Kvantitatiivisten menetelmien käytön perusteet tieteellisessä tutkimuksessa

Auvo Rauhala, LL, FT, dos.

Päivitetty: 14.10.2025

SPTY, Asiakas- ja potilasturvallisuuskeskus, Åbo Akademi

Tästä oppaasta on aina tuorein päivitetty versio verkkosivulla:
https://www.spty.fi/tutkijoille/metodologian-linkkeja/

Sisällysluettelo (linkkeinä)

1. Johdantoa

Yleistä teorialuennoista ja opasdokumentista

- **Tässä oppaassa** esitetään tärkeimmät **perusasiat** ja yleisimmät **käytännön ongelmat** ”**keittokirjamaisesti**”, mutta menetelmien ”**idea**” kerrotaan (helpottaa ymmärtämistä)
- Monet menetelmät eivät ole tähän mahtuneet.
- Opas on kuitenkin pyritty laatimaan myös **itseopiskeluun** sopivaksi ja **käsikirjamaiseksi hakuoppaaksi** ja se sisältää myös esitettäviä asioita syvällisemmin käsitteleviä **linkkejä**
- Lyhyt esittely myös **nettikalkylaattoreista**, joita saattaa tarvita lisäapuna.
- Statistiikassa on **näkemyseroja menettelytavoista** etenkin jos, muuttujat ja jakaumat eivät ole ideaalisia ja aineisto on pieni – ja menetelmän ehdot täyttyvät vain osittain
- **Ongelmissa googlaus englanniksi mahdollisimman eksaktilla tekstillä tuo netistä avun** miltei aina – joskus tosin vasta kymmenes kokeilemasi jippo hoitaa pois ongelmasi!
- Ellei kuvien lähdettä ole mainittu, ne ovat joko itse tuotettuja tai Wikipediasta
- Jos huomaat tekstissäni **puutteita, virheitä** ym. muokkaustarpeita, **ilmoita minulle** niistä! Esim. auvo.rauhala(at)gmail.com

Tiivis kooste kvantit. aineiston käsittelyn eri vaiheista

Aineisto & keruu

- **Otoskoon** laskenta
- **Kaavake**suunnittelu

Aineiston tarkistus ja muokkaus

- **Muuttujat**=sarakkeet: skaala (esim. Likert=>intervalli tai ordinaali?), normaalijakautunut?
- ”**Cases**”=tilastoyksiköt=rivit
- **Havainnot**=solut: puuttuvat (➔imputoinnit?), virheelliset, poikkeavat, useampi samassa solussa
- **Tiedostorakenteen muokkaus**: muoto long <> wide, case vs. control, monitaso (esim. yksikkö-potilas-vauriokohdat)
- **Muuttujien muokkaus**: erilaisia summamuuttujia, muokattuja luokkia
- **Excelin funktioilla** huomattavaa ajansäästöä testauks. ja muokk:ssa!
- SPSS:ssä, jamovissa ym. myös muokk.

Analyysin valinta ja suoritus lisätesteineen

- Pelkkä **deskriptiivinen**
- **Eron testaus**
  - Luokkamuutt. (ristiintaul): Chi Square, Fisher
  - Intervalli, ei-parill.
    - normaalijak: t-testi (2 muutt), ANOVA (≥3)+post-hoc
    - ei-norm, (tai ordin.): M-W U-testi (2), Kruskal-W(≥3)
    - Parill. omat t-testit ja M-W U-testin sijaan Wilcoxon
- **Yhteisvaihtelu** 2 muuttujalla: korrelaatio (mikä monista?, esim. Pearson, Spearman, polychoric, point-biserial)
- **Output/response-muuttujan selittäminen** muilla (explanatory variables): univar/multivar regressioanalyysit,
  - tavallisimmat lineaarinen ja (bin./multinom.) logistinen
  - Analyysi ”selittää” vain matemaattisesti, todellisuudessa havainnoivassa tutkimuksessa ”assosioituu” tms
- **Eksploratiivinen faktorianalyysi (EFA)**: Etsitään piileviä latentt. faktoreita, jotka selittävät esim. mittarin us. itemiä
- **Muita analyysejä**
  - Luottamusvälit, efektikoot, valid. Ja reliab. testaukset
  - Rinnakkaisluokitusten yhdenmukaisuus (kappa ym)
  - Mittarin ennustekyky: ROC, Brier-score, goodness of fit
- **Analyysien ehtojen testaus**
  - Nämä osin analyysikohtaisia, osin yleisempiä: esim. otoskoko, muuttujien skaala, normaalisuus

Raportointi

- Deskr. taulukko
- Korr. matriisi
- Monimuutt. taul.
- Grafiikka
- Tekstiosa
- Kausaliteettia RCT:n ja kvasi-eksperimentaal. Aineistojen ulko-puolella vain harvoin!

Kvantitatiivisten menetelmien käytön vaiheet

1. Tavoitteiden teoreettisten **käsitteiden operationalisointi** mitattaviksi ominaisuuksiksi
2. **Muuttujien valinta**: **selitettävät** (output, response, dependent), **selittävät** (explanatory, independent), **muut** huomioitavat (esim. monimuuttuja-analyyseissä ”kontrolloidaan”
3. **Muuttujien mitta-asteikot** (nominaali-ordinaali-lukumäärä-intervalli…) ja luokkamuuttujien **luokat**
4. Tutkimus**asetelma**
5. **Otos** (tyyppi, esim. satunnaisotos ja otoskoko)
6. Tietojen **keruu** (esim. kyselykaavake)
7. Aineiston **tarkistus ja muokkaus**
8. Aineiston **analyysi**
9. Aineiston **raportointi**

❖ Aineisto ja metodologia kannattaa **suunnitella hyvissä ajoin** mahdollisimman pitkälle, ettei myöh. törmää ongelmiin, joille ei voi enää mitään.

❖ Myös **systemaattinen** välivaiheiden **tiedostodokumenttien nimeäminen ja tallentaminen** ja analyysipöytäkirjan ym. **apudokumenttien** tallennus tärkeää!

Muuttujien tyypit/mitta-asteikot

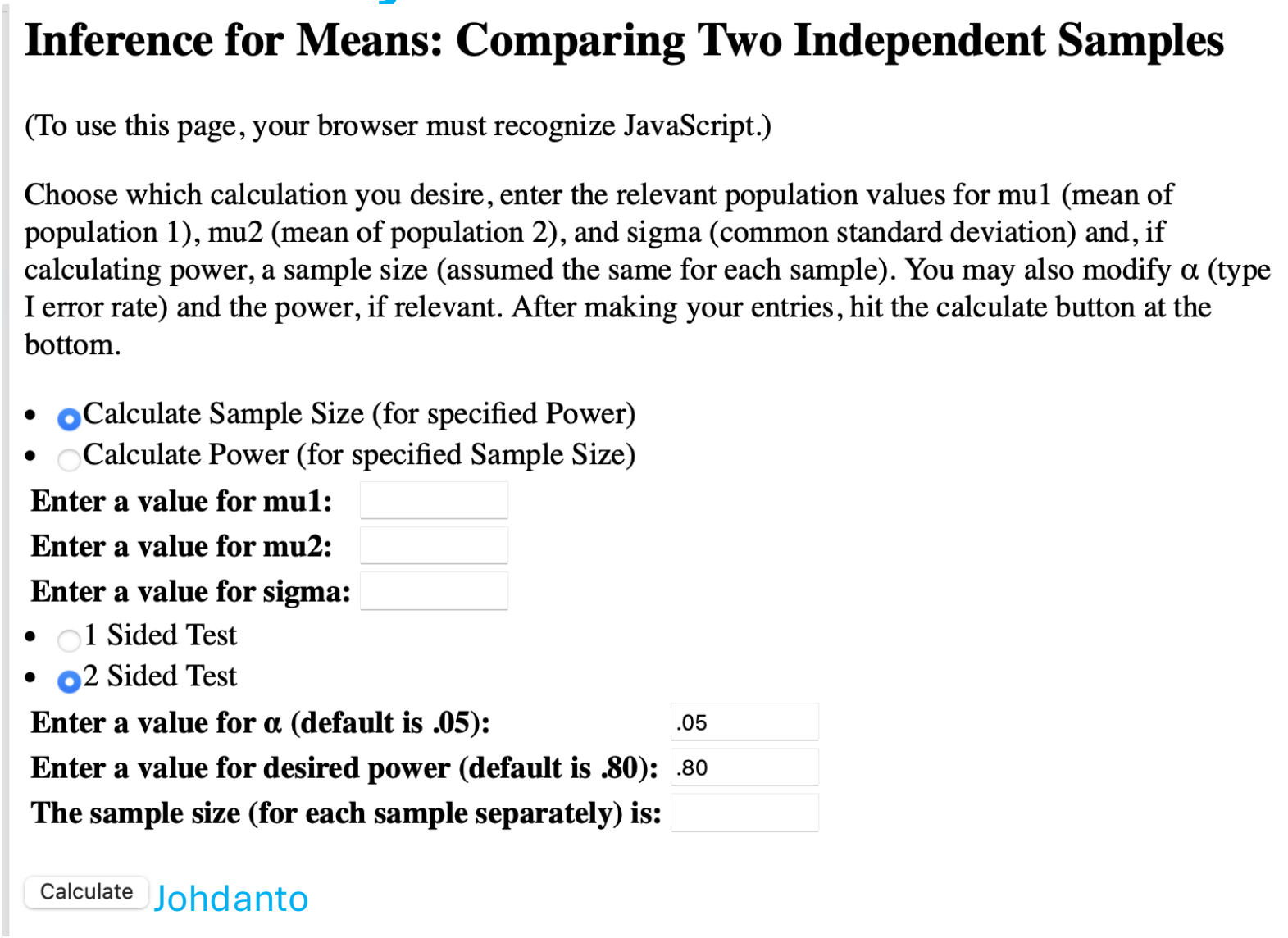
- **Mitta-asteikot/skaalat**
    1. **Nominaali-** eli **luokka**-asteikko (sukup., siv.sääty), luokissa ei mielekästä järjestystä. Erityistapaus **dikotominen** (2 luokkaa)
    2. **Ordinaali-** eli järjestys (Likert, PAPA, riski-Lk)-luokissa järjestys!
    3. **Intervalli-** eli välimatka, ilman absol. nollapistettä (Celsius)
    4. **Suhdeasteikko**, jossa absoluuttinen nollapiste (paino, Kelvin)
- **Jatkuvat** (tarkka ikä), **epäjatkuvat** eli diskreetit (esim. lukumäärät)
- Aineistoihin liittyy paljon skaalaongelmia ja rajatapauksia
- Käytetään myös nimityksiä laadulliset (Lk1) ja määrälliset(Lk3-4) muuttujat sekä ei-numeeriset (Lk1) ja numeeriset (Lk3-4) muuttujat. Järjestysasteikollisten kohdalla näissä hieman horjuvuutta eri lähteissä
- Määrällinenkin muuttuja voi ilmaista laatua (esim. sukan reikien määrä) ja laadullisen luokat voidaan koodata numeerisesti

Erityistapauksia/ongelmia mitta-asteikoissa

- **Likert**: Yksittäisten muuttujien ordinaaliset Likert/Likert-type asteikot: jos ≥5 (-7) numeerista luokkaa (ja luokiteltu ilmiö luonteeltaan jatkuva), voidaan käsitellä myös jatkuvana **intervallimuuttujana** (mutta näkemykset vaihtelevia!)
  Kts esim. https://davidfoxcroft.github.io/lsj-book/ *(Kirjoita Etsi-ruutuun: Likert)*
- **Dikotomisissa** kannattaa käyttää **”dummy”-koodeja, 0** *(=referenssitaso, johon verrataan)* ja **1** *(=tutkittava/verrattava taso)* etenkin, jos käytetään korrelaatio- tai regressioanalyysiä
- **Luokka-asteikkojen tyyppivirhe**: pidetään numerokoodattu nominaal. asteikko, jossa ≥3 luokkaa, **esim. siviilisääty korrel.- tai regressioanalyysissä yhtenä muuttujana**, jolloin esim. 1naimaton-2naimisissa-3eronnut-4leski-luokkien analyyseissä oletuksena olisi, että *”yksi leski vastaisi 2 naimisissa olevaa!!!”*
- **Intervallimuuttujan katkaisu luokkiin** tehdään **tilastollisessa päättelyssä** vain poikkeuksellisesti ja ilman perusteluja sitä voidaan jopa epäillä p-hacking-manipulaatioksi eli signif. p-arvojen metsästämiseksi! Siinä myös menetetään paljon analyysivoimaa. **Deskriptiivisesti** voidaan tietenkin esittää **luokkinakin**!

Otoskoon laskenta I/II - tyyppiesimerkki t-testin otoskoon määrittely nettilaskurilla

- Usein suositeltu laskuri: https://www.stat.ubc.ca/~rollin/stats/ssize/n2.html
- Otoskoon laskennan pitäisi perustua pää-tuloksen vaatimaan otokseen ja siinä käytettävään testiin
- Käytännössä t-testi usein pohjana, kuten esimerkissämmekin



Inference for Means: Comparing Two Independent Samples

(To use this page, your browser must recognize JavaScript.)

Choose which calculation you desire, enter the relevant population values for mu1 (mean of population 1), mu2 (mean of population 2), and sigma (common standard deviation) and, if calculating power, a sample size (assumed the same for each sample). You may also modify α (type I error rate) and the power, if relevant. After making your entries, hit the calculate button at the bottom.

- (•) Calculate Sample Size (for specified Power)
- ( ) Calculate Power (for specified Sample Size)

Enter a value for mu1:
Enter a value for mu2:
Enter a value for sigma:

- ( ) 1 Sided Test
- (•) 2 Sided Test

Enter a value for α (default is .05): .05
Enter a value for desired power (default is .80): .80
The sample size (for each sample separately) is:

Calculate

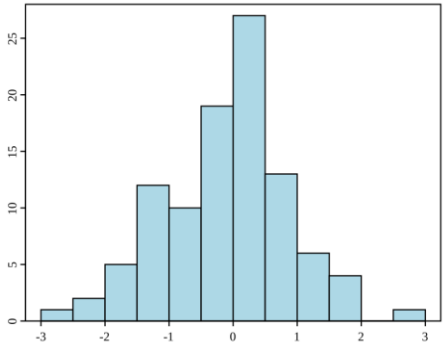
Otoskoon laskenta II/II – ohje edellisen sliden nettilaskurin käytölle

1. Käytä **desimaaliPISTETTÄ**!
2. Valitse analyysiksi **sample size (**tai **power**)
3. Merkitse verrattavien 2 ryhmän **keskiarvot** kohtiin **mu1** ja **mu2**. Huomaa, että keskiarvoilla ei sinänsä ole merkitystä, ohjelma käyttää vain niiden välisen ERON arvoa!
4. **Sigma** on **SD**, Jos ryhmillä on eri SD, käytä niiden keskiarvoa
5. Nuo **p-arvo** ja **power** ovat esiasetettu tavanomaisiin, mutta kannattaa kokeilla myös $p<0.01$:llä
6. Testit tietenkin pääsääntöisesti **2-sided**
7. Paina **CALCULATE**, niin saat sample sizet molemmille ryhmille.
8. Jos haluat t-testissä tietyn **(Cohenin d) efektikoon**, niin laskukaava on: **Efektikoko=keskiarvojen ero / SD** ja tavanomainen moderaatti efektikoko, on 0,5.

