarianssi.
    - $s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}$
  - Varianssi on yhden muuttujan sisäinen kovarianssi.
    - $Cov(X, X) = s_{xx} = s_x^2 = Var(X)$
- Yllä on luonnollisestikin laskettu otoskovarianssia, sillä perusjoukon kovarianssissa σ_{xy} jakajana olisi n, yllä käytetyn $(n-1)$'n sijaan.

Pearsonin korrelaatiokerroin (r)

- $r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{n}(y_i - \overline{y})^2}}$
  - s_x on x:n keskihajonta ja s_y on y:n keskihajonta.
- On kuitenkin helpompi käyttää johdettua kaavaa:
  - $r = \frac{n\sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{\left(n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2\right)\left(n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2\right)}}$

Vaihtoehtoinen lähestymistapa

- Korrelaatiokerroin voidaan laskea myös ns. z-pistemäärien avulla:
  - $z_{x_i} = \frac{x_i - \overline{x}}{s_x}$ ja $z_{y_i} = \frac{y_i - \overline{y}}{s_y}$
  - z-pistemäärien laskenta tarkoittaa havaintoarvojen standardointia.
- Tällöin korrelaatiokertoimen kaava on:
  - $r = \frac{\sum_{i=1}^{n}(z_{x_i} \cdot z_{y_i})}{n-1}$
- Kaavojen yhteys:
  - $r = \frac{s_{xy}}{s_x s_y} = \frac{\frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_x s_y (n-1)}$
    $= \frac{\sum_{i=1}^{n}\frac{(x_i - \overline{x})(y_i - \overline{y})}{s_x s_y}}{n-1} = \frac{\sum_{i=1}^{n}\frac{(x_i - \overline{x})}{s_x}\frac{(y_i - \overline{y})}{s_y}}{n-1} = \frac{\sum_{i=1}^{n}(z_{x_i} \cdot z_{y_i})}{n-1}$
- Tällainen korrelaatiokertoimen laskentatapa on teoreettisesti mielenkiintoinen, mutta laskennallisesti kovin työläs.

Esimerkki

Henkilö	Ikä (x_i)	Verenpaine (y_i)	x_i^2	y_i^2	$x_i \cdot y_i$
1	25	115	625	13225	2875
2	27	110	729	12100	2970
3	33	126	1089	15876	4158
4	36	136	1296	18496	4896
5	42	128	1764	16384	5376
6	45	143	2025	20449	6435
7	56	159	3136	25281	8904
8	60	147	3600	21609	8820
$\sum$	324	1064	14264	143420	44434

Esimerkki

Henkilö	Ikä (x_i)	Verenpaine (y_i)	x_i^2	y_i^2	$x_i \cdot y_i$
$\sum$	324	1064	14264	143420	44434

$$r = \frac{n\sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{\left(n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2\right)\left(n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2\right)}}$$

$$= \frac{8 \cdot 44434 - 324 \cdot 1064}{\sqrt{(8 \cdot 14264 - 324^2)(8 \cdot 143420 - 1064^2)}}$$

$$= \frac{355472 - 344736}{\sqrt{9136 \cdot 15264}} \approx 0.91$$

Selitysaste (r^2)

- Selitysaste mittaa muuttujan selitysvoimaa.
- Yritetään selittää muuttujan y vaihtelua muuttujalla x.
  - Jos $r = 0.5$, niin $r^2 = 0.25$, eli x selittää 25% y:n vaihtelusta.
- Jäljelle jäävä vaihtelu selittyy joko muilla selittäjillä tai sattumalla.
- Kun $r = 0$, myös $r^2 = 0$.
  - y'n kanssa korreloimaton x ei selitä y'n vaihtelusta mitään.
- Kun $r = 1$ tai $r = -1$, niin $r^2 = 1$
  - Muuttuja x selittää täysin y'n vaihtelun.
- Korrelaatiokertoimen etumerkillä ei ole väliä selitysasteen kannalta.
  - Jos $r = 0.7$, niin $r^2 = 0.49$ ja jos $r = -0.7$, niin $r^2 = 0.49$.
- Ikä-verenpaine –esimerkki:
  - $r \approx 0.91$, joten selitysaste on noin 0.83.

