Kvantitativa metoders användning i vetenskaplig forskning

Auvo Rauhala, ML, FT, doc.

Uppdaterad: 14.10.2025

SPTY, Klient- och patientsäkerhetscentret, Åbo Akademi

Den senaste uppdaterade versionen av denna guide finns alltid på webbplatsen:
https://www.spty.fi/tutkijoille/metodologian-linkkeja/

Innehållsförteckning (som länkar)

1. Inledning

Allmänt om teoriföreläsningar och handledningsdokumentet

- Denna guide presenterar de viktigaste **grunderna** och de vanligaste **praktiska problemen** på ett ”**kokboksaktigt**” sätt, men metodernas ”idé” förklaras också (underlättar förståelsen).
- Många metoder har inte fått plats här.
- Guiden är dock utformad för att även passa för **självstudier** och som en **handbok** för informationssökning, och innehåller även **länkar** till mer djupgående material.
- Kort presentation av **onlinekalkylatorer** som kan vara till hjälp som tillägg.
- Inom statistik finns **olika synsätt på tillvägagångssätt**, särskilt när variabler och fördelningar inte är ideala och datamängden är liten – och metodens villkor endast delvis uppfylls.
- **Vid problem ger googling på engelska med så exakt text som möjligt nästan alltid hjälp från nätet** – ibland först efter det tionde knepet du provar!
- Om bildkällan inte anges, är bilderna antingen egenproducerade eller från Wikipedia.
- **Om du upptäcker brister, fel eller andra behov av redigering i texten, meddela mig gärna!** T.ex. auvo.rauhala(at)gmail.com

Sammanfattning av olika steg i hantering av kvantitativt material

Datainsamling

- **Beräkning av urvalsstorlek**
- **Formulärdesign**

Datagranskning och bearbetning

- **Variabler = kolumner**: skala (t.ex. Likert ⇒ intervall eller ordinal?), normalfördelad?
- **”Cases”=statistiska enheter= rader**
- **Observationer = celler**: saknade ➔imputeringar?), felaktiga, avvikande, flera i samma cell
- **Filstrukturens omformning**: format long ↔ wide, case vs. control, flernivå (t.ex. enhet–patient–skadeställen)
- **Bearbetning av variabler**: olika summavariabler, omkodade kategorier
- **Excel-funktioner** sparar mycket tid vid testning och bearbetning!
- **Även i SPSS, jamovi m.fl. kan man bearbeta data**

Val och genomförande av analys med tilläggstester

- **Endast deskriptiv**
- **Test av skillnader**
  - Kategorivariabler (korsklassning): Chi Square, Fisher
  - Intervall, oparade
    - Normalfördelade: t-test (2 variabler), ANOVA (≥3) + post-hoc
    - Ej normalfördelade (eller ordinala): M-W U-test (2), Kruskal-W (≥3)
    - Parade: egna t-test, istället för M-W U-test → Wilcoxon
- **Samvariation** mellan två variabler: korrelation (vilken typ? t.ex. Pearson, Spearman, polychoric, point-biserial)
- **Förklaring av output-/responsvariabel** med andra (förklarande variabler): univariat/multivariat regressionsanalys
  - Vanligast: linjär och (binär/multinom.) logistisk regression
  - Analysen ”förklarar” endast matematiskt – i observationsstudier handlar det snarare om ”association”
- **Explorativ faktoranalys** (EFA): söker latenta faktorer som förklarar t.ex. flera items i ett mätinstrument
- **Andra analyser**:
  - Konfidensintervall, effektstorlekar, validitets- och reliabilitetstestning
  - Överensstämmelse mellan parallella klassificeringar (kappa m.m.)
  - Mätinstrumentets prediktiva förmåga: ROC, Brier-score, goodness of fit
  - Testning av analysvillkor
- Delvis analysberoende, delvis **generella**: t.ex. urvalsstorlek, variabelskala, normalitet

Rapportering

- Deskriptiv tabell
- Korrelogram
- Multivariat tabell
- Grafik
- Textdel
- Kausalitet utanför RCT och kvasi-experimentella datamaterial är sällsynt!

Steg i användningen av kvantitativa metoder

1. **Operationalisering** av teoretiska begrepp till mätbara egenskaper
2. **Val av variabler**: beroende (output, response), oberoende (explanatory), samt andra att beakta (t.ex. kontrollerade i multivariatanalyser)
3. **Skalnivåer för variabler** (nominal–ordinal–antal–intervall…) och kategorier för klassvariabler
4. Forsknings**design**
5. **Urval** (typ, t.ex. slumpmässigt urval, samt urvalsstorlek)
6. **Datainsamling** (t.ex. enkätformulär)
7. Data**granskning** och **bearbetning**
8. **Dataanalys**
9. **Rapportering** av data

❖ Det är viktigt att **planera** datainsamling och metodologi **i god tid** för att undvika senare problem som inte går att åtgärda.

❖ **Systematisk** namngivning och lagring av mellanliggande dokumentfiler samt analysprotokoll och hjälpdokument är också viktigt!