2. Aineiston tarkistus ja muokkaus

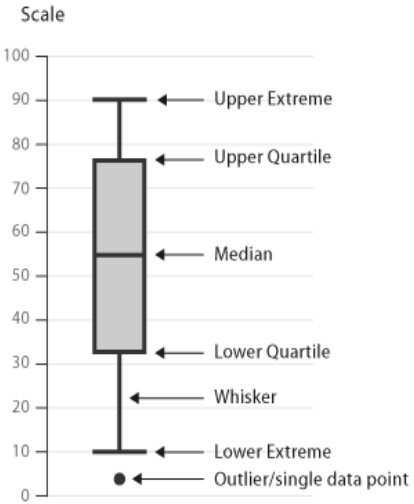
Aineiston kanssa työskentelyyn kuluvasta ajasta aineiston tarkistukset ja muokkaukset vievät jopa 90% ja itse analyysit vain ≥10%

Yleisesti - mitä ja miten tarkistetaan



- **Graafisesti**: Histogrammit, laatikko-jana-kuviot intervallimuuttujilla, luokkamuuttujilla pylväskuviot
- **Taulukoista**: silmämääräisesti, tunnusluvuilla

1. **Rakenne?**: sarakkeet/muuttujat, **rivit**/”caset” ja **solut**
2. **Puuttuvat, poikkeavat, tunnusluvut?**
3. **Tunnusluvut:** keskiluvut *(mean, median, mode)* ja hajonta-luvut *(SD, kvartiilit eli neljännekset, persentiilit, min., maks., vaihteluväli, kvartiiliväli)*
   - Lasketaan statistiikkaohjelmalla *(SPSS, jamovi)* tai (yleensä) Excelillä
4. **Muuttujien väliset assosiaatiot?:** sirontakuviot, korrelaatiot
5. **Jakautumat?:** normaalijakauma, vino, huipukas, 2-huippuinen?



Kuva Wikipedia

Aineiston rakenne

- **Sotkuinen ”messy data”➔ siisti ”tidy data”**
- **Tidy data**= muuttujat omissa sarakkeissaan, tilastoyksiköt *(=caset, havainnot)* omilla riveillään, arvot omissa soluissaan
- **Rakennemuokkauksia**
  - Jonkin **muuttujan luokilla** saattaa olla **omat sarakkeensa** (esim. naisille ja miehille oma erillinen sarakkeensa painolle➔uusi sarake ”sukupuoli” ja vain yksi sarake ”paino”
  - Muuttujien **yksiköt epäyhtenäiset** ➔ korjataan
  - Soluissa joskus useampi arvo ➔ vain 1 arvo (toiset delet. tai lisäsarake)
  - Poikk: **case-control**-tutkimuksessa samalle riville sekä case että control; **aikasarjat**
  - **Monitaso-datat** hankalia, hierarkk. muuttujina esim. yksikkö>potilas>vauriokohdat
  - Muuttujille pitää olla luotettavan tunnistuksen yksilöllinen **avainmuuttuja**
  - Toisinaan kannattaa myös tallentaa esim. alkuperäinen rivijärjestys omana **järjestysnumero**muuttujanaan, jotta sen järjestyksen voi aina halutessaan palauttaa

Puuttuvat arvot I/II

-**tee kaikkesi** välttääksesi puuttuvia arvoja!!!

- **Puuttuvien määrä ja luonne tutkitaan** ja **raportoidaan** myös toimenpiteet *(Tarkistuslistat Strobe, Consort)*
- **Satunnainen vai systeeminen kato muuttujissa -** *esim. Little's MCAR Test SPSS:llä tai R-kielellä*
  - **MCAR (***missing completely at random):* Todennäköisyys puuttumiselle ei liity muiden havaittujen (tai havaitsemattomien) muuttujien arvoihin eikä itse puuttuvan datamuuttujan arvoihin
  - **MAR** (*missing at random*): puuttumistodennäköisyys on sama vain havaitun datan (muut muuttujat) määrittelemien ryhmien/luokkien sisällä. MAR on paljon laajempi luokka kuin MCAR
  - **MNAR** (*missing not at random*): puuttumistodennäköisyys vaihtelee meille tuntemattomista syistä, koska puuttuvan datan todennäköisyys liittyy systemaattisesti puuttuviin hypoteettisiin arvoihin
- **Vrt**: jos **vastaus-% huono**, paljon **total nonresponse**-tapauksia ➔ non-response bias eli participation bias

Puuttuvat arvot I/II

-**tee kaikkesi** välttääksesi puuttuvia arvoja!!!

- **Seuraukset puuttuvista arvoista** (elleivät aivan yksittäisiä harvoja)
  - Jos puuttuvat täysin satunnaisia *(**MCAR**)*, **jättämällä pois analyyseistä puutteelliset rivit (”incomplete cases”) (**=list-/casewise deletion) tai vain puuttuvat **arvot (**pairwise deletion**)** menetetään pelkästään voimaa.
  - Muissa *(**MAR** **ja** **MNAR**)* seuraa **myös bias**ta ja tilanne vielä pahenee monimuuttuja-analyyseissä
- **Puuttuvien arvojen korvaaminen eli imputaatio**
  - **Mean imputation:** korvaus rivi- tai sarake**keskiarvolla** - simppeli ja käytetään, vaikka ei juuri suositella!
  - **[+/-*Stochastic*] regression imp.**: korvaus regressioyhtälön ennusteella [+/- *satunnaisvirhetermillä*] - parempi, mutta sekä mean että regr. Imputaatiossa SD liian pieni ➔signifikanssit ja luottamusvälit valheellisen hyviä
  - paras moni-imputaatio (**multiple imputation),** (käy jos MCAR tai MAR), kts. **MI:stä tarkemmin Slide 72-!**
- *Lisätietoa esim*: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/

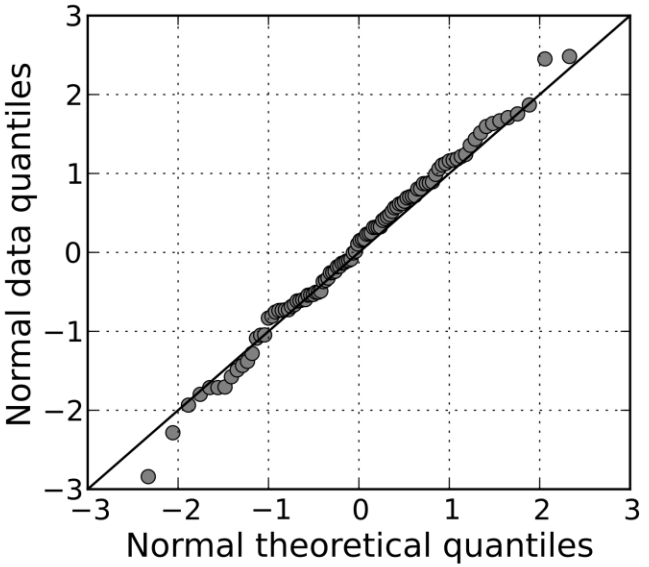
Poikkeavat arvot

- **Poikkeavat** arvot havaitaan joko graafisesti tai erilaisilla testeillä

1. **Outliers**: poikkeava arvo y-akselin suunnassa
2. (high) **leverage**
3. (high) **influence,** kun on molemmat e.m. poikkeavuudet
   - Vaikuttaa paljon esim. lineaarisen regression regressiolinjaan
   - **Cook's distance** >1 on suuri influence. Jos käyttää yhtä lukua, niin tätä!
   - Voi myös katsoa, paljonko analyysitulos muuttuu, jos k.o. arvo poistetaan

- **Mitä tehdä?**
  - Pääsääntö: **virheellisiä?** (=poistetaan/korjataan), **todellisia?** (=jätetään)

Normaalijakaumaoletus



Kuva Wikipedia

- **Normaalijakaumaoletus** keskeinen, koska silloin voidaan käyttää tehokkaampia ns. parametrisia testejä,
  - käyttävät esim. mean- ja SD-parametreja (t-test, Pearson-korr, ANOVA ym.)
- **Milloin (riittävän) normaalijakautunut?** *(ref. Lauri Nummenmaa: Tilastotieteen käsikirja)*
  1. Jos normaalijakauma**testi** (Shapiro-Wilk, Kolmogorov-Smirnov) p>0.05, aineisto normaalijakautunut (ellei pieni, muutaman kymmenen aineisto). Testit huonohkoja: Isoissa aineistoissa liian herkkiä ja pienissä liian epäherkkiä poikkeavuuksille normaalijakaumasta
  2. **TAI**: Jos testin perusteella ei normaalijakautunut, näytöksi normaalijakaumasta riittää, että **visuaalisesti normaalijak. PLUS** että **vinous ja huipukkuus** ovat noin **välillä -1…+1**
  - Visuaalinen tarkastelu: esim. histogrammi ja/tai QQ-plot, (=kvantiilikuvio) **KUVA**!
  - Mitä isompi aineisto, sitä perustellummin voi normaalijakaumaoletuksen tehdä (central limit theorem)
  - Alarajana normaalisuusoletukselle: n on luokkaa (20 -) 30
- **Vinoissa** jakaumissa voidaan joskus yrittää muunnoksilla (√, log, 1/x ym.) saada jakaumaa normaaliksi – mutta voi törmätä uusiin hankaluuksiin

Muuttujien analyysiä ja muokkausta Excelillä

- Näitä voidaan tehdä statistiikkaohjelmissa
- Excel käytännössä helpompi ja monipuolisempi erilaisille tarkistuksille ja muokkauksille
- Excelin tavallisimpiin funktioihin kannattaa tutustua, niillä voi työläässä aineistossa säästää aikaansa viikkokausia!
- SPTY:n sivustolta https://www.spty.fi/tutkijoille/statistiikan-linkkeja/ löytyy laatimani Tutkijan Excel-manööverien basicit_v.211023.xlsx
- Tavallisimpia funktioita: **SUMMA(), KESKIARVO(), MAKS(), MIN(), JOS(), MEDIAANI(), LASKE.JOS(), LASKE.A()** (=laskee solut joissa sisältöä), **LASKE.TYHJÄT()**
  - Esim. ”kombo”, (composite), joka monen muuttujan summa tai keskiarvo
- Muita hyödyllisiä: **makrot**, funktioiden **monistaminen/kopioiminen**

3. Yleistä statistista analyyseistä

Tilastotiede – mitä ja miksi?

- **Menetelmätiede:** miten havainnoista johdetaan substanssitieteen ilmiöistä päteviä päätelmiä, joiden luotettavuus (=P) ja haarukka (=CI) tiedetään
- Käyttää **seuraavia menetelmiä** (sama testi voi kuulua moneen eri ryhmään):
  1. **Deskript. statistiikka**: tunnusluvut *(keskiarvo, keskihajonta ym.)*, graf.
  2. **Tilastoll. päättely (statistical inference)**, (H0-testaus, P, CI): voidaanko otoksen tulokset yleistää kohdepopulaatioon (*t, ANOVA*)
  3. Muuttujien **yhteisvaihtelun** tutkiminen (*korrel., faktorianalyysi*)
  4. **Ennustaminen, mallintaminen** (*regressioanalyysit ym.*)
  5. **Lajittelu ja ryhmittely** (*logistinen regressio, erottelu- ja ryhmittelyanalyysi*)

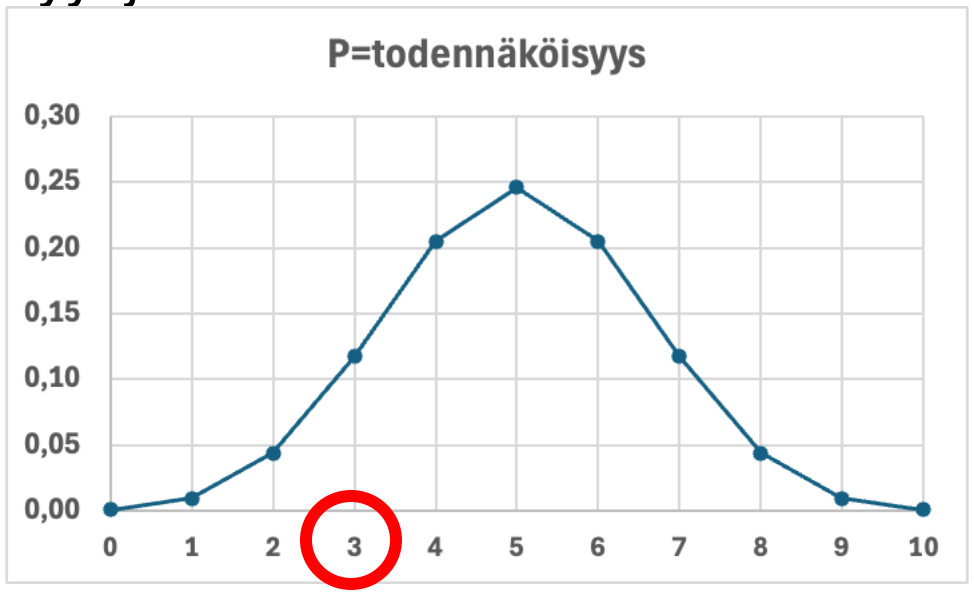
Sample – Study population – Target population

- **Otos/sample** on se osa tutkimuspopulaatiota, jota tutkitaan, testataan, analysoidaan tiedon keräämiseksi
- **Tutkimuspopulaatio/study population** on se ryhmä, mihin **realistisesti** tulokset todella voidaan soveltaa
- **Kohdepopulaatio/target population** on laajempi ryhmä, johon **ideaalisesti** haluamme tuloksia soveltaa, vaikka kyseessä on populaatio, joka on yleensä aivan liian laaja toimiakseen otoksen poimintapohjana

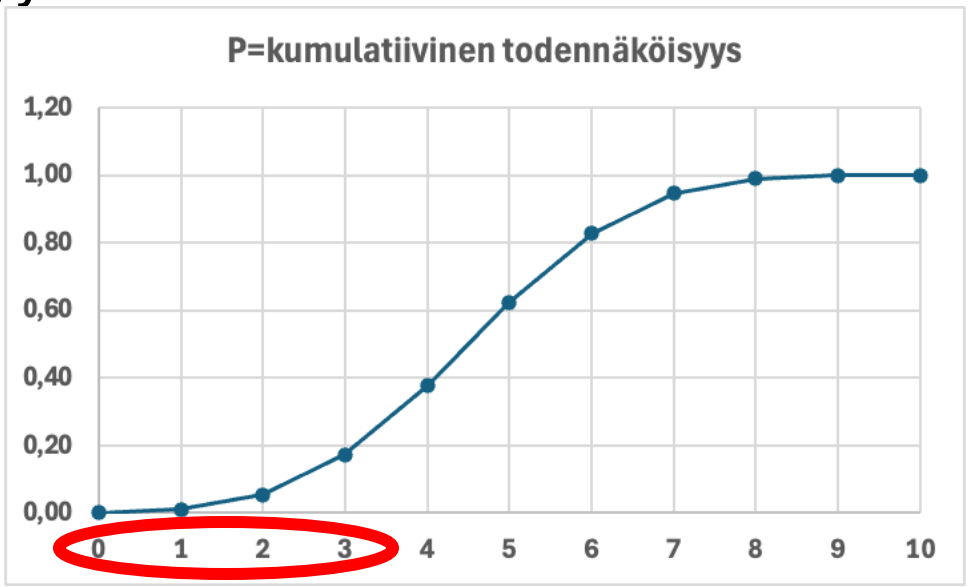
Todennäköisyysjakaumat ovat statistiikan kovaa ydintä!