Riippuvuustunnusluvut mitta-asteikoittain

-

Laatueroasteikko	Järjestysasteikko
χ^2, C, V	C, V, ρ, τ_a , τ_b, τ_c, γ

-

Välimatka-asteikko	Suhdeasteikko
ρ, τ, r	ρ, τ, r

-

5. Lineaarinen regressiomalli

5.1 Lineaarinen riippuvuus

- Yhden muuttujan riippuvuutta muista muuttujista voidaan analysoida lineaarisella regressiomallilla.
  - Yhden selittäjän mallissa kuvaajana on suora.
  - Kahden selittäjän mallissa kuvaajana on taso.
  - Useamman selittäjän mallissa kuvaajana hypertaso.
- Selitettävän muuttujan (y) arvoja pyritään ennustamaan selittävän/selittävien muuttujan (x_i) arvoilla.
- Mallin muuttujat ovat vähintään välimatka-asteikollisia.
- Tällä kurssilla keskitytään yhden selittäjän malleihin, koska usean selittäjän mallit ovat laskennallisesti vaativia.
  - Yhden selittäjän mallit ovat kuitenkaan harvoin riittäviä.

5.2 Regressiomalli

- Regressiomallissa määritetään regressiokertoimet pienimmän neliösumman menetelmällä (PNS-menetelmä).
  - PNS-menetelmässä minimoidaan ennusteen $\hat{y}$ ja selitettävän muuttujan y erotusten neliötä.
  - Havaitun arvon ja ennusteen erotusta kutsutaan residuaaliksi: $\epsilon = y - \hat{y}$
- Regressisuoran yleinen yhtälö on:
  - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$
- Yhden selittäjän mallissa yhtälö on yleensä muodossa:
  - $y = \alpha + \beta x + \varepsilon$
    - α on mallin vakio.
    - β on regressiokerroin.
    - ε on virhetermi, jota ei aina merkitä yhtälöön.

Suora

- Suora määritellään yhtälöllä $y = \alpha + \beta \cdot x$.
  - α on y-akselin leikkauspiste.
    - eli x'n arvolla 0 y saa arvon α.
  - β on suoran kulmakerroin.
    - eli β kertoo kuinka paljon y muuttuu x'n muuttuessa yhden yksikön verran.
- Suoran piirtäminen koordinaatistoon:
  - Sijoitetaan kaksi x'n arvoa suoran yhtälöön, jolloin saadaan yhtälöstä laskettua kaksi y'n arvoa.
  - Saadut pisteet (lukuparit) merkitään koordinaatistoon ja piirretään niiden kautta kulkeva suora.
  - x -muuttujan minimi ja maksimi ovat yleensä toimivia valintoja.

PNS-suora

- PNS-suora määritellään yhtälöllä $y = a + bx$, jossa
- $a = \overline{y} - b\overline{x}$, jossa
  - $\overline{y}$ ja $\overline{x}$ ovat y'n ja x'n keskiarvot ja
- $b = r\frac{s_y}{s_x}$, jossa
  - s_y ja s_x ovat y'n ja x'n keskihajonnat ja
  - r on x'n ja y'n välinen korrelaatiokerroin.
- Luonnollisesti kulmakerroin b lasketaan ensin ja sen avulla saadaan laskettua vakio a.

Esimerkki (jatkoa)

Ikä ($\overline{x}$)	Verenpaine ($\overline{y}$)	s_x	s_y	r_{xy}
40.5	133	12.773	16.510	0.909

- $b = r\frac{s_y}{s_x}$
  - $b = 0.909 \cdot \frac{16.510}{12.773} \approx 1.175$
- $a = \overline{y} - b\overline{x}$
  - $a = 133 - 1.175 \cdot 40.5 \approx 85.41$
- $y = a + bx$
  - $y = 85.41 + 1.175 \cdot x$
- Jos ikä on 50 vuotta, niin verenpaine-ennuste on:
  - $\hat{y} = 85.41 + 1.175 \cdot 50 \approx 144.2$