Variabeltyper / Skalnivåer

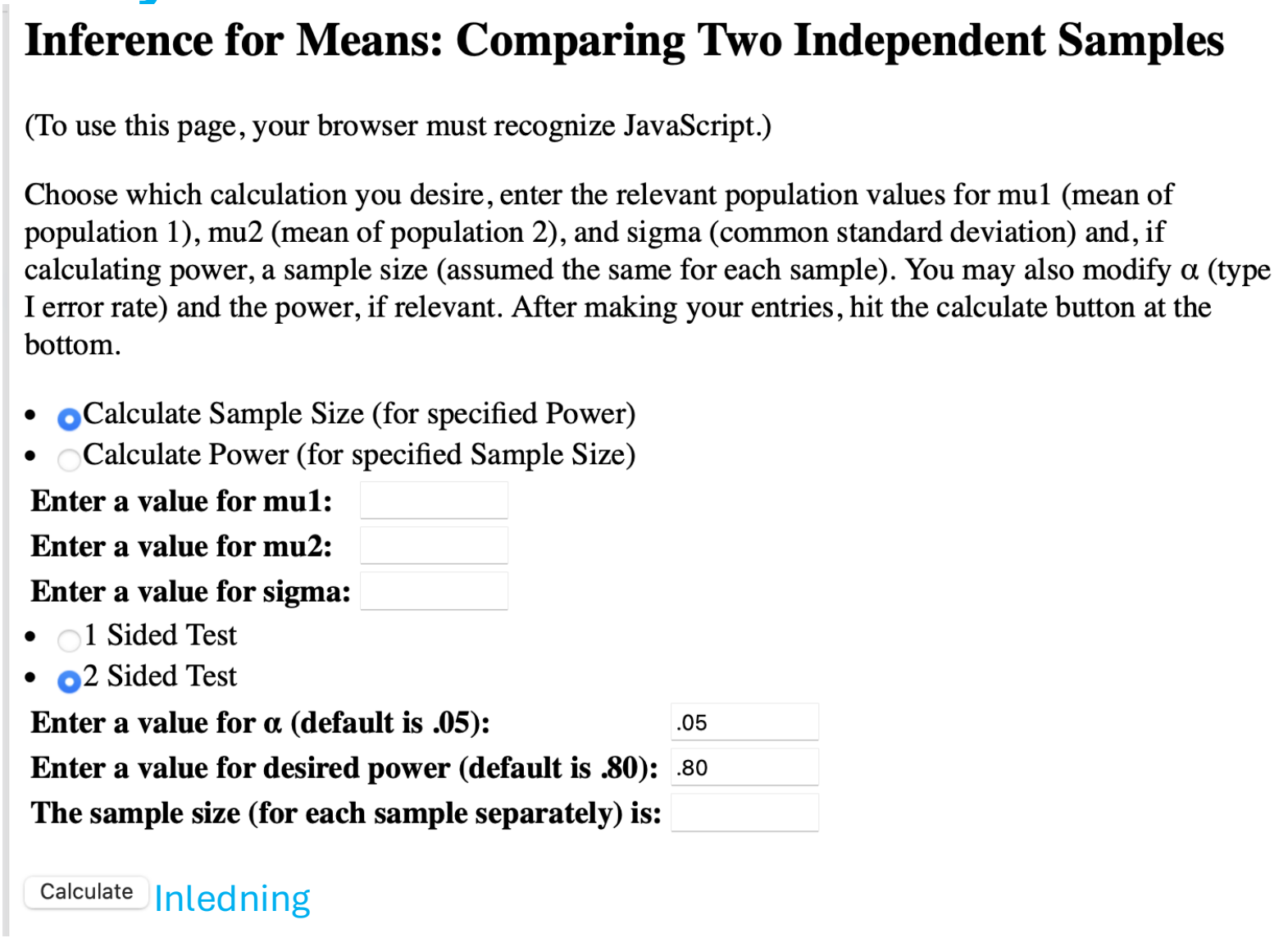
- **Skalnivåer / skalor**
  - **Nominalskala** (t.ex. kön, civilstånd) – ingen meningsfull ordning mellan kategorier. Särfall: dikotom (2 kategorier)
  - **Ordinalskala** (t.ex. Likert, PAPA, riskklass) – kategorierna har en ordning
  - **Intervallskala** – avstånd utan absolut nollpunkt (t.ex. Celsius)
  - **Kvotskala** – med absolut nollpunkt (t.ex. vikt, Kelvin)
- **Kontinuerliga** (exakt ålder), **diskreta** (t.ex. antal)
- Skalproblem och gränsfall är vanliga i datamaterial
- Används även termer som kvalitativa (nivå 1) och kvantitativa (nivå 3–4) variabler, samt icke-numeriska (nivå 1) och numeriska (nivå 3–4). Ordinala variabler kan vara otydliga i vissa källor
- En kvantitativ variabel kan uttrycka kvalitet (t.ex. antal hål i en strumpa), och kvalitativa kategorier kan kodas numeriskt

Särfall / problem med skalnivåer

- **Likert**: Enskilda ordinala Likert/Likert-typ variabler: om ≥5–7 numeriska kategorier (och fenomenet är kontinuerligt), kan behandlas som intervallvariabel – men åsikterna varierar!
  Se t.ex. https://davidfoxcroft.github.io/lsj-book/ (Sök: Likert)
- **Dikotoma variabler**: använd ”dummy”-koder, 0 (=referensnivå) och 1 (=undersökt nivå), särskilt vid korrelations- eller regressionsanalys
- **Felaktig användning av nominalskala**: t.ex. civilstånd kodat som 1=ogift, 2=gift, 3=skild, 4=änka – om detta används som en enda variabel i analysen, antas t.ex. att ”en änka motsvarar två gifta”!
- **Att dela upp intervallvariabler i kategorier** bör undvikas i statistisk inferens utan goda skäl – kan t.o.m. misstänkas som p-hacking (jakt på signifikanta p-värden) och minskar analyskraft. Deskriptivt kan man förstås visa data i kategorier.

Beräkning av urvalsstorlek I/II – exempel med t-test och onlinekalkylator

- Rekommenderad kalkylator: https://www.stat.ubc.ca/~rollin/stats/ssize/n2.html
- Urvalsstorleken bör baseras på huvudresultatets krav och den testmetod som används
- I praktiken används ofta t-test som grund, som i vårt exempel



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

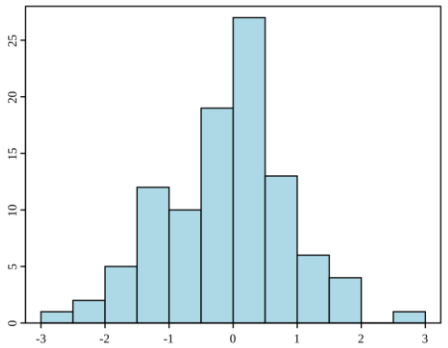
Beräkning av urvalsstorlek II/II – instruktioner för kalkylatorn

- Använd **decimalpunkt**!
- Välj analys: **sample size** (eller **power**)
- Ange **medelvärden** för de två grupperna i fälten **mu1** och **mu2**. Själva medelvärdena är inte viktiga – kalkylatorn använder endast **skillnaden** mellan dem
- **Sigma** = **SD**. Om grupperna har olika SD, använd medelvärdet av dem
- **p-värde** och **power** är förinställda till vanliga värden, men prova gärna med p < 0.01
- Tester är i regel **tvåsidiga**
- Klicka på **CALCULATE** för att få urvalsstorlek för båda grupperna
- Om du vill ha en viss **effektstorlek (Cohen's d)** i t-testet, använd formeln: **Effektstorlek = skillnad i medelvärden / SD**, där en vanlig måttlig effektstorlek är **0,5**

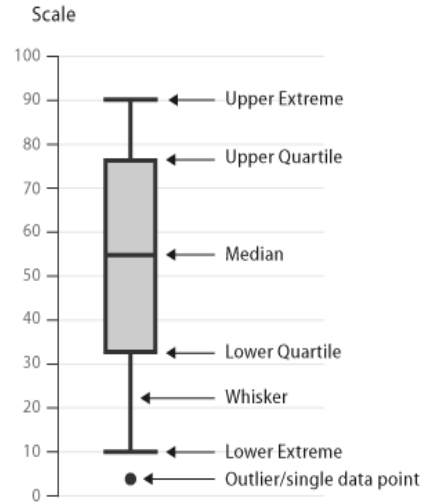
2. Granskning och bearbetning av datamaterialet

Av den tid som går åt till att arbeta med datamaterialet kan upp till 90 % gå åt till granskning och bearbetning, medan själva analyserna utgör endast ≥10 %.