- Kun heitetään kolikkoa tai noppaa, voidaan todennäköisyydet tuloksille, vaikkapa 3 kruunaa 10 heitolla, laskea tarkasti (ns. *binomitodennäköisyys*).
- Todennäköisyysjakauma kertoo ne todennäköisyydet tässä **suoraan**

Kruunat	P=toden-näköisyys
0	0,00
1	0,01
2	0,04
3	0,12
4	0,21
5	0,25
6	0,21
7	0,12
8	0,04
9	0,01
10	0,00



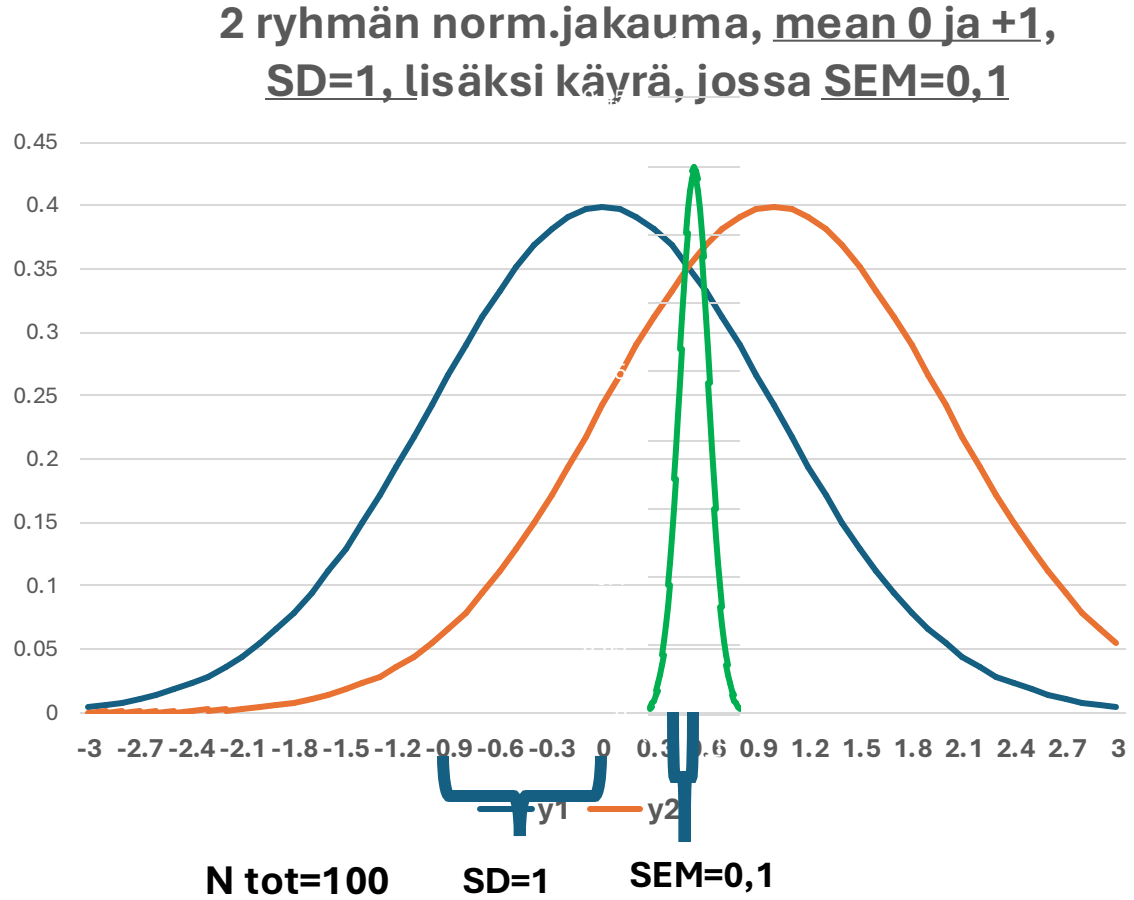
”Tiheysfunktio”=todennäköisyys saada vaikkapa 3 kruunaa 10 heitolla



”**Kertymäfunktio**”=todennäköisyys saada vaikkapa 0-3 (=max.3) kruunaa 10 heitolla

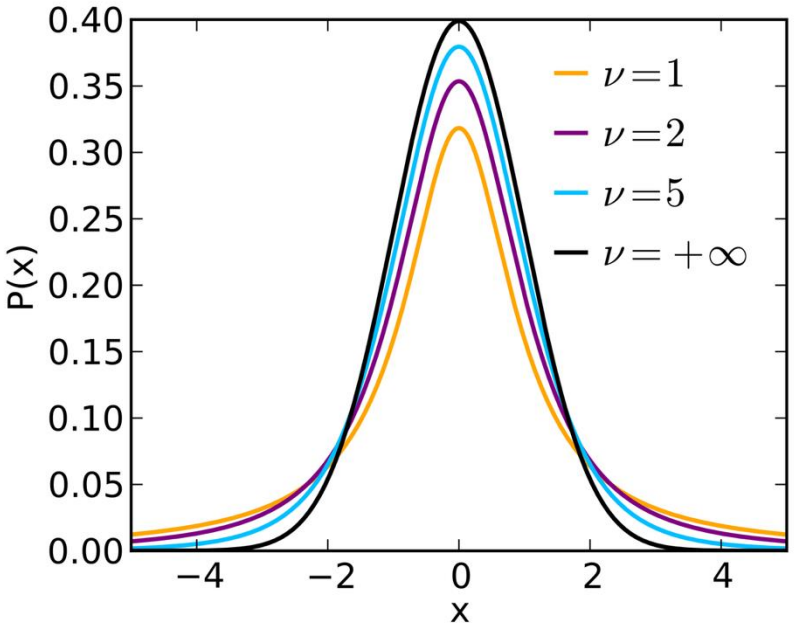
Esim: P-arvon laskemisen idea 2 riippumattoman otoksen **t-testissä**, sinisen ja punaisen ryhmän eron p - ei rakettitiedettä!

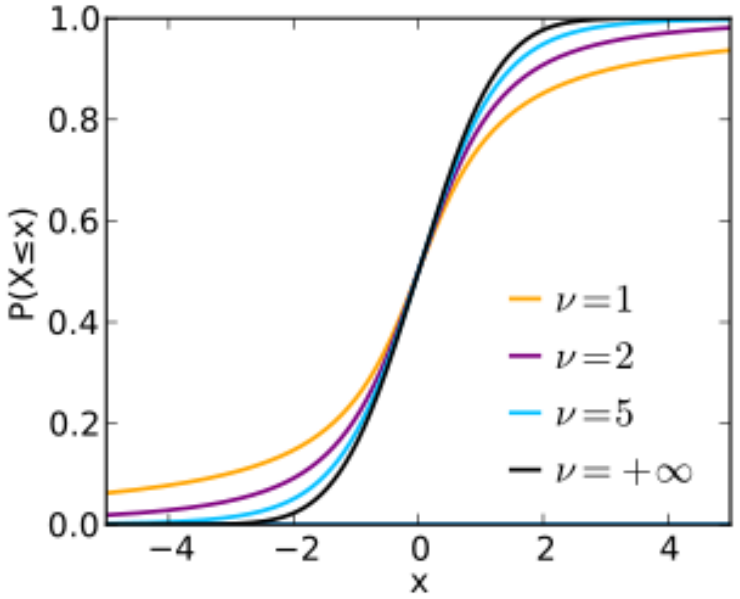
1. Todennäk.jakauman **testisuure t:n kaava**
2. **t = (mean1-mean2)/(keskivirhe)**
   jossa (keskiarvon**) keskivirheen kaava:**
   SEM = SD / √n
3. eli: *t=”montako keskivirheyksikköä on 2 ryhmän keskiarvojen ero”*
4. **SEM= 1/√100** = 1/10 **SEM=0,1**
   (*SEM= SE= standard error of mean)*
5. **t =(1-0)/0,1 eli t=10, df = n1+n2-2 df=98**
   ➔ 2 ryhmän keskiarvojen ero=10 SEM’iä!
6. Poimitaan Studentin t-**todennäk. jakaumasta** vastaava **P-arvo P<0,001**



t-testi esimerkkinä P-arvon määrittämisestä, jatkoa…

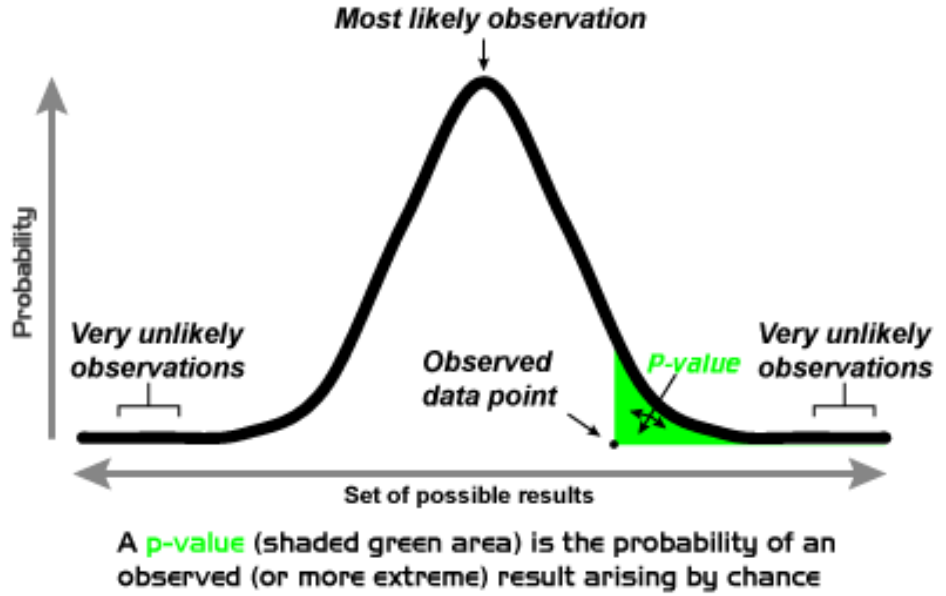
- ***”t-testissä (Studentin t-jakaumaa käyttäen) testisuure t=10,0, vapausaste=98, p=<0.001”***
- Todennäköisyysjakaumia on muitakin: *normaali-, χ^2- ja F-jakauma ym.*
- t-jakauma (tiheysfunktio alla, kertymäfunktio oik.) lähenee isoilla n:illä normaalijakaumaa.





Kuvat Wikipedia

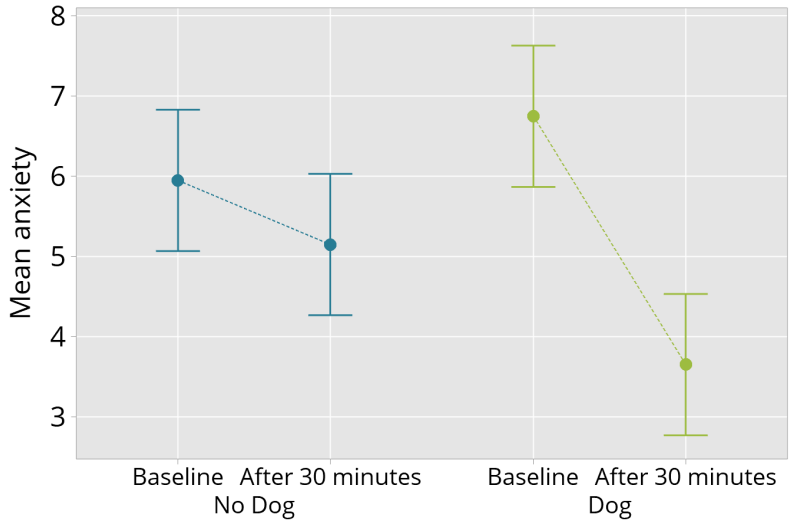
P-arvosta
-mikä eron statist. signifikanssi?



Kuva Wikipedia

- H0-hypoteesi = eroa ei ole ryhmien välillä
- H1-hypoteesi = eroa on
- **P-arvo kertoo, kuinka todennäköistä on kyseisen tuloksen tai vielä sitäkin poikkeavamman saaminen silloin, kun 0-hypoteesi on voimassa!**
- Maagista kriittistä arvoa ei ole, perinteisesti rajana <0.05, nyt yl. **<0.01**
- Suurissa aineistossa todelliset erot tulevat useammin esiin, pienissä aineistossa **Power**/tilastollinen voima on pienempi.
  - *Joka 20. homeopatiatutkimuksessakin eron P <0.05, sattumalta!*
- **Monitestausongelma**: 20 P-arvoa➔tod näk. 1:ssä sattumalta P<0.05
- **P ei kerro mitään eron suuruudesta**, siksi sen lisäksi tai sijaan on enemmän alettu käyttää esim. **95% luottamusvälejä**

Mikä on eron suuruus?



A. Keskiarvon 95% luottamusväli (CI)

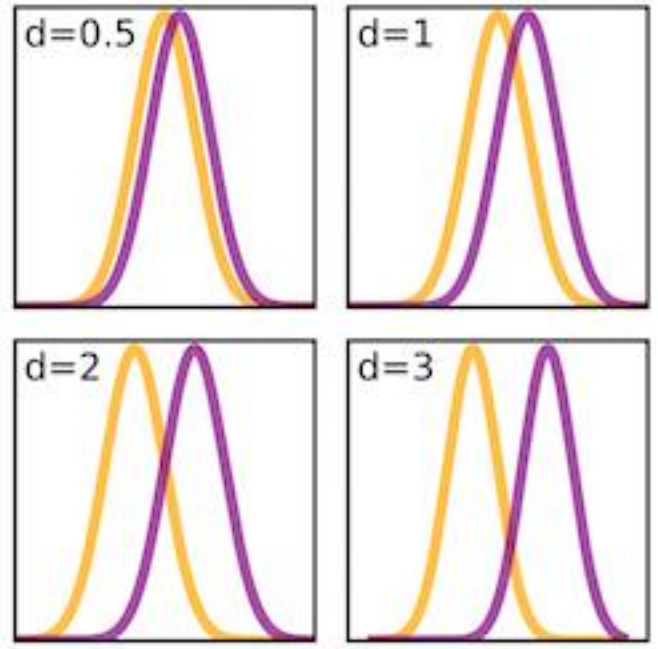
- 95%CI kuvaa väliä, jolle 95% keskiarvoista osuisi, jos vastaavia otoksia poimittaisiin äärettömän monta (kuva)

Kuvat Wikipedia

- Jos 2 keskiarvon luottamuusvälit eivät kohtaa, myös P varmasti <0.05

B. Efektikoko (Effect size)

- Tapa 1: Kuinka monen **keskihajonnan verran** ryhmät eroavat (Cohenin d; kuva)
- Tapa 2: muuttujien välisen **korrelaatiokertoimen avulla**, näin saatavat arvot vajaa puolet edellisistä



Erilaisia todennäköisyyden, riskin ja riskien suhteen epidemiologisia mittareita

- **Todennäköisyys** 0 - 1 (eli 0% - 100%)
- **Riski** = (ikävän) tapahtuman todennäköisyys
- **Risk ratio**, **relative risk**=RR=Risk1/Risk2. **Hazard ratio,HR**: tietyssä ajassa
- Riski A 10%, B 5%: **Abs. riskiero** 0,10-0,05=0,05 eli **5%**, **Rel. riskiero** 0,05/0,1=**50%**
- **Odds** =vetokerroin, vastasuhde, (veto, vedonlyöntisuhde), =todennäk., että voittaa (vedossa)/ että ei voita
- **Odds ratio (OR, vetosuhde,** ristisuhde**)** = Ryhmän Odds / Refer.ryhmän Odds
  - *Ehkä selkeintä käyttää Odds'ista ja Odds ratio'sta vain näitä engl. termejä!?!*

1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10

Riski=2/10=0,2 **Odds=2/8=0,25**

Riski=1/10=0,1 **Odds=1/9=0,111…**

RR=0,2/0,1=2 **OR=0,25/0,11=2,25**

Mitä statistista ohjelmaa käyttäisi?