PNS-suoria on kaksi

- Y'n PNS-suora X'n suhteen:
  - $y - \overline{y} = r\frac{s_y}{s_x}(x - \overline{x})$
- X'n PNS-suora Y'n suhteen:
  - $x - \overline{x} = r\frac{s_x}{s_y}(y - \overline{y})$
- Nämä suorat leikkaavat havaintoaineiston painopisteessä $(\overline{x}, \overline{y})$.
- Jos $r > 0$, molemmat suorat ovat nousevia.
- Jos $r < 0$, molemmat suorat ovat laskevia.
- Kun $r = 0$, niin muuttujat ovat lineaarisesti riippumattomia.
  - Kun x'llä selitetään y'tä, niin suora on vaakasuorassa.
  - Kun y'llä selitetään x'ää, niin suora on pystysuorassa.
- Suorat ovat samat vain, jos $r = 1$ tai $r = -1$.

Lineaarisen regressiomallin rajoitteita

- Muuttujien on oltava vähintään välimatka-asteikollisia.
- Ennuste pätee vain muuttujien vaihteluvälillä tai sen lähellä.
- Poikkeavat havainnot saattavat vääristää tuloksia.
- Riippuvuuden tulee olla lineaarista, eli suoraviivaista.
  - Jos muuttujien välinen riippuvuus ei ole lineaarista, sitä voi yrittää linearisoida esimerkiksi logaritmoimalla.
- Yhtälössä on yleensä useita selittäjiä ja virhetermi ε.
  - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$
  - Virhetermi (ts. residuaalit) oletetaan normaalijakautuneeksi odotusarvolla 0 ja vakiovarianssilla, eli $\varepsilon \sim N(0, \sigma^2)$.
    - Yksi yleinen ongelma on mallin heteroskedastisuus, jolloin residuaalien vaihtelu PNS-suoran ympärillä ei ole tasaista.

Lineaarisen regressiomallin tulkinta

- Tavoitteena on yleensä löytää muutama vaihtoehtoinen malli, josta sovellusalan tutkija voi valita, sen jolle löytyy järkevin tulkinta.
- Hyvän mallin ominaisuuksia:
  - Mallin selitysaste on korkea.
  - Kaikki malliin jätetyt selittäjät ovat merkitseviä.
  - Mallissa on järkevä määrä selittäjiä.
  - Malli on harhaton.
- Malli kannattaa pitää mahdollisimman yksinkertaisena:
  - Jos selittäjä ei kasvata selitysastetta se kannattaa poistaa mallista.
  - Jos selittäjän muunnos ei anna merkittävää lisäarvoa mallille/tulkinnalle, kannattaa muunnos yleensä jättää tekemättä.
- Vakio kannattaa säilyttää mallissa, jollei ole erityistä syytä poistaa sitä.
- Jos kyseisen sovellusalan teoria/käytäntö puoltaa muuttujamuunnosta tai jonkun tietyn selittäjän mukana oloa, kannattaa testata myös tällainen malli, vaikkei se valikoituisikaan tilastollisesti parhaaksi.

Standardoitu lineaarinen regressiomalli

- Kun standardoiduilla muuttujilla lasketaan PNS-suora/hypertaso, niin saadaan standardoidut β-kertoimet.
  - $z = \beta_1 \cdot z_1 + ... + \beta_k \cdot z_k + \varepsilon$
  - Muunnetun mallin vakio on 0, eli PNS-"suora" kulkee origon kautta.
- Eräänä hyötynä on, että nyt β-kertoimia voi verrata suoraan toisiinsa, eli mitä korkeampi kertoimen itseisarvo on, niin sitä merkitsevämpi se on myös selittäjänä.
- Eräänä haittana on se, että ennusteet ovat myös standardoituja, eli ne pitää muuntaa takaisin, jotta saadaan aidot ennusteet.
  - $\hat{y_j} = \hat{z}_j \cdot s_y + \overline{y}$
- Yhden selittäjän mallissa β-kerroin on tällöin muuttujien välinen korrelaatiokerroin r_{xy} ja vakio α on edelleenkin 0.
  - $z_y = \beta \cdot z_x + \varepsilon <=> z_y = r_{xy} \cdot z_x + \varepsilon$