Allmänt – vad och hur granskas



- **Grafiskt**: Histogram och låd-diagram för intervallvariabler, stapeldiagram för kategorivariabler
  Tabeller: Visuell granskning och med hjälp av nyckeltal



Kuva Wikipedia

1. **Struktur**: kolumner/variabler, rader/"fall" och celler
2. **Saknade värden, avvikande värden, nyckeltal?**
3. **Nyckeltal**: lägesmått (*medelvärde, median, typvärde*) och spridningsmått (*standardavvikelse, kvartiler, percentiler, min, max, variationsbredd, kvartilavstånd*)
   Beräknas med statistikprogram (SPSS, jamovi) eller oftast med Excel
4. **Samband mellan variabler?**: spridningsdiagram, korrelationer
5. **Fördelningar?**: normalfördelning, sned, spetsig, tvåtoppad?

Datamaterialets struktur

- Från **"messy data" → till "tidy data"**
- ***Tidy data*** = variabler i egna kolumner, statistiska enheter (=fall, observationer) i egna rader, värden i egna celler
- **Strukturförändringar**
  - En variabels **kategorier** kan ha egna kolumner (t.ex. vikt för kvinnor och män i separata kolumner → ny kolumn ”kön” och en gemensam kolumn ”vikt”)
  - **Enheterna för variablerna** kan vara inkonsekventa → korrigeras
  - Flera värden i en cell → endast ett värde (övriga tas bort eller flyttas till ny kolumn)
    - Undantag: i fall-kontrollstudier både fall och kontroll på samma rad; tidsserier
  - **Multinivådata** är svåra, hierarkiska variabler t.ex. enhet > patient > skadeställe
  - Variabler bör ha en tillförlitlig **identifieringsnyckel**
  - Ibland lönar det sig att spara t.ex. **ursprunglig radordning** som en egen variabel, så att den kan återställas vid behov

Saknade värden

- **gör allt du kan** för att undvika saknade värden!!!

- **Antalet och typen av saknade data undersöks och rapporteras** också (*Checklistor Strobe, Consort*).
- **Slumpmässigt** eller **systematiskt** saknad data i variabler - - t.ex. Little's MCAR-test i R eller SPSS
  - **MCAR** (missing completely at random): Sannolikheten för saknade data är inte relaterad till värdena på andra observerade (eller oobserverade) variabler eller till värdena på själva variabeln saknade data.
  - **MAR** (missing at random): Sannolikheten för saknade data är densamma endast inom de grupper/kategorier som definieras av de observerade data (andra variabler). MAR är en mycket bredare kategori än MCAR.
  - **MNAR** (missing not at random): Sannolikheten för saknade data varierar av skäl som är okända för oss, eftersom sannolikheten för saknade data är systematiskt relaterad till saknade hypotetiska värden.
- **Jfr**: om **svarsfrekvensen är låg** finns det många fall av total non-response ➔ **non-response bias**, dvs. participation bias.

Saknade värden

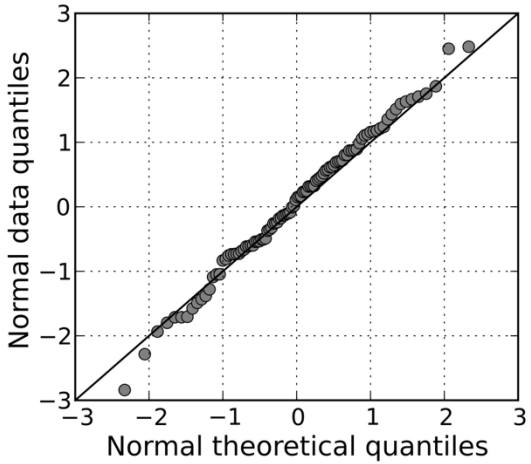
- **gör allt du kan** för att undvika saknade värden!!!

- **Konsekvenser av saknade värden** (såvida de inte är mycket få)
  - Om saknade värden är helt slumpmässiga (***MCAR***), förlorar utelämnande av ofullständiga fall (=list-/fallsvis borttagning) eller endast saknade värden (parvis borttagning) från analyserna helt enkelt sin styrka.
  - I andra (***MAR*** och ***MNAR***) uppstår även bias och situationen är ännu värre i multivariata analyser.
- **Ersättning av saknade värden, dvs. imputation**
  - **Medelvärdesimputation**: ersättning med rad- eller kolumnmedelvärde - enkelt och använt, men inte speciellt rekommenderat i statistisk litteratur!
  - **[+/- Stokastisk] regressionsimputation**: ersättning med regressionsekvationsprediktion [+/- slumpmässig felterm] - bättre, men både i medelvärde och regr. imp. Blir SD för litet➔signifikanser och konfidensintervall falskt bra
  - Bästa är **multipel imputation** (använd om MCAR eller MAR), **se slide 72 för mer information om MI!**
- Mer information t.ex.: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/

Avvikande värden

- Avvikande värden upptäcks grafiskt eller med olika tester
  1. **Outliers**: avvikande värde i y-led
  2. **(High) leverage**
  3. **(High) influence**, när båda ovanstående förekommer
     - Påverkar t.ex. regressionslinjen i linjär regression
     - **Cook's distance** >1 = stor påverkan. Om man använder ett enda mått, så detta!
     - Man kan också undersöka hur mycket analysresultatet förändras om värdet tas bort
  - Vad göra?
    - Huvudregel: **felaktiga?** (=tas bort/korrigeras), **verkliga?** (=behålls)

Normalfördelningsantagandet



- **Antagandet om normalfördelning är centralt**, eftersom det möjliggör användning av effektivare så kallade parametriska tester
  - Dessa använder t.ex. medelvärde och standardavvikelse som parametrar (t-test, Pearsons korrelation, ANOVA m.fl.)

Bild Wikipedia

- **När är en fördelning (tillräckligt) normalfördelad?** (ref. Lauri Nummenmaa: *Tilastotieteen käsikirja*)
  1. Om p > 0,05 i ett normalitetstest (Shapiro-Wilk, Kolmogorov-Smirnov), anses materialet vara normalfördelat (förutsatt att det inte är ett mycket litet material, några tiotal). Testerna är dock bristfälliga: i stora material alltför känsliga, i små alltför okänsliga för avvikelser från normalfördelning
  2. **ELLER**: Om testet visar att fördelningen inte är normal, räcker det som bevis för normalitet att den visuellt ser normal ut **OCH** att skevhet och kurtosis ligger ungefär mellan -1 och +1
  - **Visuell granskning**: t.ex. histogram och/eller QQ-plot (kvantil–kvantil-diagram) – BILD!
  - Ju större material, desto mer berättigat är antagandet om normalfördelning (centrala gränsvärdessatsen)
  - **Nedre gräns för antagande om normalitet: n ≈ (20–)30**
- Vid **sned**fördelningar kan man ibland försöka **transformera** ($\sqrt{}$, log, 1/x osv.) för att uppnå normalitet – men detta kan leda till nya problem