- **Pragmaattisesti**: sitä, mikä on tarjolla, mitä osaa käyttää ja joka sisältää tarvitt. menetelmät
- **Excel**: Excelin käyttöä tieteellisissä analyyseissä pidetään ”sivistymättömänä”, vaikka se sinänsä sisältää funktiot lukuisille analyyseille
- **SPSS**: Laajalti käytetty, mutta opiskelija- ja yliopistolisenssien ulkopuolella se on kallis, joka moduulilla oma vuosimaksunsa. Monipuolinen, mutta hieman kömpelö ja hidas, kankeampi aineiston muokkauksessa kuin Excel
- **jamovi**: (*pienellä alkukirjaimella!*) ilmainen ohjelma, ladattavissa https://www.jamovi.org, jossa runsaat ohjeet. Perusmoduulin lisäksi kohtalaisesti laajennusmahdollisuuksia moduulikirjaston kautta. Helppo oppia alkeet, nopea käyttää. Valinnanmahdollisuudet jkv rajoittuneita, grafiikka ”korutonta” ja vailla riittäviä muokkausmahdollisuuksia, eli valtaosin ei esitysgrafiikkatasoa
- **R-kieli**: Jos tarvitsee erikoisanalyysejä, tilasto-ohjelmointi R-kielen yli 2000 kirjaston avulla mahdollistaa ”kaiken”, mutta se on hieman työläs opetella ja ylläpitääkseen taidon pitäisi työskennellä sen kanssa melko usein. RStudio tarjoaa R-kielelle erinomaisen käyttöliittymän.

4. Analyysimenetelmän valinta

Millä perusteilla analyysimenetelmä valitaan? I/II

- Yleensä tarjolla useampi teoriassa mahdollinen vaihtoehto
- Pelkässä deskriptiivisessä statistiikassa ei valita mitään statistisen signifikanssin tuottavaa testiä
- Yleensä **kannattaa käyttää yksinkertaisinta** ehdot täyttävää metodia, etenkin jos n on pieni
- Mitä sofistikoidumpi analyysi, sitä isompi aineisto yl. tarvitaan
- Oma **osaaminen** ja tilasto-ohjelman version sisältämät metodit
- ”Yllättävät” valinnat kannattaa **perustella** huolella ja asianmukaisella terminologialla metodiosassa!
- Valintakaavioita on olemassa lukuisia, sisällöltään hieman vaihtelevia

Millä perusteilla analyysimenetelmä valitaan? II/II

- **Muuttujien määrä**: 2 vai useampia *(esim. t-testi vs. ANOVA)*
- **Muuttujien asteikko**: nominaali-ordinaali-intervalli tai enemmän
- **Jakauman normaalisuus (N),** jos intervalliasteikollinen
  - On normaalijak.: **parametriset** testit (= voivat käyttää mean ja SD), esim. t-testi, ANOVA, Pearson-korrelaatio
  - Ei normalijak.**: nonparametriset** (perustuvat järjestykseen; hieman heikompia), esim. t-testi ➔ Mann-Whitney U-testi tai Wilcoxon (jos parillinen), ANOVA ➔ Kruskal-Wallis, Pearsonin korr. ➔ Spearman
- **Havaintojen määrä** per ryhmä ja/tai per selittävä muuttuja (esim. monimuuttuja-analyyseissä)
- Onko 2 muuttujan **pelkkä yhteisvaihtelu** *(➔korrel.)* vai **”toinen selittää toista”?** *(➔regressio)*
- Täyttyvätkö analyysin **erityisehdot** (tutkitaan ”diagnostiikalla”)

5. Tavallisimmat statistiset menetelmät

Deskriptiivinen statistiikka määrällisillä muuttujilla - mitä tunnuslukuja raportoidaan?

-(nominaali (ja ordinaali) asteikollisilla: luokkafrekvenssit ja %)

Normaalijakautunut

- Keskiarvo
- Keskihajonta (SD)
- Vaihteluväli, min, max
- *Usein aineisto ”hybridi” normaalista ja ei-normaalista ja joudutaan erilaisiin kompromisseihin tunnusluvuissa*

Ei-normaalijakautunut

- Mediaani (md, puolet suurempia, puolet pienempiä)
- Moodi (mo, yleisin arvo)
- Alakvartiili (Q1), yläkvartiili (Q3)
- Kvartiiliväli (IQR)=Q3-Q1
- Vaihteluväli, min, max
- Vinous ja huipukkuus

Ristiintaulukoidut luokkamuuttujat

(Crosstable)

Ristiintaulukoitu 2 luokkamuuttujan data

- Ristiintaulukoiduille nominaaliasteikollisille eli luokkamuuttujille vain niukasti testejä: **Khiin neliö (χ^2) riippumattomuustesti** (*Engl. The Chi-Square Test of Independence TAI Chi-Square Test of Association*) **ja Fisherin testi**
- Testataan**, ovatko 2 luokkamuuttujaa yhteydessä toisiinsa?** (assosiaatio)
- Samoja testejä voi **käyttää myös järjestysasteikollisille** muuttujille, mutta silloin se käsitellään nominaaliasteikollisena ja aineiston järjestykseen sisältyvä suurempi analyysivoima menetetään
- **Erityistilanteita**
  - jos verrataan samaa otosta ennen-jälkeen (**paired samples**), tai esim. tapaus-verrokki-asetelmassa, käytetään ”parillista testiä”: **McNemarin testi**
  - Jos yhden luokkamuuttujan jakaumaa **verrataankin oletukseen**, esim. onko datassa F:M suhteessa 1:1, käytetään ***χ^2 yhteensopivuustestiä*** *(goodness-of-fit)*

Pearsonin Khii toiseen, χ^2,(Chi Square)-riippumattomuustesti

		Muuttuja 2 sukupuoli		
		mies	nainen	Yht
Muuttuja 1	usein	20	30	50
	toisinaan	15	15	30
	harvoin	10	5	15
	Yht	45	50	95

- Ovatko 2 luokkamuuttujaa yhteydessä/ riippuvuudessa toisiinsa?
- Testin **idea**
  - taulukon reunajakaumien perusteella lasketaan odotetut (Expected, E) frekvenssit, joita verrataan havaittuihin (observed, O) frekvensseihin ja tästä saadaan testisuure χ^2
  - Mitä enemmän O ja E poikkeavat toisistaan, sitä isompi χ^2 ja sitä pienempi P
  - Testi perustuu χ2 –todennäköisyysjakaumaan, josta siis poimitaan P
- **Käytön edellytykset** (ohjelma raportoi kyseiset %-luvut):
  - < 5 suuruisia odotettuja (E) frekvenssejä saa olla viidesosa (20 %) kaikista odotetuista frekvensseistä ➔ eli 2x2 taulukossa ei yhtään!
  - < 1 suuruisia odotettuja (E) frekvenssejä ei saa olla ollenkaan
- *N.s. Yatesin jatkuvuuskorjaimen käyttöä ei χ^2:ssa suositella*
- *P.S. Taul. esimerkissäkin Muuttuja 1 on järj.asteikollinen ➔analyysivoimaa menetetetään*

Khiin neliö eli χ^2-testin suoritus ja tulkinta

- **Valitaan:** χ^2-ruksi ja rivi- & sarakemuuttuja, niiden järjestys ei vaikuta P-arvoon!
- Jos on vaihtoehdot 1- ja 2-suuntainen p, 1-suuntainen valitaan vain poikkeustapauksissa
- **Tulos katsotaan:** χ^2-riviltä, ohjelma ilmoittaa **χ^2**-arvon, vapausasteen (=**d.f.**) ja **P**-arvon
- Myös Fisherin P saattaa näkyä ohjelmasta ja valinnoista riippuen
- Terminologiaa: ristiintaulukointi = kontingenssitaulut

χ^2 Tests			
	Value	df	p
χ^2	42.215	22	0.00589
N	558		

Kuva jamovista

Fisherin eksakti testi

- Poikkeuksellinen idea: ei käytä mitään todennäköisyysjakaumaa (*norm, t, χ^2, F*), vaan ohjelma pyörittää läpi kaikki hyvinkin lukuisat mahdolliset vaihtoehdot ja laskee nykyistä vielä poikkeavimpien osuuden, mikä =P
- ➔Ohjelma voi raksuttaa minuutteja, ehkä valitaan rajaksi 5 min?
- Tulos ei siten ole todennäköisyysjakaumasta saatu pyöristys eli approksimaatio, vaan täysin eksakti arvo, ellei annettu aika ollut riittämätön
- ➔Fisher kannattaa valita **aina**, jos PC jaksaa vääntää, eli ainakin pienissä 2x2, 2x3, 3x2, 3x3-taulukoissa ja kokeilla isommissakin
- Muita rajoituksia ei ole kuin PC:n vääntö, joka nykyään harvoin ongelma
- ➔**”Käytä ristiintaulukoissa aina Fisheriä, kokeile ainakin ensin!”**

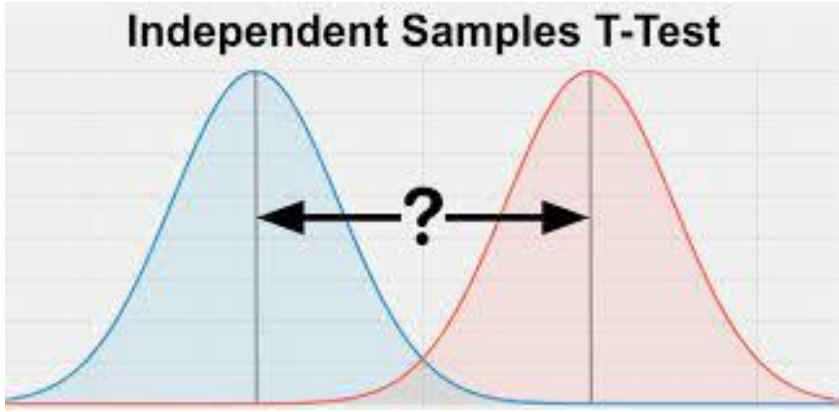
2 tai useamman määrällisen muuttujan ryhmän välinen vertailu

Stat. Menetelmät:≥2 ryhmää, määrällinen muuttuja

≥2 määrällisen muuttujan ryhmän välinen vertailu - yleistä

Mitta-asteikko	**Ordinaali**	**Intervalli**	**Intervalli**
Jakauma	**–**	**Ei normaalijakautunut**	**Normaalijakautunut** (ja ryhmien koko ≥20-30)
Riippumattomat otokset (2 kpl)	**Mann-Whitneyn U-testi** (**eli** Mann-Whitney-Wilcoxonin rank sum)	**Mann-Whitneyn U-testi**	**t-testi (riippumattomat otokset)** ***Student/Welch***
Riippuvat eli parittaiset otokset (2 kpl) (ennen-jälkeen, case-control)	**Merkkitesti** (sign test**) tai** **Wilcoxonin merkittyjen järjestyslukujen testi** (signed rank)	**Merkkitesti** (sign test**) tai** (jos jakauma symmetrinen) **Wilcoxonin merkittyjen järjestyslukujen testi** (signed rank)	**parittainen t-testi (riippuvat otokset)**
≥3 riippumatonta otosta	*Kruskal-Wallis-testi*	*Kruskal-Wallis-testi*	*1-way-ANOVA (varianssianalyysi)*

(Studentin) riippumattomien otosten t-testi
– parametrinen testi, 2 ryhmän vertailu



Kuva Wikipedia

Milloin mahdollista käyttää Studentin t-testiä?

1. **Normaalijakautuneita intervalli**muuttujia molemmat ryhmät
2. **Riippumattomat otokset** (*myös olemassa toinen versio: parittainen mittaus, kts myöh.*)
   - Havainnot ryhmässä eivät korreloi toisten kanssa/eivät millään lailla yhteydessä
   - Ei riippuvuuksia ryhmien välilläkään (esim. samoja havaintoja)
3. **Varianssit yhtä suuria** (”homogeenisuus”; eli Levenen testi p>0.05).
   - Jos eri suuria (p<0.05), kutsutaan **Studentin t-testin** sijaan myös **Welch-testi**ksi
4. Ryhmien **minimikoko 20 - 30 kpl**

Muuta

- **Miltei aina valitaan 2-suuntainen p** (ellei toinen suunta ”mahdoton” tms)

Studentin t-testin sijaan kannattaa yleensä käyttää Welchin testiä

- Ideaalitilanne, että varianssit ovat täysin identtiset on harvinainen
- Muutkaan t-testin ehdot (normaalisuus, havaintojen riippumattomuus muista intra-/inter-group) harvoin toteutuvat ideaalisesti
- Welchin testi huomioi varianssien eron, siksi sen power jkv heikompi
- SPSS:ssä p-arvo katsotaan erisuuruisten varianssien tapauksessa alemmalta riviltä. Jamovissa kruksataan Studentin t-testin ohjelmassa myös ”Welch’s” ja saadaan sen(kin) tulokset
- Jos siis aineistosi on ideaalinen (eli uskot myös joulupukkiin…), käytä Studentin t-testiä, mutta **reaalimaailmassa kannattaa aina käyttää Welchin testiä** (jota saatetaan myös kutsua Studentin testiksi, jossa varianssit eivät homogeeniset)

Mann-Whitney U testi

– nonparametrinen 2 ryhmän vertailu, määrällinen muuttuja

Milloin käytetään?