Standardointi lineaarisessa regressiossa

- Ns. standardoidussa lineaarisessa regressiomallissa sovitetaan PNS-suora aineistoon, jonka muuttujat on standardoitu.
  - Kaikista muuttujista vähennetään keskiarvo ja jaetaan keskihajonnalla.
    - $z_{ij} = \frac{x_{ij} - \overline{x_i}}{s_{x_i}}$ sekä $z_j = \frac{y_j - \overline{y}}{s_y}$
  - Muunnos tehdään siis sekä selitettävälle että selittäville muuttujille.
- Muunnettujen muuttujien keskiarvo on tällöin 0 ja varianssi 1.
- Standardoitujen muuttujien tapauksessa kovarianssi=korrelaatiokerroin.
  - $r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{Cov(X,Y)}{1 \cdot 1} = Cov(X,Y)$
- Sekä standardoitujen että standardoimattomien muuttujien välinen korelaatiokerroin on sama, mutta niiden kovarianssit poikkeavat.
- Standardoinnin suurin hyöty saavutetaan usean selittäjän mallien laskennassa, kun tarvittavat matriisit yksinkertaistuvat.

5.3 Lineaaristen mallien yleistykset

- Yleistetyt lineaariset mallit $\eta = \mathbf{XB} + \mathbf{U}$
  - Yleiset/klassiset lineaariset mallit $\mathbf{Y} = \mathbf{XB} + \mathbf{U}$
    - mm. regressioanalyysi, varianssianalyysi, t-testit, F-testi
  - Logistinen regressiomalli $\mathbf{log(Y)} = \mathbf{XB} + \mathbf{U}$
    - Käytetään silloin, kun selitettävä muuttuja on binäärinen.
  - Log-lineaariset mallit
    - Käytetään luokitellun aineiston frekvenssien mallintamiseen.
  - Muillekin ns. eksponenttiperheen jakaumille voidaan muodostaa vastaavat mallit.
- Valitettavaa on, että sekä Yleistetyistä lineaarisista malleista (Generalized linear models) että Yleisestä lineaarisesta mallista (General linear model) käytetään lyhemmystä GLM.

5.* Usean muuttujan tutkailua

- Multippelikorrelaatio $(R_{y,x_1 \ldots x_k})$
  - Usean muuttujan välinen korrelaatiokerroin
  - Lineaarisen mallin selitysasteen neliöjuuri (tai oikeastaan päinvastoin).
  - $R_{y,x_1 \ldots x_k} = \sqrt{R^2}$
- Osittaiskorrelaatio $(r_{xy.z})$
  - Muuttujan z vaikutuksesta puhdistettu korrelaatio.
  - $r_{xy.z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}}$
- Kendallin konkordanssi (W)
  - Usean muuttujan järjestyskorrelaatio
  - $W = \frac{12 \sum_{i=1}^{n} R_i^2 - 3k^2 n(n+1)^2}{k^2 n(n^2-1)} = \frac{\overline{\rho}(k-1)+1}{k}$
  - $R_i =$ järjestyslukujen summa, $n =$ havaintojen lukumäärä, $k =$ muuttujien lukumäärä, $\overline{\rho} =$ järjestyskorrelaatioiden keskiarvo

Monimuuttujamenetelmät

- Monimuuttujamenetelmät ovat nimensä mukaisesti menetelmiä, joissa tutkitaan useiden muuttujien välisiä riippuvuusrakenteita.
- Klassisia lineaarisia malleja ei yleensä sisällytetä monimuuttuja-menetelmiin, vaikka niissä tavallisesti on selittäjinä useita muuttujia.
- Ylivoimaisesti tunnetuin monimuuttujamenetelmä on faktorianalyysi.
- Faktorianalyysi
  - Psykometriassa ja yhteiskuntatieteissä paljon käytetty menetelmä.
  - Isosta joukosta muuttujia pyritään löytämään latentteja taustamuuttujia, eli faktoreita.
    - Rotatoinilla etsitään parasta mahdollista faktorikombinaatiota.
    - Saaduille faktoreille pyritään löytämään mielekkäät tulkinnat.
  - Faktorianalyysi voi olla joko eksploratiivista tai konfirmatorista.
    - Eksploratiivisessa etsitään uutta tuntematonta faktorirakennetta.
    - Konfirmatorisessa varmennetaan teoreettista faktorirakennetta.
  - Muuttujat ovat usein järjestysasteikollisia (esim. likert-asteikko)