Variabelanalys och bearbetning i Excel

- Dessa analyser kan göras i statistikprogram
- **Excel är i praktiken enklare och mer mångsidigt för olika kontroller och bearbetningar**
- Det lönar sig att bekanta sig med **Excels vanligaste funktioner** – de kan spara veckor av arbete i stora dataset!
- På SPTY:s webbplats: https://www.spty.fi/tutkijoille/statistiikan-linkkeja/ finns min fil *Tutkijan Excel-manööverien basicit_v.211023.xlsx*
- Vanliga funktioner:
  - SUMMA(), MEDEL(), MAX(), MIN(), OM(), MEDIAN(), ANTAL.OM(), ANTALV() (räknar celler med innehåll), ANTAL.TOMMA()
  - T.ex.: ”kompositvariabel”, som är summan eller medelvärdet av flera variabler
- Andra användbara: **makron**, **kopiering av funktioner** sparar mycket arbete

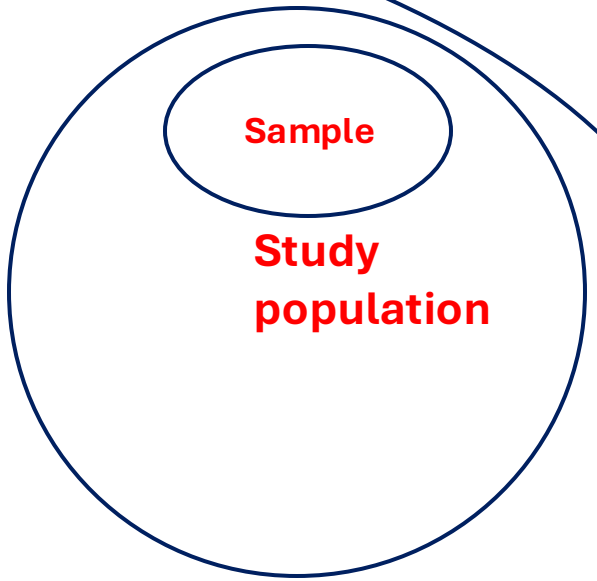
3. Allmänt om statistiska analyser

Statistik – vad och varför?

- En **metodvetenskap:** hur man från observationer drar giltiga slutsatser om fenomen inom ämnesvetenskapen, med känd tillförlitlighet (=P) och konfidensintervall (=CI)
- Använder **följande metoder** (samma test kan tillhöra flera grupper):
  1. **Deskriptiv statistik**: nyckeltal (medelvärde, standardavvikelse m.m.), grafik
  2. **Statistisk inferens, (statistical inference)**, att dra slutsatser
     - nollhypotes (H0)-prövning, P, CI), kan urvalets resultat generaliseras till målpopulationen? (t-test, ANOVA)
  3. **Analys av samvariation** mellan 2 variabler (korrelation, faktoranalys)
  4. **Prediktion och modellering** (regressionsanalyser m.m.)
  5. **Klassificering och gruppering** (logistisk regression, diskriminant- och klusteranalys)

Sample – Study population – Target population

Urval/sample: den del av studiepopulationen som undersöks, testas och analyseras
Studiepopulation/study population: den grupp till vilken resultaten realistiskt kan generaliseras
Målpopulation/target population: den bredare grupp till vilken man idealt vill tillämpa resultaten, även om den ofta är för stor för att fungera som urvalsram

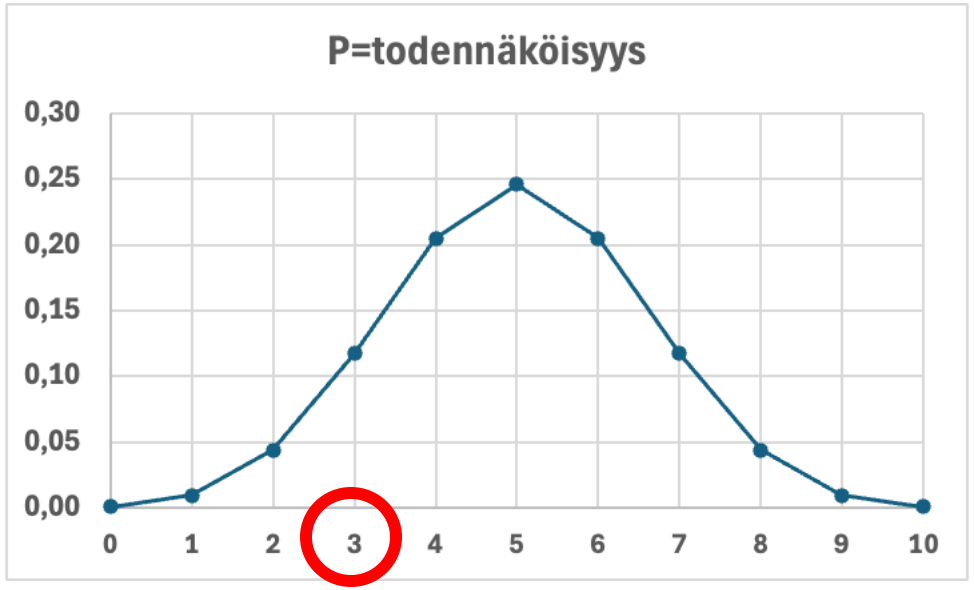


Target population

Sannolikhetsfördelningar är statistikens kärna!

- När man kastar en slant eller tärning kan sannolikheten för t.ex. 3 kronor på 10 kast beräknas exakt (binomialfördelning)
- En **sannolikhetsfördelning** visar dessa sannolikheter **direkt**