- Määrällinen muuttuja **vähintään ordinaaliasteikollinen**
- Jos muuttuja intervalliasteikollinen, se **ei** ole **normaalijakautunut ja/tai ryhmien koot <20-30**
- **Ryhmiä 2**: riippumattomat otokset **Mann-Whitney U test**,

Muuta

- **Ryhmät** voivat siis olla **pieniäkin**, <20, mutta jos <10, vaikea saada enää statistisesti signifikanttia eroa
- SPSS: Analyze/nonparametric
- Nonparametriset testit hieman heikkotehoisempia kuin parametriset
- Parittaisissa vertailuissa Wilcoxon) ja jos >=3 ryhmää: U testin sijaan Kruskal-Wallis test

Riippuvien otosten (eli parittaiset) testit I/II

(engl. *paired, matched, related*)

- Vertaillaan 2 ryhmää, niin että samaa tai sitä muistuttava tilastoyksikköä mitataan kahdesti
  1. vertaillaan samoja tilastoyksiköitä kahdesti (**ennen – jälkeen** -asetelma)
  2. **Tapaus-verrokki**-tutkimuksessa käytetään kaltaistettuja verrokkeja
- Käytetään paljon harvemmin kuin riippumattomien otosten testejä
- Etuna: kun samaa tilastoyksikköä verrataan itseensä tai kaltaiseensa, satunnaisvaihtelua on vähemmän➔ Power↑
- Parittaisia testejä
  1. Luokka-asteikko, 2 Lk, eli **dikotominen asteikko** (esim. luokat 0 ja 1): **McNemarin testi**
  2. **Järjestysasteikko: Merkkitesti** (*sign test*) tai **Wilcoxonin testi**
  3. **Intervalliasteikko, ei-normaalijakaumaa** (mutta silti symmetrinen): **Wilcoxonin testi**
  4. **Intervalliasteikko, normaalijakautunut: parittainen t-testi**

Varianssianalyysi (ANOVA) - yleistä

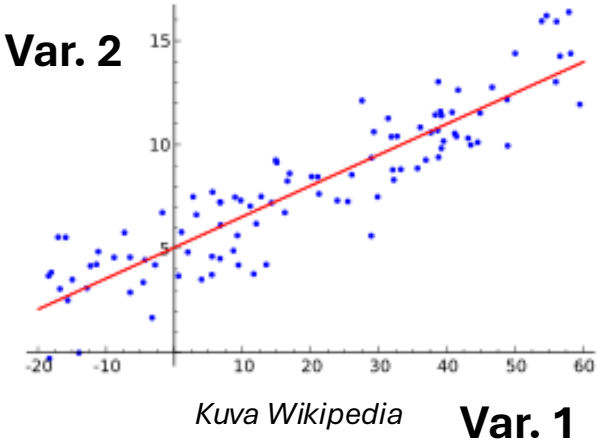
-kun verrattavia ryhmiä ≥3

- **Idea**:
  - Kokonaisvaihtelu jaetaan ryhmien sisäiseen ja ryhmien väliseen.
  - Mitä suuremman osan vaihtelusta ryhmien välinen vaihtelu kattaa, sitä suurempi on ero ryhmien välillä
- **Eri tyyppisiä** testejä *(ryhmittelijöiden lukumäärän mukaan)*
  - 1-suuntainen varianssianalyysi (1-way-ANOVA), *esim. sukupuolen mukaan*
  - 2-suunt. varianssianalyysi (2-way-ANOVA), *esim. sukupuolen ja kotipaikan mukaan*
- **Post hoc-analyysit**
  - ANOVA ei suoraan kerro, minkä ryhmien välillä ero on
  - Ryhmien välisissä jatkoanalyyseissä tehdään useita parittaisia vertailuja
  - Lukuisia menetelmiä valittavana
  - Tällöin pitää huomioida myös monitestausongelma! *(slide 24)*

Yhteisvaihtelun analysointi

-korrelaatio ja regressio, faktorianalyysi

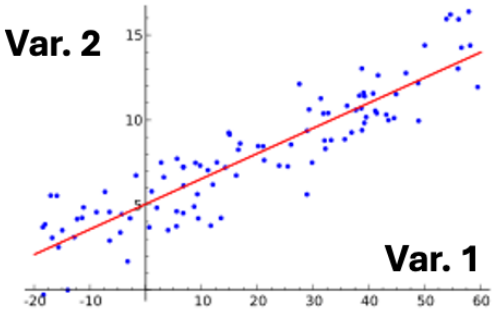
Korrelaatio, yleistä



Kuva Wikipedia

- 2 kvantitatiivista muuttujaa, näiden välillä **yhteisvaihtelua**
- Korrelaatiossa **ei oleteta, että muuttuja x selittää muuttujaa y**, vaan niiden välillä on yhteys *(jos oletettaisiin yhteydelle suunta, kyseessä olisi jo regressioanalyysi!)*
- **Kerroin r** kuvaa yhteisvaihtelun voimakkuutta ja suuntaa, -1 – 0 – +1
  - Useita **tasoluokitteluja** siitä, mikä matala ja mikä korkea r, mutta nämä tasot eroavat eri tieteenalojen välillä paljonkin
  - Korrelaatiot, joissa r itseisarvoltaan **<0.1,** ovat yleensä aika **merkityksettömiä** (Pearsonin r:ssä sitä vastaava selitysaste <0.01 eli < 1%!)
- **P** kuvaa stat. signifikanssia sille, eroaako r nollasta
- Esitetään yksittäisinä tai isommissa korrelaatio**matriiseissa**
- Intervallimuuttujan arvojen vertailu 2 ryhmässä voidaan edellä esiteltyjen testien lisäksi analysoida myös korrelaationa, riippuen siitä, kiinnostaako enemmän ryhmien keskiarvojen ero (t-testi ym.) tai muuttujan arvojen välinen yhteisvaihtelu 2 ryhmässä!!

Tavallisimpia korrelaatioita



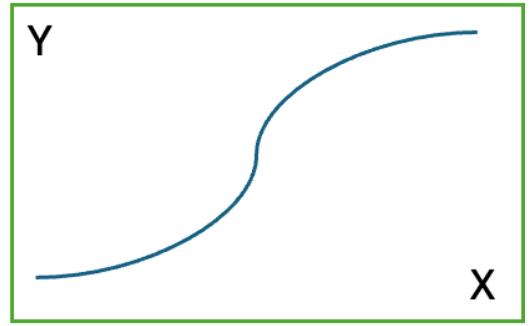
Lineaarinen yhteys

Pearsonin korrelaatio

- **Metodin idea**: Huomioi järjestyksen lisäksi myös välimatkat
- **Edellytys**: normaalijakautuneet intervallimuuttujat, myös dikotom. muuttujat
- Jos muuttujien **yhteys lineaarinen**, eikä hyvin poikkeavia arvoja, Pearson vahvin
- Pearsonin kerroin r potenssiin 2 (=R^2) = **selitysosuus**

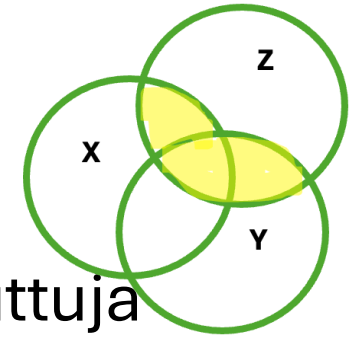
Spearmanin (järjestys)korrelaatio

- **Metodin idea**: Huomioi VAIN järjestyksen
- **Käy**: sekä (1) intervalli-, (2) ordinaali- että (3) dikotom. luokkamuuttujille, mutta ei automaattisesti paras missään näissä.
- **Monipuolinen**, usein korrelaatiomatriiseissa eri mitta-asteikkoisia muuttujia, jolloin Spearman hyvä ”yleiskorrelaatio”
- Pearson vahvempi lineaarisissa yhteyksissä, mutta muissa epälineaar. yhteyksissä (esim. monotoninen kts kuva!) Spearman yleensä vahvempi

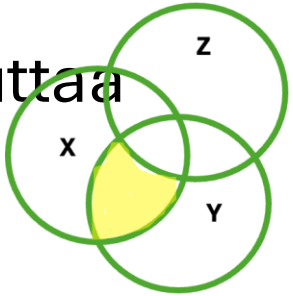


Ns. monotoninen yhteys

Muita korrelaatioita

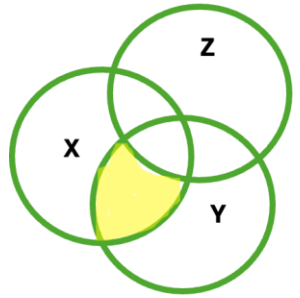


- **Multippelikorrelaatio:** kuinka paljon x ja y tai vielä useampi muuttuja yhdessä ”summavaikuttavat” z-muuttujaan
- **Osittaiskorrelaatio:** kuinka paljon x yksinään ”puhdistetusti” vaikuttaa y-muuttujaan, kun z:n vaikutus on poistettu



- **Erikoistapauksia** *(esim. käsikirjoituksen reviewer toisinaan kysyy/edellyttää!!)*
  - Järjestysasteikolliset muuttujat: **polychoric corr. ”tehokkain”,** harvoin käytetty
  - Aito dikotominen ja jatkuva muuttuja: **piste-biseriaalinen** korr. ideaali, mutta jos dikotominen koodattu 0 ja 1, Pearson antaa aivan identtisen tuloksen!
  - Dikotomisoitu muuttuja ja jatkuva muuttuja: **biseriaalinen** korr., mutta ensisijassa: ei suositella keinotekoista dikotomisointia!
  - ≥3 luokkaista nominaalimuuttujaa ei voi pistää korrelaatioon sellaisenaan, ellei luokkia saa kiistattomaan järjestykseen – jolloin kyseessä jo ordinaalimuuttuja!

Osittaiskorrelaatiosta lisää



- **Osittaiskorrelaatio:** kuinka paljon x yksinään ”puhdistetusti” vaikuttaa y-muuttujaan, kun z:n vaikutus on poistettu. Esimerkiksi monet muuttujat korreloivat keskenään osittain siksi, että molemmat korreloivat iän kanssa
- **Pearson**in lisäksi myös esim. **Spearman**in menetelmällä, joka löytyy esim. **jamovi**stakin
- **Milloin – voi pelastaa pulasta monenlaisissa tilanteissa!**
  - Useiden **muuttujien väliset vahvat korrelaatiot** hankaloittavat korrelaatioiden selvittelyä
  - Spearmanin osittaiskorrelaatio myös, kun **lineaarinen regressio ei mahdollista** (esim. ordinaaliasteikko, liian pieni otos, poikkeava jakauma, yhteys ei ole lineaarinen, outlier’eitä tms. poikkeavia arvoja)
- Voidaan myös käyttää **useita kontrolloivia** Z-muuttujia
- Perinteinen korrelaatio ja osittainen korrelaatio ilman Z-muuttujia ovat tuloksiltaan identtisiä
- Osittaiskorrelaatio r, jossa on kontrolloiva Z-muuttuja, on korkeintaan sama, mutta yleensä vaihtelevasti pienempi, periaatteessa ei koskaan suurempi
- Jos Z korreloi voimakkaasti X:n ja Y:n kanssa, X:n ja Y:n välinen korrelaatio pienenee voimakkaasti Z:n korrel. ”puhdistuksella”, heikosti korreloiva Z taas vähentää sitä vain vähän

Regressioanalyysi - yleistä

- Regressioanalyyseissä selitetään/mallinnetaan matemaattisesti **x**-muuttujien avulla (=**selittäjä**, **explanatory**, selittävä; 1-useita kpl) **y**-muuttujaa (=**vaste, response, output,** selitettävä)
- Pitää selkeästi **erottaa 2 täysin erilaista ”selittämisen” muotoa!:**
  1. **Regressiomallissa** x-muuttujat **selittävät** automaattisesti y-muuttujaa **vain matemaattisesti**
  2. **Reaalimaailmassa** tuo ”selitys” tarkoittaa automaattisesti **vain assosiaatiota**. Kausaalinen selitys edellyttää lisäksi sopivaa tutkimusasetelmaa (esim. RCT, kvasikokeellinen), esim. observationaalinen poikkileikkausasetelma ei kelpaa!
- **Käyttö**: y:n selittäminen, ennustaminen, mallintaminen, luokittelu
- *Näiden esittely vain pintaraapaisu, koska niin paljon erilaisia vaihtoehtoja ja optioita*

Eri tyyppisiä regressioanalyysejä

- **Simple tai multiple** regressio: sen mukaan, onko 1 vai us. selittäjä
- **Lineaarinen regressio:** lineaarinen (eli ”suoraviivainen”) yhteys selittäjien ja jatkuvan vastemuuttujan välillä – *regressioiden perusmuoto*
- **Binäärinen logistinen regressio:** yhteys selittäjien ja 2-luokkaisen vastemuuttujan välillä (multinomiaalisessa ≥3 luokkaa), yhteys raportoidaan Odds Ratio (OR) käyttäen
- **Muita regressioanalyysin muotoja (Kts. Myös slidet 73-74!)**
  - *General linear model (GLM) – yksi yleisesti käytetty lineaariselle regressiolle vaihtoehtoinen metodi*
  - *Ordinal regression – vastemuuttuja järjestysasteikollinen, ehdot (”parallel slopes”) eivät useinkaan täyty*
  - *Cox regression – elinaika/survival on vaste, jota selitetään erilaisilla tekijöillä*
- **Mallin rakentamisessa** tarvitaan myös **ymmärrystä substanssista**, jotta mukaan otettavat selittäjät ovat relevantteja, eikä pelkästään matemaattisesti sopivia tai p-hacking-perusteisia. Jatkuvia muuttujia dikotomisoidaan vain perustellusta syystä.
- Malliin voidaan **joskus** ottaa myös esim. 2 selittäjän signif. **yhdysvaikutuksia** eli **interaktioita** (merkitään: *muuttuja1 x muuttuja 2*), jotka siis poikkeavat muuttujien erillisistä vaikutuksista. Mutta malleista tulee tällöin usein liian monimutkaisia, tulkinnanvaraisia ja hankalia!