Muita monimuuttujamenetelmiä

- Pääkomponenttianalyysi
  - Pyritään tiivistämään suuren muuttujajoukon informaatio muutamiin johdettuihin muuttujiin, eli pääkomponentteihin.
- Ryhmittelyanalyysi
  - Tavoitteena on löytää perusjoukosta mahdollisia ryhmityksiä.
- Erotteluanalyysi
  - Tavoitteena on löytää etukäteen tiedetyt luokitukset aineistosta.
- Korrespondenssianalyysi
  - Etsitään luokitellusta aineistosta samankaltaisia muuttujia.
- Moniulotteinen skaalaus
  - Havainnot pyritään sijoittamaan 2- tai 3-ulotteiselle kartalle.
- Rakenneyhtälömallit
  - Konfirmatorisen faktorianalyysin ja regressioanalyysin yhdistelmä.

Aikasarja-analyysi

- Tutkii muutoksia ajassa, joten aika on yksi muuttujista.
- Ekonometriassa käytetyin tilastollinen menetelmä.
- Vaihtelu hajotetaan kahteen komponenttiin:
  - deterministiseen ja satunnaiseen.
- Deterministisestä komponentistä pyritään erottelemaan lisäksi:
  - Trendi, joko lineaarinen tai epälineaarinen nousu tai lasku
  - Kausivaihtelu, vuoden sisäinen säännöllinen vaihtelu
  - Suhdannevaihtelu, vuosien välinen vaihtelu
- Keskeisimmät mallit ovat:
  - autoregressiivinen malli (AR)
  - liikkuvan/liukuvan keskiarvon malli (MA)
  - sekä näiden johdokset (ARMA ja ARIMA)
- Useista selittävistä muuttujista huolimatta aikasarja-analyysiäkään ei yleensä sisällytetä monimuuttujamenetelmiin.

6. Normaalijakauma

Normaalijakauma

- Tärkein ja käytetyin jatkuva jakauma tilastotieteessä.
- Monet luonnonilmiöt noudattavat normaalijakaumaa.
- Esimerkkejä normaalijakautuneista satunnaismuuttujista:
  - ihmisestä mitatut ominaisuudet
  - koneiden ja laitteiden käyttöikä
  - lämpötilat ja sademäärät
- Normaalijakautuneessa ilmiöissä keskimääräisiä havaintoja on enemmän kuin äärimmäisiä havaintoja.
- Muita yleisesti käytettyjä nimityksiä ovat Gaussin jakauma ja kellokäyrä.
- Ns. normaalisuusoletus tarkoittaa, että muuttuja noudattaa normaalijakaumaa joillain parametereillä.

Normaalijakauman parametrit

- Normaalijakauman muodon määrittävät kaksi parametria:
  - Odotusarvo: μ määrää jakauman keskikohdan eli paikan.
  - Varianssi: σ^2 määrää jakauman muodon.
- Kun satunnaismuuttuja X noudattaa normaalijakaumaa parametrein μ ja σ^2 eli $X \sim N(\mu, \sigma^2)$, niin X:n tiheysfunktio on muotoa $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$.
- Satunnaismuuttuja Z noudattaa standardoitua normaalijakaumaa, kun $\mu = 0$ ja $\sigma^2 = 1$, eli $Z \sim N(0, 1)$.
  - Tällöin X:n tiheysfunktio on muotoa $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$.

Normaalijakauman ominaisuuksia

- Normaalijakauma on yksihuippuinen ja symmetrinen.
- Normaalijakauman ja x-akselin rajaama pinta-ala on yksi.
- Normaalijakauma on keskeisin jatkuva todennäköisyysjakauma.
- Jatkuvien todennäköisyysjakaumien yleisiä ominaisuuksia:
  - Pinta-ala määrittää todennäköisyyden.
  - Kaikki pistetodennäköisyydet ovat nollia.
  - Kokonaispinta-ala (eli kokonaistodennäköisyys) on yksi.
- Taulukoista nähdään yleensä kertymäfunktion arvot, eli siihen pisteeseen mennessä kertynyt todennäköisyys.
  - Toisinaan taulukoidaan myös ns. häntätodennäköisyyksiä.