Kronor	P=sannolikhet
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

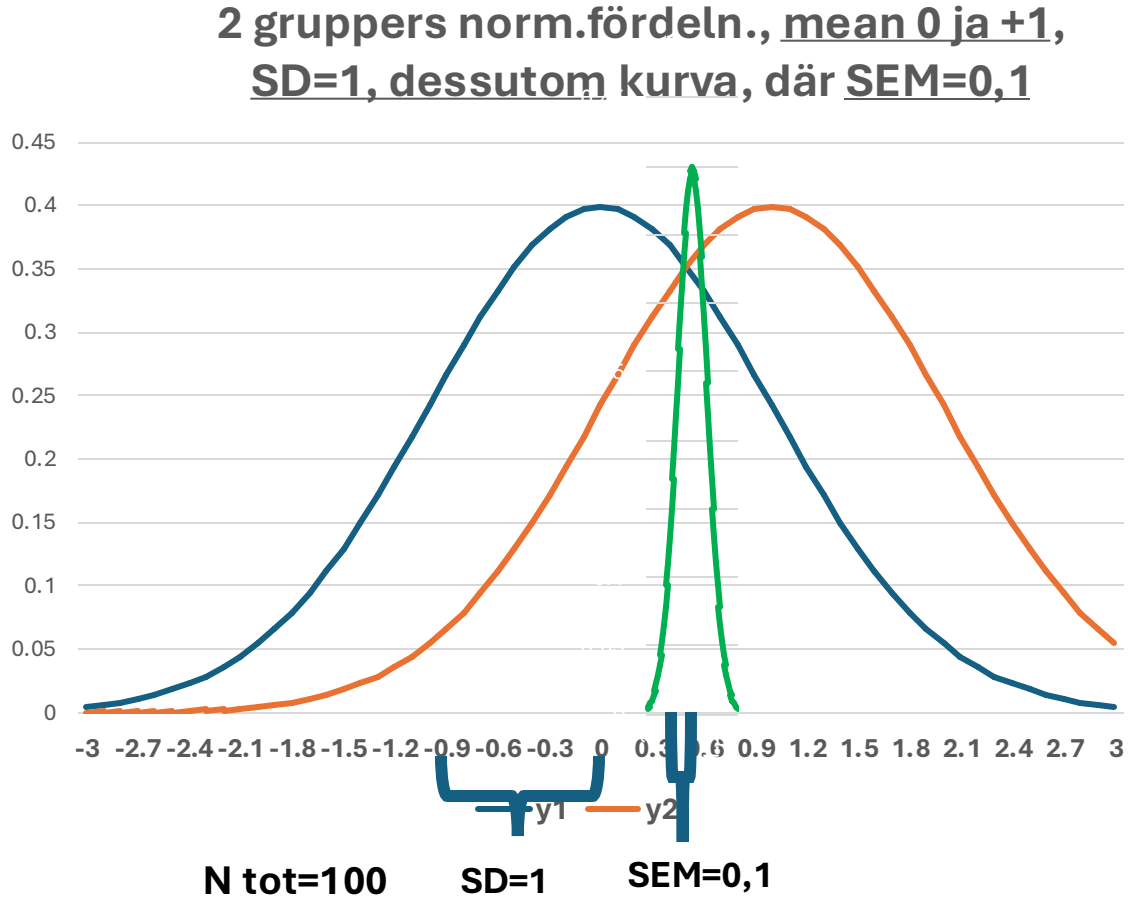


Täthetsfunktion = sannolikheten att få t.ex. exakt 3 kronor på 10 kast la

cumulative distribution function = sannolikheten att få t.ex. högst 3 kronor på 10 kast

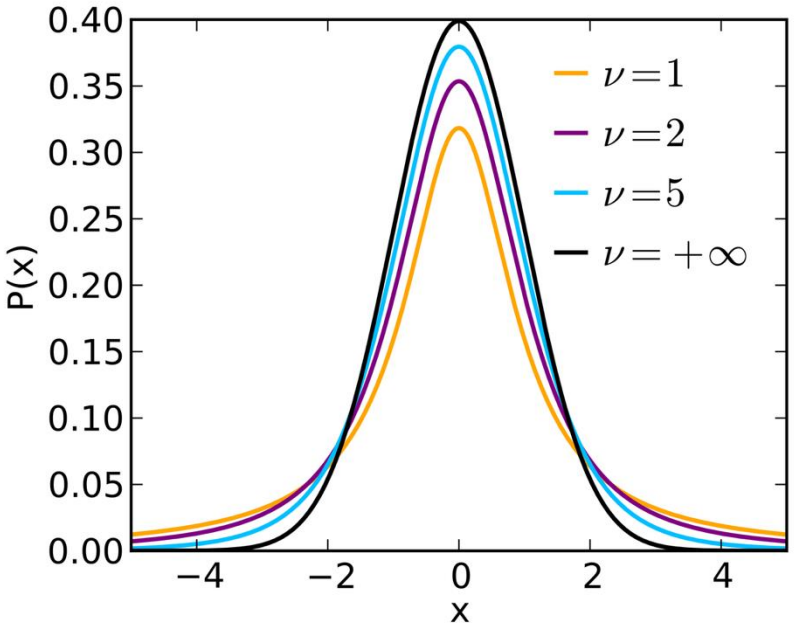
Exempel: Idén bakom P-värdesberäkning i t-test för två oberoende urval (blå och röd grupp)

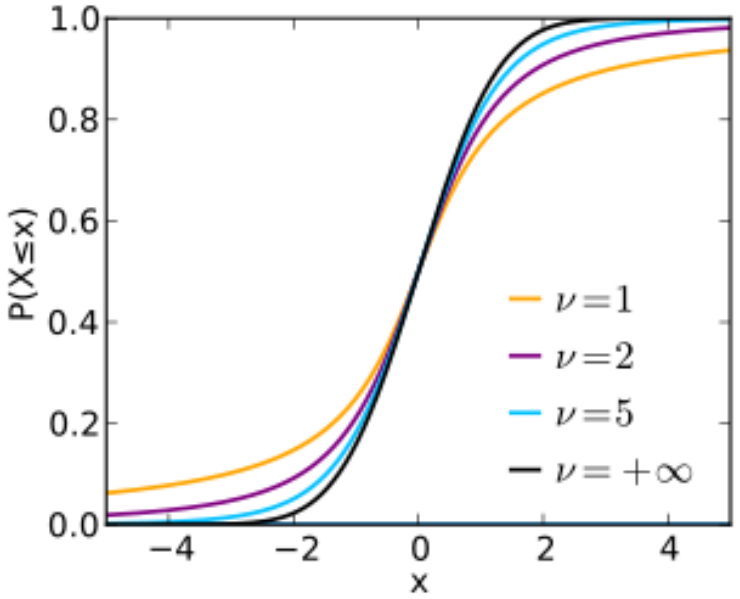
1. **Formel för testvariabeln t** i sannolikhets-fördelningen
2. $t = \frac{medelvärde_1 - medelvärde_2}{standard\ error}$
   där standard error (SEM) = $\frac{SD}{\sqrt{n}}$
3. Alltså: t = **”hur många standardfel är skillnaden mellan gruppernas medelvärden”**
4. Exempel i bilden: SEM = $1/\sqrt{100} = 1/10 = 0{,}1$
5. $t = \frac{1-0}{0{,}1} = 10$ t=0,11−0=10, df = $n_1 + n_2 - 2 = 98$
   → Skillnaden mellan gruppernas medelvärden = 10 SEM!
6. P-värdet hämtas från Student’s t-fördelning:
   P < 0,001