Lineaarinen regressioanalyysi - yleistä

- **Idea**: **y=a+bx**
  - 1 tai us. **x-selittäjä** ja sen **b-kerroin** (b_1x_1, $b_2x_{2...}$) sekä **a-vakio** määrittävät **y-vasteen** arvoa
  - Selittäjinä voi olla **myös luokkamuuttujia**, jolloin yksi luokka toimii vertailutasona ja muut luokat pitää (ohjelman tai tutkijan) ”dummy-koodata” Y/N koodeilla 1/0
  - Jäännöstermit eli **residuaalit** kuvaavat sitä osaa vaihtelua, jota malli ei selitä
- **Oletukset/edellytykset**
  1. **Lineaariset** yhteydet muuttujien välillä *(muuten ei selitä y-vastetta ollenkaan)*
  2. **Ei multikollineariteettia:** selittäjien välillä ei saa olla korkeaa korrelaatiota
  3. **Havaintoja** vähintään (10-) 20-50/selittävä muuttuja
  4. **Ei** voimakkaasti **poikkeavia** havaintoja
  5. **Residuaalien tarkastelu keskeistä!**
     - (A) ne ovat normaalijakautuneita
     - (B) niiden varianssit samat (eli homogeeniset) koko y-muuttujan alueella
     - (C) ne ovat riippumattomia toisistaan (ei ”outoja”, vaikkapa U- kuvioita residual ploteissa)

Lineaarinen regressioanalyysi – suoritus - *esimerkkinä jamovi*

- **Valitaan analyysiohjelma**
- **Syötetään muuttujat**
  - ***vaste****muuttuja* ➔*Dependent variable*
  - ***selittävät jatkuvat*** *muuttujat* ➔*Covariate-kenttään*
  - ***selittävät luokka****muuttujat* ➔*Factors-kenttään (joka koodaa ne 0/1- dummyiksi automaattisesti)*
- **Lisävalinnat (***kuvassa oikealla vain osin avattuna)*
  - Luokkamuuttujien (Factors) **viitetason** valinnat *(=reference levels, johon muita tasoja verrataan)*
  - **Oletusten** tsekkaus *(kts slide edellä ja 2 seuraavaa!)*
  - **Selitysasteet** R2, mallin ja kerrointen **P-testaukset**
  - Ilmoitetaanko myös estimaatin **CI** ja **standardoitu** muoto ym.
- **Raportoidaan** mallin signifikanssi (P<0.05?), selitysaste (Adj)R^2, selittäjien estimaatit ja niiden signifikanssi (P<0.05?), kollineaarisuus, jäännöstermien (reciduals) analyysit

Stat. menetelmät : regressio

Lineaarinen regressio-analyysi – suoritus jamovissa

Selitysosuudet

Vaste-muuttuja

Vastemuuttujan muutos, kun selittäjän arvo kasvaa yhdellä

Selittävän muutt:n p

Selittäjät vakioitu, vaikutuksien suuruudet verrattavissa

Model Fit Measures

Model	R^2	Adjusted R^2
1	0.31 1	0.31 0

Model Coefficients - dinhälsa43_hög_är_god

Predictor	Estimate	SE	t	p	Stand. Estimate	95% Confidence Interval Lower	95% Confidence Interval Upper
Intercept [a]	2.969	0.047	63.606	< .00001			
Sum_14_utför_extrov	0.093	0.007	13.397	< .00001	0.179	0.152	0.205
modersmål4_dummy_Sv1_FiRef:							
1 – 0	-0.167	0.026	-6.463	< .00001	-0.165	-0.216	-0.115
utbild_Lev4_dum:							
1 – 0	0.122	0.053	2.319	0.02044	0.121	0.019	0.223
utbild_Lev5_dum:							
1 – 0	0.216	0.036	6.047	< .00001	0.214	0.145	0.284
kännsfys27a_dummy_Yngre1_LikaRef:							
1 – 0	0.634	0.027	23.420	< .00001	0.630	0.577	0.683
kännsfys27a_dummy_Äldre1_LikaRef:							
1 – 0	-0.887	0.054	-16.280	< .00001	-0.881	-0.987	-0.775
ålder	-0.154	0.010	-15.649	< .00001	-0.201	-0.227	-0.176

[a] Represents reference level

Vakio

Covariate

Factor kieli, dikotominen

Factor koulutus, 2 dummya+ref)

Factor koettu ikä, 2 dummya+ref)

Covariate ikä

Onko poikkeavia arvoja? Silloin arvo >1

Data Summary

Cook's Distance

			Range	
Mean	**Median**	**SD**	**Min**	**Max**
0.000	0.000	0.000	0.000	0.007

Korreloivatko residuaalit, DW arvot 0-4; n. 2 on ideaali

Assumption Checks

Durbin–Watson Test for Autocorrelation

Autocorrelation	DW Statistic	p
-0.028	2.056	0.06600

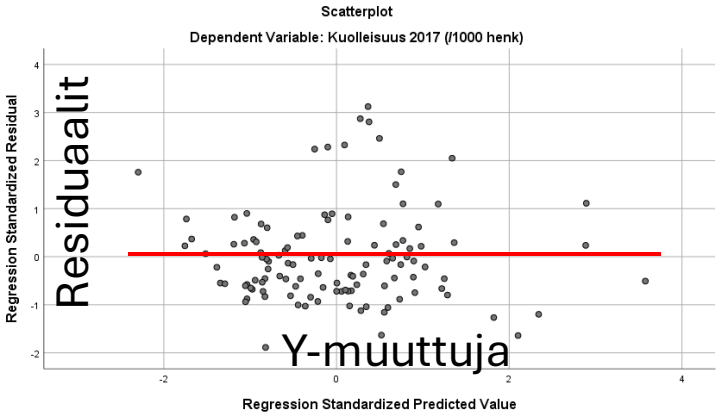
Onko multikollineaarisuutta, Jolloin toler.<0,2; ideaali n. 1. Huom. myös graaf. tarkast.

Collinearity Statistics

	VIF	Tolerance
Sum_14_utför_extrov	1.154	0.867
modersmål4_dummy_Sv1_FiRef	1.012	0.988
utbild_Lev4_dum	1.030	0.971
utbild_Lev5_dum	1.046	0.956
kännsfys27a_dummy_Yngre1_LikaRef	1.077	0.928
kännsfys27a_dummy_Äldre1_LikaRef	1.062	0.941
ålder	1.076	0.930

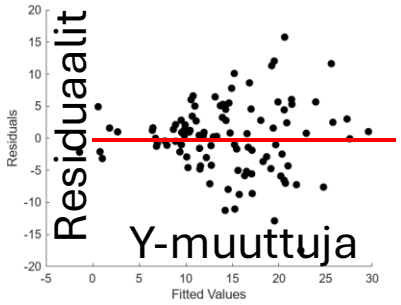
Stat. menetelmät : regressio

Lin. regr.analyysi – Residuaalien graaf. tulkintaa

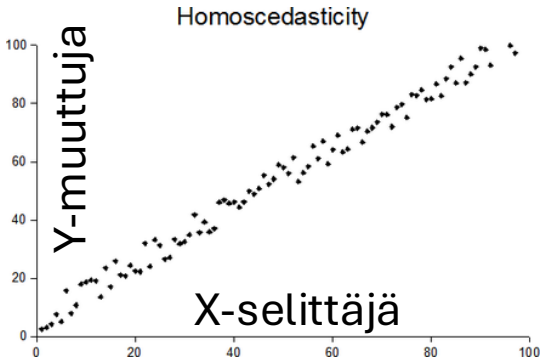


Residuaalit **normaalijak.**, tiheimmillään 0-kohdassa

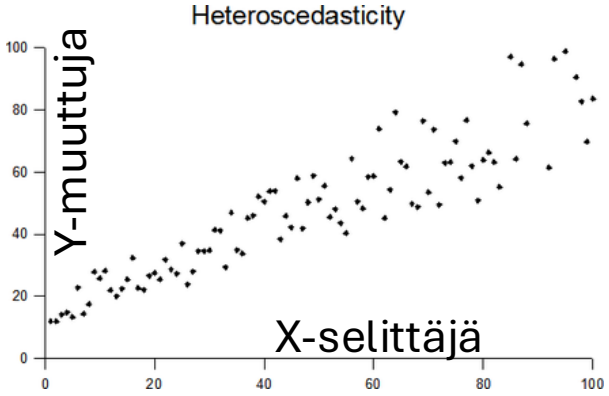
Lähde: Tietoarkisto, Treen yliopisto, sivun muut kuvat Wikipedia



Residuaalit **normaalijak.**, tiheimmillään 0-kohdassa, mutta ”pilvi” laajenee oikealle mentäessä, eli SD kasvaa eli varianssi **ei homogeeninen**



X-selittäjän ja Y-vasteen yhteydessä vaihtelu kauttaaltaan tasaista eli residuaalien SD tasainen eli **homogeeninen**



X-selittäjän ja Y-vasteen yhteydessä vaihtelu kasvaa oikealle päin eli residuaalien SD ei tasainen eli **ei homogeeninen**

Binäärinen logistinen regressioanalyysi - idea

- **Milloin** käytetään lineaarisen sijaan: **Kun vastemuuttuja 2-luokkainen** (eli dikotominen)
- Mallin melko vaikeaselkoinen idea pelkistettynä rautalankamallina:
  1. **Lineaarisessa** regressiossa laskukaava on: **y=a+bx** (tai monen x-muuttujan tapauksessa y= $b_0+b_1x_1+b_2x_2+...$), ja yhteydet lineaarisia
  2. **Logistisessa** kaavan **oikea puoli**, jossa ovat **selittävät** x-muuttujat on se ihan sama **=a+bx**, joten oikea puoli voi edelleenkin saada ihan mitä tahansa positiivisia tai negatiivisia arvoja
  3. ➔**Ongelmana**, että kaavan **vasen puoli eli vastemuuttuja** ei voikaan saada mitä tahansa arvoja, vaan 2-luokkaisena **ainoastaan 2 arvoa**, tavallisimmin 0 (=ref. luokka) ja 1
  4. ➔**Miten ratkaistu, osa 1: Vas. puolelle Y:n tilalle** vaihdetaankin **Odds** (kts slide 26!), eli *”voittaa vedossa/ei voita vedossa”* eli **p/(1-p)**, ➔ Vas. puoli voi saada jo kaikki positiiviset luvut
  5. ➔**Miten ratkaistu, osa 2**: V**as. puolelle Oddsin tilalle** vaihdetaan vielä **ln Odds** (ln= luonnollinen logaritmi), eli ➔ vasen puoli voi saada jo kaikki negatiivisetkin luvut ➔**malli toimii!!!**
- **Hintana** matemaattisesti toimivasta, mutta kompleksista mallista on, että sitä on **vaikeampi ymmärtää konkreettisesti** ja että yhteys x:n ja y:n välillä **ei ole lineaarinen**, vaan S-muotoinen
- **Tuloksia** ei **ilmoiteta** selittäjien kertoimien avulla vaan **Odds ration avulla**: ”Kun selittävän muuttujan arvo kasvaa yhdellä yksiköllä, niin Odds ratio=esim. 1,5, että vastemuuttujan arvo onkin 0 sijaan 1” (OR: kts myös slide 26!)

Binäärinen logistinen regressioanalyysi – oletukset ja suoritus

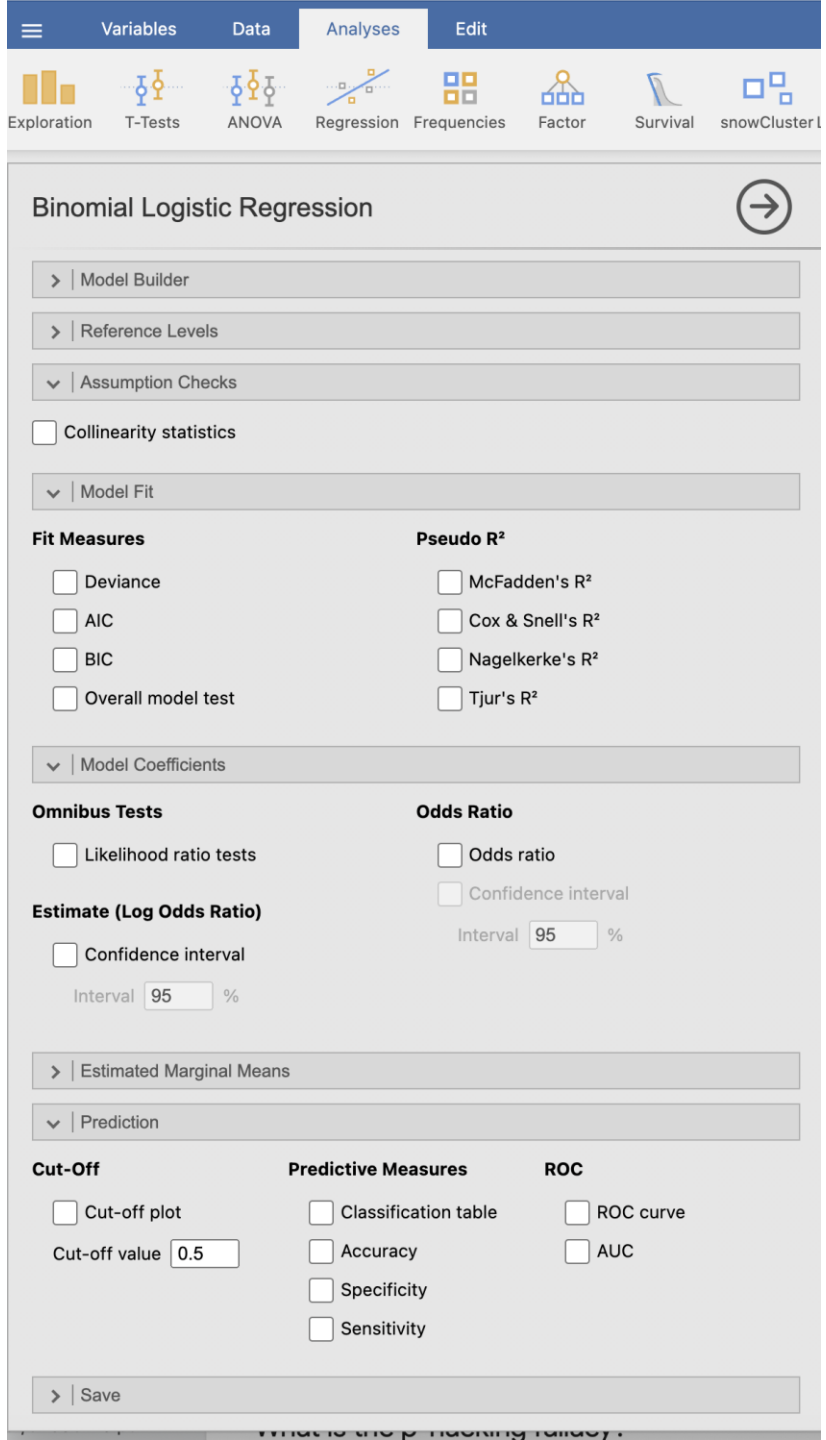
- **Oletukset**:
  - Ei merkittävää multikollineaarisuutta
  - Mielellään ei paljon poikkeavia havintoja (outliers etc)
  - Eri ryhmien välillä pitäisi olla risteäviä havaintoja, eli kaikki muuttujakombinaatiot katettu, jotta luokittelu onnistuisi hyvin (pienet aineistot tämänkin vuoksi ongelmallisia)
  - Residuaaleista oletukset kuten lineaarisessa regressiossa: normaalijakautuneet, varianssit homogeeniset, riippumattomuus toisistaan
  - Aineiston koko ≥2x se, mitä lin. regr:ssa, muuten riski n.s. ylimallinnuksesta. TAI: selittävien muuttujien lukumäärä max. 1/3 (tai max. 1/10) vastemuuttujan pienemmän luokan koosta
  - Muita oletuksia vähemmän, **EI oleteta**: selittäjien normaalisuus, yhteyden lineaarisuus
- **Suoritus**
  - **Muuttujat** syötetään samannimisiin kenttiin kuin lineaar. regressiossa, valitaan luokkamuuttujien ref-tasot
  - **Estimaatin** lisäksi pyydetään **Odds Ratio** *(joka estimaatin eksponenttifunktio)* ja sen **CI**
  - **Lisävalinnat**: oletusten tarkistukset, koko mallin P, muuttujien P:t, mallin (pseudo-)R2-testit, mallin ennustearvo (mm. sensit. ja spesif.), etc. tilanteen mukaan