Normaalijakauman standardointi

- $X \sim N(\mu, \sigma^2)$, jossa μ on odotusarvo ja σ^2 on varianssi.
- Kun normaalijakautunut muuttuja X standardoidaan, niin siitä vähennetään odotusarvo μ ja se jaetaan keskihajonnalla σ.
- Saatu uusi muuttuja Z noudattaa standardoitua normaalijakaumaa, eli $Z \sim N(0, 1)$, jossa $z = \frac{x-\mu}{\sigma}$
- Esimerkki: $X \sim N(2, 4)$.
  - Mikä on todennäköisyys, että x on pienempi kuin 3?
  - Standardoidaan: $z = \frac{3-2}{\sqrt{4}} = 1/2$, joten $\phi(0.5) = 0.6914$, eli todennäköisyys on noin 69%
- Huom. Normaalijakauman taulukoissa on siis standardoidun normaalijakauman $Z \sim N(0, 1)$ kertymäfunktion arvoja.

Normaalijakauman taulukon käyttö

- $P(Z < z) = \phi(z)$, eli katsotaan taulukosta z'aa vastaava todennäköisyys.
- $P(Z < -z) = 1 - \phi(z)$, eli katsotaan taulukosta z'aa vastaava todennäköisyys ja vähennetään se yhdestä.
- $P(Z > z) = 1 - \phi(z)$, eli katsotaan taulukosta z'aa vastaava todennäköisyys ja vähennetään se yhdestä.
- $P(Z > -z) = 1 - (1 - \phi(z)) = \phi(z)$, eli katsotaan taulukosta z'aa vastaava todennäköisyys.
- Esim. $z = 1.22$
  - $P(Z < 1.22) = \phi(1.22) = 0.8888$
  - $P(Z < -1.22) = 1 - \phi(1.22) = 1 - 0.8888 = 0.1112$
  - $P(Z > 1.22) = 1 - \phi(1.22) = 1 - 0.8888 = 0.1112$
  - $P(Z > -1.22) = 1 - (1 - \phi(1.22)) = 1 - 1 + 0.8888 = 0.8888$

68-95-99.7 -sääntö

- Normaalijakautuneille aineistoille pätee:
  - 68% havainnoista on yhden hajonnan mitan päässä keskiarvosta.
    - eli 68% havainnoista on välillä $(\mu - \sigma\ ,\ \mu + \sigma)$.
  - 95% havainnoista on kahden hajonnan päässä keskiarvosta.
    - eli 95% havainnoista on välillä $(\mu - 2\sigma\ ,\ \mu + 2\sigma)$.
  - 99.7% havainnoista on kolmen hajonnan päässä keskiarvosta.
    - eli 99.7% havainnoista on välillä $(\mu - 3\sigma\ ,\ \mu + 3\sigma)$.
- Tämä sääntö perustuu seuraaviin normaalijakauman taulukkoarvoihin:
  - 95% havainnoista on välillä, jonka alaraja on 0.025 ja yläraja 0.975.
  - Taulukosta katsottuna $0.975 = \phi(1.960)$, joka on noin kaksi.
  - Vastaavasti $0.840 = \phi(0.994)$ ja $0.9985 = \phi(2.968)$, jotka ovat lähellä yhtä ja kolmea.
- Tähän sääntöön perustuu se, että jotkut käyttävät testeissä tai luottamusvälejä laskiessaan arvoa 2 oikeamman arvon 1.960 sijasta.

Normaalijakauman keskeisiä ominaisuuksia

- Normaalijakauma on otoskeskiarvojen jakauma, eli kun otetaan samasta perusjoukosta useita otoksia, näiden otosten keskiarvot noudattavat normaalijakaumaa.
- Muuttujat, joiden arvot riippuvat useista vaikuttavista tekijöistä noudattavat usein normaalijakaumaa.
  - Tähän perustuen satunnaisvirheet ovat normaalijakautuneita.
- Keskeisen raja-arvolauseen perusteella riippumattomien samoin jakautuneiden muuttujien summat ovat approksimatiivisesti normaalijakautuneita.
  - Keskeiseen raja-arvolauseen perusteella voidaankin binomi- ja poissonjakautuneita muuttujia approksimoida normaalijakaumalla, kunhan otoskoot ovat riittävän suuria.