Exempel: t-test och P-värde

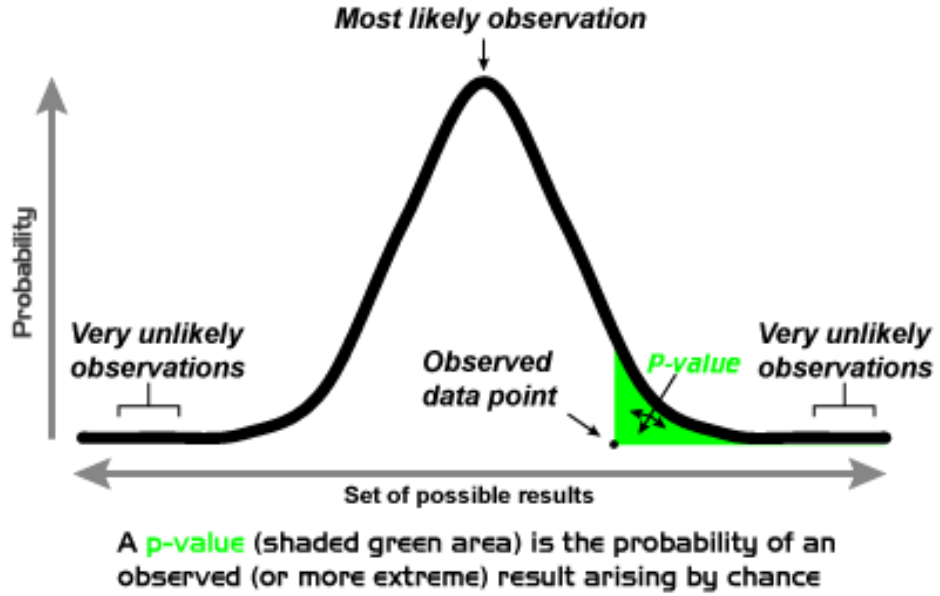
- ”I ett t-test (med Student’s t-fördelning): testvariabeln t = 10,0, frihetsgrader = 98, p < 0,001”
- Det finns även andra sannolikhetsfördelningar: normalfördelning, χ^2-fördelning, F-fördelning m.fl.
- t-fördelningen (täthetsfunktion till vänster, kumulativ distrib. funktion till höger) närmar sig normalfördelningen vid stora n





Kuvat Wikipedia

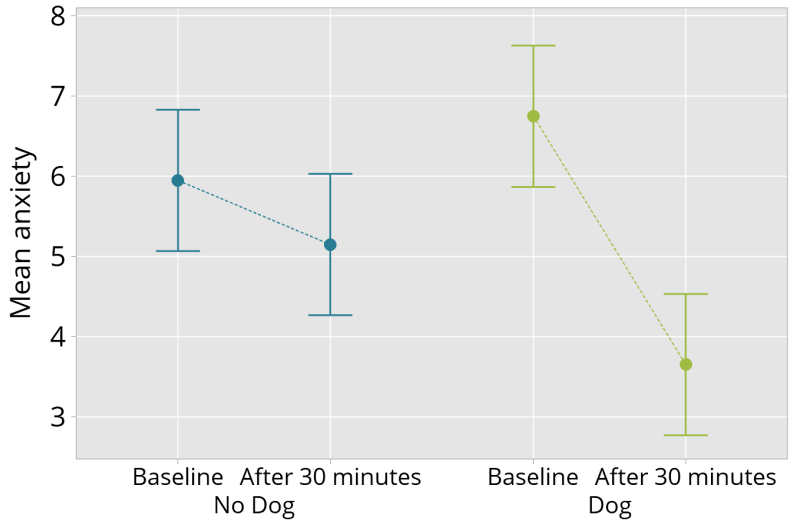
Om P-värdet – vad är statistisk signifikans?



Kuva Wikipedia

H_0-hypotesen = ingen skillnad mellan grupperi
H_1-hypotesen = det finns en skillnad
P-värdet anger hur sannolikt det är att få det ob
ett ännu mer extremt), **förutsatt att nollhypote**
Det finns inget magiskt gränsvärde, men traditionellt används p < 0,05 – numera ofta p < 0,01
I stora material upptäcks verkliga skillnader oftare; i små material är den statistiska styrkan (power) lägre
Exempel: även i var 20:e homeopatistudie kan p < 0,05 uppstå – av ren slump!
Multipla tester: om man gör 20 tester, är det sannolikt att minst ett ger p < 0,05 av en slump
P-värdet säger **inget** om skillnadens storlek – därför används allt oftare **95 % konfidensintervall**

Hur stor är skillnaden?



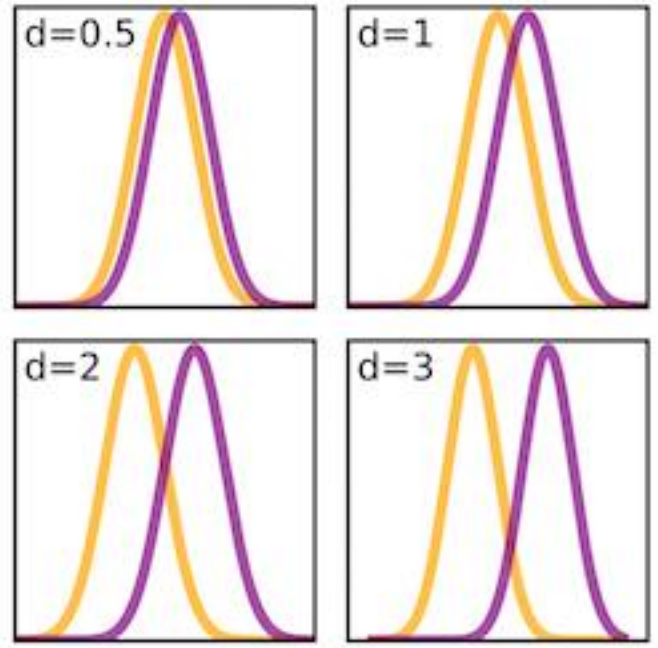
Kuvat Wikipedia

A. 95 % konfidensintervall (CI) för medelvärde

CI visar det intervall där 95 % av medelvärdena
man tog oändligt många liknande urval (illustration)
Om två CI inte överlappar, är p garanterat < 0,05

B. Effektstorlek (Effect size)

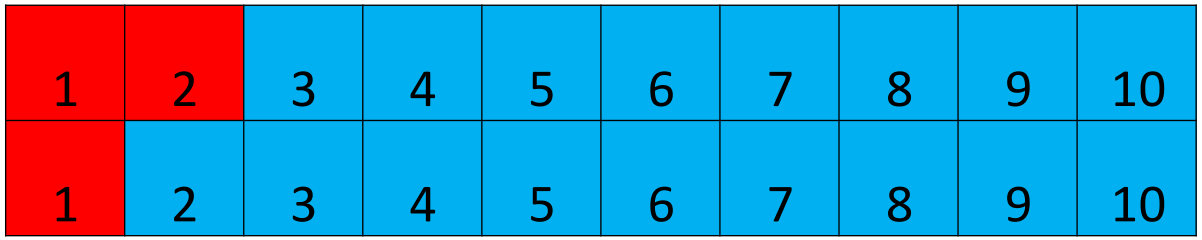
Metod 1: Hur många standardavvikelser skiljer gruppe
(Cohen's d; illustration)
Metod 2: Med hjälp av korrelationskoefficient mellan
ungefär hälften så stora värden som metod 1



Epidemiologiska mått för sannolikhet, risk och riskkvoter

Sannolikhet: 0–1 (dvs. 0 %–100 %)
Risk = sannolikheten för en (negativ) händelse
Risk ratio / relativ risk (RR) = $Risk_1$ / $Risk_2$
Hazard ratio (HR) = risk inom viss tid
Exempel: Risk A = 10 %, Risk B = 5 %

- Absolut riskdifferens = 0,10 – 0,05 = 0,05 = 5 %
- Relativ riskdifferens = 0,05 / 0,10 = 50 %
  Odds = sannolikhet att vinna / att inte vinna
  Odds ratio (OR) = Odds i grupp / Odds i referensgrupp
  Det kan vara tydligast att använda engelska termerna *odds* och *odds ratio*



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**

Vilket statistikprogram skulle man använda?