Binäärisen logistisen regression valinnat

Kuvan (jamovista) yläpuolella on alue, jonne syötetään **Muuttujat** kuten lin. regressioonkin

- Dependent – 2-luokkainen vastemuuttuja
- Covariates – jatkuvat muuttujat
- Factors – luokkamuuttujat

Reference-levels-kohdasta valitaan luokkamuuttujien eri luokista referenssi, johon muita verrataan. Tämä koskee sekä selittäviä että vastemuuttujaa



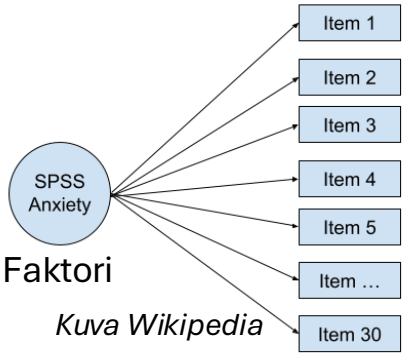
Kokoavasti regressioanalyysien käytöstä…

- **Mitä enemmän selittäjiä** (=X), sitä enemmän on myös korrelaatioita, interaktioita, monimutkaisia kausaaliketjuja eri suuntiin, vaikutuksen välittäjiä, analyysin ehtojen täyttymättömyyttä, bias'ta, mittausvirhettä, ylimallinnusriskiä…
- ➔Kannattaa **käyttää melko yksinkertaisia regressiomalleja, joissa selittäjät valitaan relevanssin**, ei askeltavan ohjelman mukaan! Voimakkaasti keskenään korreloivista mukaan vain relevantimpi. Esim. ≥8 selittäjää muuttuu helposti hallitsemattomaksi. Ei p-hacking-motiivia!
- ➔**Tuloksiin** kannattaa **suhtautua vain suuntaa-antavina;** kaikki analyysin lukuisat selittäjät tuskin ovat reaalimaailmassa täysin "mutually adjusted"!
- **Kausaliteetti on vaativaa, yleensä vain assosiaatio** (muita tilanteita: mallintaa, luokittelee, ennustaa)
- Toisinaan kannattaa **kokeilla useampaa regressiomenetelmää**, vaikka tulokset esitetäänkin vain yhdestä. Jos tulokset useammalla tavalla melko yhdenmukaisia ja R2 riittävä, se tukee tulosten luotettavuutta. Julkaisussakin voi mainita k.o. asiasta.
- **Pienillä aineistoilla suositeltavin on lineaarinen regressio** (jos ehdot täyttyvät) ja vähän selittäjiä
- **Ellei mikään regressioanalyysi mahdollista**: esim. korrelaatiot, alaryhmäanalyysit, grafiikat…

Eksploratiivinen faktorianalyysi (EFA) - yleistä

- *Tässä mahdollista esitellä EFA vain aivan yleisluontoisesti*
- EFA on matem. lisäksi paljolti substanssin pohjalta tehtäviä **vaikeita tulkintoja ja valintoja**
- EFA:ssa **pyritään löytämään havaintoyksikön ominaisuuksia kuvaavasta muuttujajoukosta piileviä yhdenmukaisuuksia eli faktoreita**, jotka pystyvät selittämään havaittujen muuttujien vaihtelua
- Tällaiset keskenään korreloivien muuttujien ”korrelaatiokimput” kuvastavat siis kukin jotakin **piilevää tekijää (=latenttia faktoria),** jotka edustavat joitakin **teoreettisia rakenteita/ konstruktioita**, joita ei voida suoraan mitata (esim. älykkyyttä voidaan mitata vain useilla verbaalisilla, spatiaalisilla, loogisilla, matemaattisilla ym. testeillä)
- Näin voidaan rakentaa esim. opiskelijoiden **osaamisen mittari (=yleisfaktori)**, joka mittaa monipuolisesti tietyn alan osaamista. Joka **osa-alueelta (=ryhmäfaktorit)** on useampia kysymyksiä/**muuttujia**, joiden vastaukset korreloivat muiden saman osa-alueen vastausten kanssa ja jkv heikommin kokonaispistemäärän kanssa
- **Konfirmatorisessa faktorianalyysissa (CFA)** *(ei käsitellä tässä)* tutkijalla on jo etukäteen teorian pohjalta muodostettu käsitys aineiston faktorirakenteesta ja analyysin tehtävänä on joko varmistaa tai kumota tämä käsitys empiirisen aineiston pohjalta.

Terminologiaa



Kuva Wikipedia

- EFA:an liittyy useita hieman hankalasti ja vaihtelevasti tulkittavia termejä
- **Faktorin lataus** (*factor loading*) **= korrelaatio.** Latauksen suuruus kertoo, kuinka paljon faktorin avulla pystytään selittämään havaitun muuttujan vaihtelusta, arvoja -1 ja 1 välillä
- **Faktorin eigenvalue/ominaisarvo =faktorin latausten neliösumma** (saadaan ottamalla kaikista faktorilatauksista neliö ja laskemalla saadut arvot yhteen)
- **Faktorin Selitysosuus** =**eigenvalue per muuttujien määrä**, 0-1
- **Muuttujan Kommunaliteetti =** kuinka hyvin faktorit selittävät yksittäisen muuttujan hajontaa **muuttujan kaikkien latausten neliösumma**, 0-1
  - **Uniqueness**: se osa variaatiota, joka on uniikkia muuttujalle ja jota faktorit eivät selitä=1-Kommunaliteetti

Oletukset/edellytykset

- Sphericity: **Bartlett's test** p<0,05
- Sampling adequacy: **the Kaiser-Meyer- Olkin (KMO) index** skaala 0-1, ⩾0,5 riittävä
- **Otoskoko**:*"100-poor sample, 300-good, 1000-excellent", Nummenmaa: ≥200, Field: ≥10-15 havaintoa/muuttujaa*

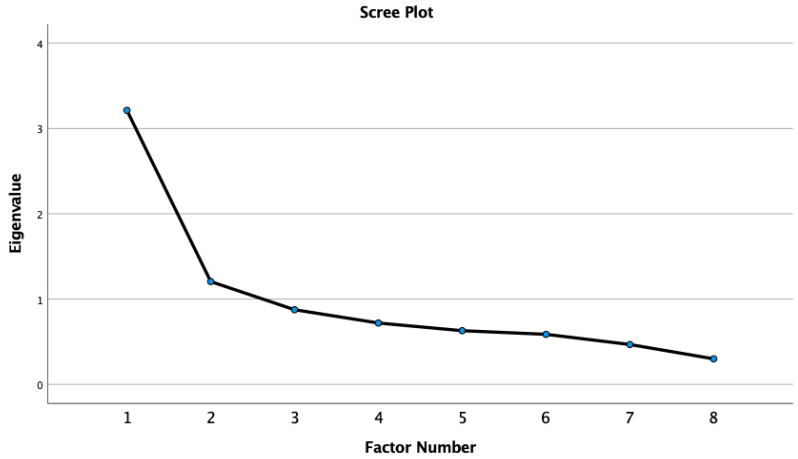
Mikä faktoriyhdistelmä valitaan??

Yleisesti: faktoreiden määrän ja yhdistelmän valintakriteerit ovat keskenään ristiriitaisia!

1. Faktorien selitysosuus mahdollisimman suuri
2. Faktoreita mahdollisimman vähän ja miel. ≥3 muuttujaa/faktori
3. Lataukset itseisarvoltaan mahdollisimman suuria ja pieniä (ei hankalasti luokiteltavia keskikokoisia mahdoll. vähän)
4. Faktoreille oltava mielekäs tulkinta

Millä kriteereillä ja faktorien LKM valinta käytännössä tehdään

1. Eigenvalue >1
2. Scree plotin (jyrkännekuvio) taitekohta
3. **parallel analysis-tekniikalla** ekstrahoituvien faktoreiden määrä (±1 tarkkuudella) (Tätä kohtalaisesti suositellaan, mutta EFA ei ole pelkistettävissä selkeiksi toimintaohjeiksi!)



Kuva Wikipedia

Muita kriteereitä, esim. muuttujien valinnoille

- Merkittävä faktorilataus on >0,3 (jos otos n. 300) tai 0,5 (jos otos n. 100)
- Merkittävä kommunaliteetti >0,3-0,5

EFA:n suoritus

- Valitaan kysymykset/**muuttujat** ja syötetään ne ohjelmaan
- Valitaan **oletusten tarkistukseen** testit
- Faktorien **ekstraktointi** = tuotetaan faktorit ja niiden lataukset,
  - Lataukset kuvaavat faktorien yhteyttä/korrelaatiota muuttujiin
  - Useita ekstrahointimenetelmiä: jamovissa ***Minimum reciduals, Principal axis*** *ja Maximum likelihood (isoihin, normaalijak. aineistoihin)*
- **Rotaatio**lla pyritään faktorilatausten rakennetta muuttamaan
  - lataukset yhteen faktoriin maksimoidaan ja muihin minimoidaan
  - Suorakulmaiset (orthogonal: *Varimax, Quartimax*) tai vinokulmaiset (oblique: ***Oblimin***, *Promax*) menetelmät
  - Yleensä oblique-rotaatiomenetelmä, etenkin jos faktorit korreloivat keskenään >0,3

Elinaika-analyysit

- Survival-analyysit **sopivat** muihinkin (kielteisiin ja myönteisiin) tapahtumiin, jos analysoidaan vain ensimmäistä/seuraavaa kertaa: sairauden uusiutuminen, muutto, työpaikan vaihto, siirtyminen laitoshoitoon, long covidista parantuminen, avioliitto
- **Seurannassa** osalle **tapahtuma**, osa poistuu seuranta-ajan jälkeen tai kesken seuranta-ajan ilman kyseistä tapahtumaa (**censored**) **(=2 ryhmää)**
- **Osa tapahtumista on komboja**: progression-free survival voi päättyä joko relapsiin tai muuhun tautiin kuolemiseen
- Jos analysoidaan erikseen esim. relapsit ja non-relapse-mortality (NRM), ideaalisesti pitäisi käyttää tarkempaa **competitive events-mallia**, jossa erotetaan toisistaan esim. poistuminen seurannasta NRM:n kautta ja muu häviäminen seurannasta, eli seurannan päättymisen **ryhmiä onkin kolme** (1. tutkittava tapahtuma, 2. muu tapahtuma ja 3. censored). Tämä vaatii R-kielisiä analyysejä, eikä sitä käsitellä tässä laajemmin, kts esim. https://www.ebmt.org/sites/default/files/2018-03/Suggestions%20on%20the%20use%20of%20statistical%20methodologies%20in%20studies%20of%20the%20EBMT.pdf

Erilaiset analyysimenetelmät

Elinaikataulu

- **Kirjataan** määrävälein (esim. vuosittain): elossa olevat jakson alussa, jakson aikana kuolleet, seurannasta pois jääneet
- **Lasketaan:** kuolemanvaarassa olevat henkilöt, kuolemanvaara jakson aikana, todennäköisyys jäädä eloon jakumulatiivinen todennäköisyys olla elossa seurantajakson alussa

Kaplan-Meierin käyrä

- **Elossaolokäyrä**, yksi käyrä tai eri alaryhmien vertailu
- Jos useampia käyriä, niiden eron P voidaan testata **Log rank-testi**llä

• Coxin regressioanalyysi

- Tämä muistuttaa muita regressioanalyysejä
- Analyysiin tarvitaan aikamuuttuja, statusmuuttuja ja selittävät muuttujat

Meta-analyysi

Yleistä

- Tässä voidaan tätä aihepiiriä sen laajuuden vuoksi **esitellä vain pintaraapaisuna!**
- Meta-analyysi on tieteellisessä tutkimuksessa käytetty menetelmä, joissa pyritään systemaattsesti keräämään ja yhdistelemään yksittäisiä kvantitatiivisia tutkimuksia
- Yleensä synteesiin kerätään RCT:itä tai muita vertailuryhmän sisältäviä kokeellisia tai observationaalisia (kohortti tai tapaus-verrokki) tutkimuksia. Pre-post-asetelmat näitä huonompia.
- Se on tieteellisten tutkimusasetelmien hierarkian huipulla, koska ne ovat yksittäisiä artikkeleita tarkempia ja niillä on suurempi voima
  - meta-analyysi >systemaatt. katsaus > kokeell.tutk. > observat. tutk.
- Niitä siteerataan muita tutkimuksia useammin johtuen mm. niiden erityisasemasta tieteellisen tiedon arvioinnista ja lääketieteessä esim. hoitosuositusten laadinnassa ym. synteeseissä.
- Se voi olla erillinen tai osana systemaattista katsausta
- Meta-analyysin kohtuullinenkin oppiminen vie merkittävästi enemmän aikaa kuin tavallisten uusien statistiikkamenetelmien opettelu. Pitää myös kerätä isohko kokoelma tarvittavia muunnoskaavoja
- Ensimmäisessä meta-analyysissä kannattaa yrittää saada kokeneempi tekijä backupiksi!

Meta-analyysin vaiheet

(ref: Muka T. et al)

- Sen alkuvaiheet ovat samanlaisia kuin systemaattisen katsauksen, jonka osa se siis voi myös olla:
  - tutkimuskysymys > hakustrategia > valintakriteerit > tietokantahaut > viitteiden ja abstr. keruu > vaiheittainen skriinaus ja tarv. tarkempi läpikäynti > artikkelien valinta > datan keruu ja muokkaus > tutkimusten laadun arviointi > tiedoston luonti > deskript. synteesi ja vuokaavio >
- Tässä vaiheessa meta-analyysin lisäpalikat
  - > kvantitat. synteesi/meta-analyysin teko > heterogeenisyyden ja julkaisuharhan analysointi < näytön laadun arviointi
- (ja lopuksi tietenkin vielä: < raportointi)

Meta-analyysin tekninen toteutus

- **Fixed-effect vai random-effects**-meta-analyysi?
  - Random-effects-mallissa oletetaan, että eri tutkimukset arvioivat hieman erilaisia ja eri kokoisia efektejä, fixed-effectissä efekti oletetaan kaikilla samaksi
- **Syötettävät tiedot** poimittava määrämuodossa (tai transformoida sellaisiksi)
  - Tämä usein hyvin työläs vaihe: tulosmuuttujia pitää ehkä transformoida yhtenäiseen formaattiin. Lisäksi niitä pitää usein yhdistellä tai poimia selektiivisesti vain osa.

  A. **Dikotomiset tulosmuuttujat** riskisuhde, riskiero, odds ratio
     - Muuttujina tapahtumien määrä ja otoskoko koe- ja kontrolliryhmissä

  B. **Jatkuvat tulosmuuttujat** keskiarvojen ero, standardisoitu keskiarvojen ero
     - Muuttujina keskiarvo SD ja otoskoko koe- ja kontrolliryhmissä

  C. **Efektikoot –** jos sekä A että B-ryhmän tulosmuuttujia, voidaan A-ryhmä transformoida ja esittää kaikki
     - Muuttujina Cohenin efektikoko ja keskivirhe (tai varianssi)
- **Jamovi-kirjaston Major-moduli** – alkuvaikeuksien jälkeen helppokäyttöinen, muita mm. STATA tai R-kielinen ohjelmointi

Meta-analyysin tulokset

- **Forest plot**: yksittäisten tutkimusten efektikoko ja sen CI sekä yhdistetty efekti ja sen CI
- **Funnel plot** ja testit, joilla arvioidaan julkaisuharhaa
- **Heterogeenisyys**
  - Kaikkinainen vaihtelu mukaan otettujen tutkimusten välillä
  - (1) Onko sitä?: tau^2, the Q-test for heterogeneity, (2) paljonko?: the I^2 statistic
- Poikkeavat arvot (influence etc)
- **Metaregressiot** tai k.o. muuttujan mukaiset **alaryhmäanalyysit**

Meta-analyysin kirjallisuutta ja linkkejä mm.