Pragmatiskt: använd det som finns tillgängligt, som du kan hantera och som innehåller nödvändiga metoder
Excel: anses ibland ”ovetenskapligt” i akademiska sammanhang, men innehåller funktioner för många analyser
SPSS: mycket använt, men dyrt utanför student-/universitetslicenser; varje modul har egen årsavgift. Mångsidigt men något tungrott, särskilt för datamanipulation jämfört med Excel
jamovi: (med liten bokstav!) gratis, kan laddas ned från https://www.jamovi.org, med omfattande dokumentation. Grundmodulen kan utökas via modulbibliotek. Lätt att lära sig, snabb att använda. Begränsade valmöjligheter, grafiken enkel och inte alltid lämplig för presentationer
R-språket: För avancerade analyser. Med över 2000 paket kan man göra ”allt”, men kräver inlärning och regelbunden användning. RStudio erbjuder ett utmärkt gränssnitt för R

4. Val av analysmetod

På vilka grunder väljs analysmetod? I/II

- Det finns oftast flera teoretiskt möjliga alternativ
- Vid enbart deskriptiv statistik används inga tester för statistisk signifikans
- Det är oftast bäst att använda den enklaste metod som uppfyller kraven – särskilt vid små n
- Ju mer sofistikerad analys, desto större datamängd krävs vanligtvis
- Det egna kunnandet och vilka metoder statistikprogrammet innehåller
- ”Överraskande” metodval bör motiveras noggrant i metodavsnittet med korrekt terminologi
- Det finns många beslutsdiagram för metodval, med något varierande innehåll

På vilka grunder väljs analysmetod? II/II

- **Antal variabler**: två eller flera (t.ex. t-test vs. ANOVA)
- **Variablernas skalnivå**: nominal, ordinal, intervall eller högre
- **Normalfördelning (N)**, om intervallskala
  - Vid normalfördelning: **parametriska** tester (kan använda medelvärde och SD), t.ex. t-test, ANOVA, Pearsons korrelation
  - Vid icke-normalfördelning: **non-parametriska** tester (baserade på rangordning; något svagare), t.ex. i stället för t-test → Mann-Whitney U-test eller Wilcoxon (vid parade data) I stället för ANOVA → Kruskal-Wallis, i korrelationer i stället för Pearson → Spearman
- **Antal observationer** per grupp och/eller per förklarande variabel (t.ex. i multivariatanalyser)
- Är det **bara samvariation** mellan två variabler (→ korrelation) eller **förklarar den ena den andra**? (→ regression)
- Uppfylls analysens **särskilda villkor**? (undersöks med ”diagnostik”)

5. Vanligaste statistiska metoderna

Deskriptiv statistik för kvantitativa variabler – vilka nyckeltal rapporteras?

(För nominal- och ordinalskalor: klassfrekvenser och procent)

Vid normalfördelning:

- Medelvärde
- Standardavvikelse (SD)
- Variationsbredd, min, max
- Ofta är datamaterialet en ”hybrid” av normal- och icke-normalfördelat, vilket kräver kompromisser i val av nyckeltal

Vid icke-normalfördelning:

- Median (md – hälften större, hälften mindre)
- Typvärde, mod (mo – vanligaste värdet)
- Nedre kvartil (Q1), övre kvartil (Q3)
- Kvartilavstånd (IQR) = Q3 – Q1
- Variationsbredd, min, max
- Skevhet och kurtosis

Korstabeller för kategorivariabler

(Crosstable)

Korstabeller för kategorivariabler

- För nominalskalor finns få tester: **Chi-två-test för oberoende** (Chi-Square Test of Independence / Association) och **Fishers exakta test**
- Testar **om två kategorivariabler är associerade**
- Samma tester **kan användas för ordinalskalor**, men då behandlas de som nominala → förlorad analyskraft
- **Särfall**:
  - Vid jämförelse av samma urval före–efter (**paired samples**) eller i fall–kontroll-studier används parat test: **McNemars test**
  - Vid **jämförelse** av distributionen **mot en förväntad fördelning** (t.ex. könsfördelning 1:1) används **Chi-två goodness-of-fit-test**

Chi-två-test för oberoende (Pearsons χ^2-test)

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

- Testar om två kategorivariabler är beroende av varandra
- **Princip**:
  - Förväntade frekvenser (E) beräknas utifrån randfördelningarna. Dessa jämförs med observerade frekvenser (O).
  - Ju större skillnad mellan O och E, desto större χ^2 och mindre P
  - Testet bygger på χ^2-fördelningen, varifrån P-värdet hämtas
- **Användningsvillkor** (programmet rapporterar dessa):
  - Högst 20 % av E får vara < 5
  - Inga E får vara < 1
- *Yates' kontinuitetskorrigering rekommenderas inte*
- *OBS: Om variabeln är ordinal → analyskraft förloras*

Utförande och tolkning av Chi-två-test

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

Bild av jamovi

- **Välj**: χ^2-test och rad- & kolumnvariabler (ordningen påverkar inte P)
- Om val mellan ensidig och tvåsidig P: välj tvåsidig, utom i undantagsfall
- **Resultat**: χ^2-värde, frihetsgrader (**df**), **P-värde**
- Fisher's P kan också visas beroende på program och inställningar
- Terminologi: korstabell = kontingenstabell

Fishers exakta test

- Unik princip: använder **ingen** sannolikhetsfördelning (norm, t, χ^2, F), utan programmet räknar igenom alla möjliga utfall och beräknar andelen som är minst lika extrema som det observerade → detta är P. Kan ta tid (minuter), ev. sätta tidsgräns (t.ex. 5 min)
- Resultatet är **exakt**, inte en approximation
- Rekommenderas **alltid** om datorn orkar – särskilt för små tabeller (2×2, 2×3, 3×2, 3×3)
- Begränsning: endast datorns kapacitet – vilket sällan är ett problem idag
- → **Använd alltid Fisher i korstabeller – åtminstone testa först!**