- Cochrane Handbook for Systematic Reviews of Interventions, etenkin Part 2: 6. Effect measures 10. Meta-analyses https://training.cochrane.org/handbook/current
- Jouko Miettunen, Presentations (löytyy 2 esitelmää): https://www.joukomiettunen.net/presentations/
- Muka T et al. A 24-step guide on how to design, conduct, and successfully publish a systematic review and meta-analysis in medical research. European Journal of Epidemiology (2020) 35:49–60
- MAJOR: Meta Analysis JamOvi R https://github.com/kylehamilton/MAJOR
- PRISMA 2020 checklist https://www.prisma-statement.org/prisma-2020-checklist
- Lisäksi R-ohjelmapaketeilla (mm. meta, metafor) omat sivunsa

TAVALLISIMMAT SYYT JAMOVIN MAJOR-PAKETILLA, KUN OHJELMA ITSEPINTAISESTI EI SUOSTU TOIMIMAAN

- Tyhjiä soluja
- Siirretyn csv-tiedoston tms. lopussa onkin vielä yksi tyhjä rivi

The general linear model (GLM)
The generalized linear model
The mixed linear model
The generalized mixed model

Ryhmä monipuolisia malleja, jotka ”taipuvat” hyvin monenlaisiin analyyseihin

Mallien luokittelu

Selittävät muuttujat	Riippuva muuttuja: Normaalijak. & jatkuva	Riippuva muuttuja: Ei-normaalijak. tai luokkamuuttuja
Riippumattomia havaintoja	General Linear Model (GML)	Generalized Linear Models
Ryhmittyneitä havaintoja	Mixed Model	Generalized Mixed Models

- Ryhmä monipuolisia malleja, joissa analyysejä ja tulosten raportointia on systematisoitu
  - **GLM** korvaa lineaarisen regression, ANOVAn ja kovarianssianalyysin (ANCOVA)
  - **Generalized Linear Model**: Riippuva muuttuja muutetaan linkkifunktion avulla jatkuvaksi ja huomioidaan sen poikkeava jakautuminen, saadaan mallit: binomiaal. ja multinomiaal. logistinen, Poisson, ordinaalinen ym. mallit
  - **Mixed models**: muuttujat eivät ole täysin riippumattomia toisistaan, jos ne kuuluvat johonkin useista ryhmistä (**multilevel** models, esim. koulujen oppilaat), jolloin samaan ryhmään kuuluvat muistuttavat enemmän toisiaan ja tavanomaiset testit antavat jkv virheellisiä tuloksia. Aikasarjamittauksissa on vastaava riippuvuus
- Kts tarkemmin esim. https://gamlj.github.io/book/gzlm.html joka samalla jamovin näissä analyyseissä käytettävien **gamlj** ja **GAMLj3**-lisämodulien opas

Moni-imputaatio puuttuvien arvojen täyttämisessä

Yleistä moni-imputaatiosta

(MI, multiple imputation)

- **Puuttuvan tiedon korvaus vain yhdellä arvolla ongelmallista**
  - SD ja siten SE (keskivirhe) liian pieni, jolloin tilastolliset päättelyt ”liian hyviä”
  - Mean-imp:ssa korvaus tietyssä sarakkeessa/rivissä kiinteä, ei ennustukseen perustuva
  - Multiple imputation välttää e.m. ongelmat
- **MI-menetelmän ja prosessin kuvaus**
  1. **Luodaan** lähtödatasta **useita** (oletusarvo =5, mutta tarv. esim. 20 tai vaikkapa 100) **imputoituja täydellisen datan versioita**, joissa puuttuvien arvojen korvaus perustuu matemaattiseen ennusteeseen, johon lisätty vielä soveltuvaa satunnaisvaihtelua. **Iteraatioiden** (toistoja) oletusarvo =5, mutta kunnes ns. konvergenssi saavutetaan, tarv. vaikkapa esim. 20-40
  2. **Jokaisesta luodusta tiedostosta analysoidaan kiinnostavat parametrit** (deskriptiivisesti tai esim. regressioanalyysistä). Näiden tulokset eroavat, riippuen imputointeja koskevan epävarmuuden määrästä
  3. **Tulokset yhdistetään** (”pooling”). Tällöin lähtödatan vaihtelun lisäksi huomioidaan imputointeihin liittyvän epävarmuuden määrä. Tälläin MCAR ja MAR-datassa ei biasta ja tilastolliset ominaisuudet asianmukaisia
- MI:n voi suorittaa R-ohjelmalla (mice-paketti) tai myös. Stata, SAS, SPSS (lisäpalikkana)

Muuta MI:n suorittamisesta

- **Predictor matrix** on taulukko, jossa ilmoitetaan, mikä muuttuja ennustaa mitäkin muuttujaa
  - Ilmoitetaan myös esim., mikä luku on matemaattisesti täydellisesti määritelty toisella (esim. mittarin score=sen kysymysten pisteiden summa)
- **Puuttuvien imputointimetodit.** Oletus esim. jatkuvat muuttujat: PMM=Predictive mean matching ja dikotomiset: logistic regr.
  - Näitä voi tarv. muuttaa. Myös ns. **post-processing,** jossa rajataan imputoitujen arvojen arvoja pelkästään k.o. muuttujalle mahdollisiin arvoihin
- Imputoitujen muuttujien mean ja SD **konvergenssikäyrien visuaalinen tarkastelu**: Tavoitteena käyrien ”kietoutuminen” toisiinsa, ilman trendejä myöhemmissä iteraatioissa.
- Menetelmän käyttö R-kielellä edellyttää kielen perusteiden osaamista ja vaatii jkv enemmän opiskelua kuin statistiikan perusmetodit. Suositeltavia lähteitä:
  - Mice README:ssa vignetit 1-4 https://cran.r-project.org/web/packages/mice/readme/README.html
  - Valittuja kohtia Vanbuuren: Flexible imputation of missing data https://stefvanbuuren.name/fimd/

6. Apua netistä

- Netin laskurit
- Netin ohjeet ym. tietolähteet

Nettilaskurit - yleistä

- Pääsääntöisesti pitää tietenkin käsitellä ja analysoida aineisto kokonaan itse (=tutkimusryhmä tai sen käyttämä statistikko)
- Joskus tarvitaan simppeli lisäanalyysi valmiiksi ristitaulukoiduista luvuista, vaihtoehtona ehkä työläs datan muokkaus analyysiin
- Nettiin ei tietenkään saa syöttää henkilöitä identifioitavaa dataa!
- Esim. seuraavissa tilanteissa käyttökelpoinen:
  - **Otoskokolaskenta** t-testiasetelmassa, kokeilu eri arvoilla (*slide 9-10*)
  - **Ristiintaulukoiduista** luvuista **P** (χ^2) riippumattomuus- tai Fisherin testillä
  - **Luottamusvälit** esim. osuudelle tai prosentille

Nettilaskurit – käytännön esim.

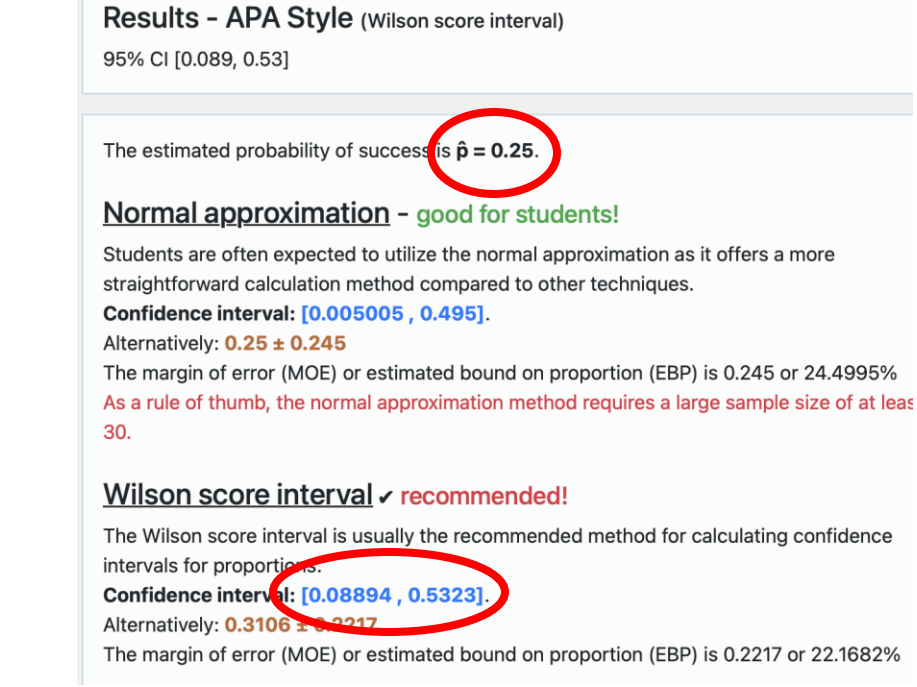
Statistics Kingdom https://www.statskingdom.com/index.html

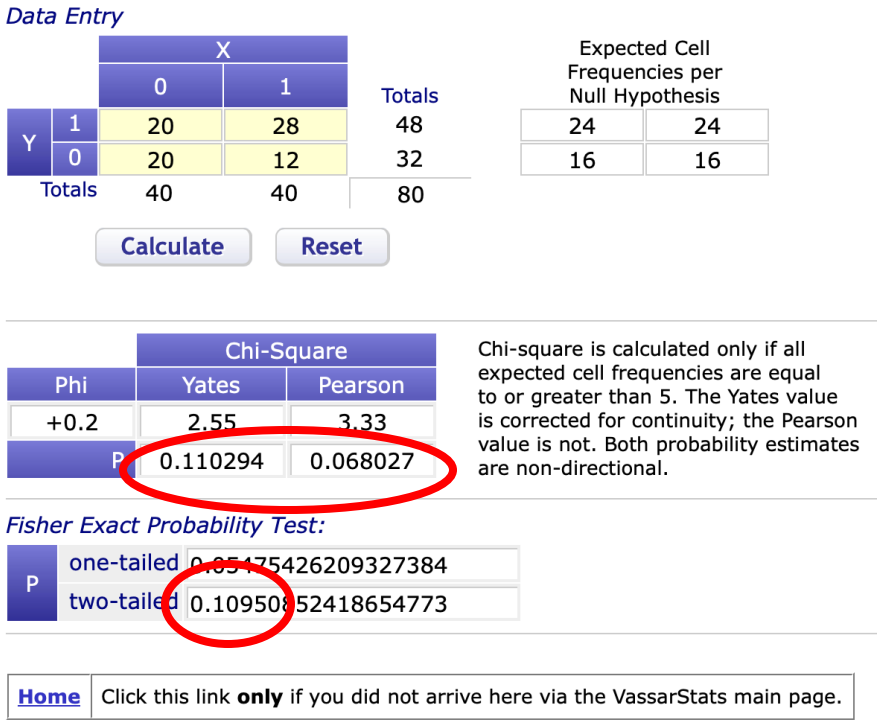
- Osuuden/prosentin CI: https://www.statskingdom.com/proportion-confidence-interval-calculator.html
  - Esim. mikä osuus ja 95%CI (ja %:t) 3 oikeaa on 12:sta? *(laskurissa juuri nämä esim. arvot, käytetään niitä!)*
  - Painetaan Calculate, katsotaan alempaa
  - Saadaan: 0,250 (95%CI 0,089-0,532), eli 25,0% (8,9-53,2%)

VassarStats http://vassarstats.net/index.html

- Chi-Square Test of Association (= Independence) reittiä:
  Frequency data>For a 2x2 Table of Cross-Categorized Frequency Data>Version 1
- Keltaisiin ruutuihin syötetty luvut ja painettu Calculate, saadaan Chi-Square Yatesin ja Pearsonin arvoina ja alinna 2-puolinen Fisher (joka periaatteessa aina tarkin!)

Apua netistä





Hyviä statistiikan laaja-alaisia nettisivuja

- **Kvantitatiivisen tutkimuksen verkkokäsikirja (Tietoarkisto)** https://www.fsd.tuni.fi/fi/palvelut/menetelmaopetus/kvanti/
  - Tutkijoille suunnattu asiallisen hyvä opas
- **Learning statistics with jamovi (Navarro&Foxcroft)** https://blog.jamovi.org/2018/10/25/learning-statistics-with-jamovi.html
  - Erinomainen jamovin ja yleisen statistiikan opas. Nettiversiokin on, mutta täältä saanee ladattua PC:lle!
- **Tilastokunto (Aleksi Reito)** https://www.tilastokunto.fi
  - Suomalaisen ortopedian dosentin monipuoliset sivut
- **Statistics notes (The BMJ)** https://www.bmj.com/specialties/statistics-notes
  - Monipuolinen artikkelikokoelma sivustolla, ei kylläkään systemaattisesti ryhmiteltynä
- **Presentations- Jouko Miettunen** http://www.joukomiettunen.net/presentations/
  - Oululaisen proffan iso setti hyviä, selkeitä esityksiä
- **Sarna: Kliinisen biostatistiikan peruskurssi** (2011) https://www.mv.helsinki.fi/home/sarna/Opetus/Moniste%20Osa%201
- **Sarna: Kliinisen biostatistiikan jatkokurssi** (2012) https://www.mv.helsinki.fi/home/sarna/Opetus/Moniste%20Osa%202
- **SPTY:n tutkimus/Statistiikan linkit** https://www.spty.fi/tutkijoille/statistiikan-linkkeja/
  - Olen nämä linkit tuottanut, sivulla on myös statist. osa-alueiden linkkejä ja laatimani Excel-koulutusdokkari

7. Raportointi

Raportointi

Menetelmät

- Sillä tarkkuudella ja detaljitasolla, että **osaava lukija pystyy arvioimaan** niiden asianmukaisuuden
- **Muuttujien luokkien yhdistely** ym. muokkaukset analyysejä varten, tarvittaessa perusteluineen

Tulokset

- **Tarkasti lehden ohjeiden mukaan** merkinnät (esim. n, %, desimaalien määrä, CI:n esitysmuoto, tarkka P tai *** etc.)
- Metodiosassa kuvattu kaikki metodit, joiden tulokset esitetty tulososassa

Diskussio

- **Eksaktit termit**, ei *correlates*, ellei tehty korrelaatioita.
- *Significant* vain statistisesti signifikantista
- ***”Explains, causes”*, vain jos osoitettu kausaaliyhteys,** *muuten nöyrästi vain: ”**association**, connection, relation, correlation”!*