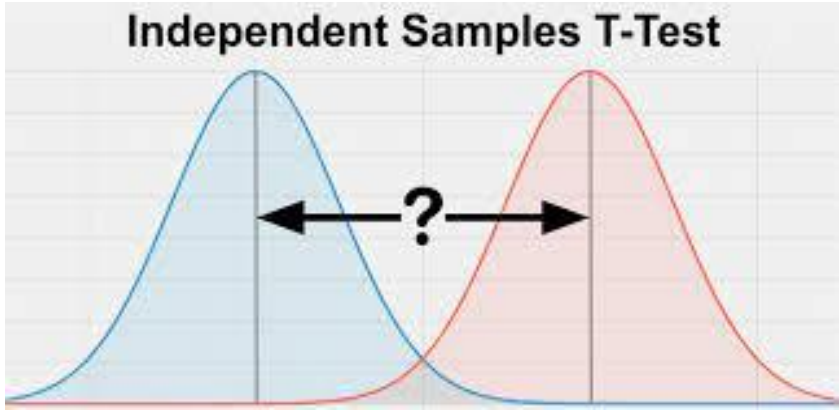
Jämförelse mellan två eller flera grupper av kvantitativa variabler

Stat. metoder: :≥2 grupper, kvantitativ variabel

Allmänt om jämförelse mellan ≥2 grupper av kvantitativa variabler

Skalnivå	**Ordinal**	**Intervall**	**Intervall**
Fördelning	–	**Ej normalfördelad**	**Normalfördelad** (och gruppstorlek ≥ 20–30)
Oberoende urval (2 grupper)	**Mann-Whitney U-test**	**Mann-Whitney U-test**	**t-test** ***(Student/Welch)***
Beroende/parade urval (2 grupper) **(2 st.) (ennen-jälkeen, case-control)**	Sign test eller Wilcoxon signed-rank	Sign test eller Wilcoxon (vid symmetrisk fördelning)	Parat t-test
≥3 oberoende grupper	*Kruskal-Wallis-test*	*Kruskal-Wallis-test*	*1-way-ANOVA*

Student’s t-test för oberoende urval
– parametriskt test för 2 grupper



Kuva Wikipedia

När kan Student’s t-test användas?

- Båda gruppernas variabler är **normalfördelade och intervallskaliga**
- **Oberoende urval** (inga kopplingar mellan observationer inom eller mellan grupper)
- **Varianshomogenitet,** dvs. varianser lika stora (Levene’s test p > 0,05)
  - Om p < 0,05 → **Welch’s t-test** i stället för Students t-test
- Gruppstorlek **minst 20–30 per grupp**
  Obs: Nästan alltid används **tvåsidigt p-värde** (utom vid mycket specifika hypoteser)

I stället för Student's t-test lönar det sig att vanligen använda Welch's test

- Identiska variansförhållanden är sällsynta
- Även andra antaganden (normalitet, oberoende) uppfylls sällan perfekt
- Welch's test tar hänsyn till olika varians → något lägre power
- I SPSS: välj nedre raden vid olika varians; i jamovi: kryssa i "Welch's"
- Welch's test kallas också Student's test med icke-homogena varianser

Mann-Whitney U-test – icke-parametrisk jämförelse av 2 grupper

När används?

- Variabeln **minst ordinalskalig**
- **Intervallskalig men ej normalfördelad** och/eller **små grupper (< 20–30)**
- **Två oberoende grupper** **Mann-Whitney U test**

Annat

- Fungerar även vid små grupper (< 20), men < 10 → svårt få signifikans
- SPSS: Analyze → Nonparametric
- Icke-parametriska tester har något lägre power
- Parvisa vertailuja: **Wilcoxon**; ≥3 grupper: **Kruskal-Wallis**

Tester för beroende (parade) urval I/II

(engl. *paired, matched, related*)

- Samma eller matchade enheter mäts två gånger
  1. **Före–efter**-design
  2. **Fall–kontroll** med matchade kontroller
- Mindre vanliga än tester för oberoende urval
- Fördel: mindre slumpvariation → högre power
- Parade tester
  - **Dikotom** (=2 klasser) nominalskala: **McNemar's test**
  - **Ordinalskala**: **Sign test** eller **Wilcoxon**
  - **Intervallskala, ej normalfördelad** men symmetrisk: **Wilcoxon**
  - **Intervallskala, normalfördelad: Parat t-test**

Variansanalys (ANOVA) – allmänt

- Används när ≥3 grupper jämförs

- Idé: total variation delas i inomgrupps- och mellangrupps-variation. Ju större andel mellangruppsvariation, desto större skillnad
- **Typer:**
  - 1-vägs ANOVA (t.ex. kön)
  - 2-vägs ANOVA (t.ex. kön och hemort)
- **Post hoc-analyser:**
  - ANOVA visar inte direkt, mellan vilka grupper skillnaden finns
  - Kräver parvisa jämförelser → multipeltestproblem!
  - Många olika alternativ, te.ex Tukey, Schèffe, Bonferroni-korrektion förstås också möjlig

Analys av samvariation

– korrelation och regression, faktoranalys

Korrelation – allmänt

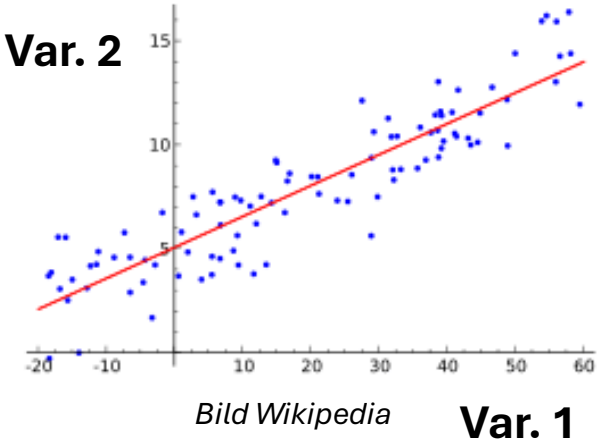
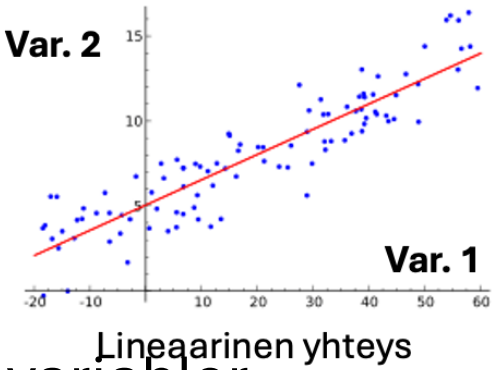


Bild Wikipedia

- Två kvantitativa variabler, undersöks för samvariation
- Ingen orsak–verkan antas (→ regression om riktning antas)
- Koefficienten **r** visar styrka och riktning: -1 … 0 … +1
  - Vad som är ”låg” eller ”hög” korrelation varierar mellan discipliner
  - r med absoluta värdet < 0,1 → ofta obetydlig (för Pearson motsvarar detta förklaringsgrad < 1 %)
- **P-värde** testar om r ≠ 0
- Presenteras som enskilda värden eller i korrelationsmatriser
- Utöver de tester som presenterats ovan kan jämförelsen av värdena för en intervallvariabel i två grupper också analyseras som en korrelation, beroende på om man är mer intresserad av skillnaden i gruppens medelvärden (t-test, etc.) eller av samvariationen mellan variabelns värden i två grupper!!

Vanliga korrelationstyper



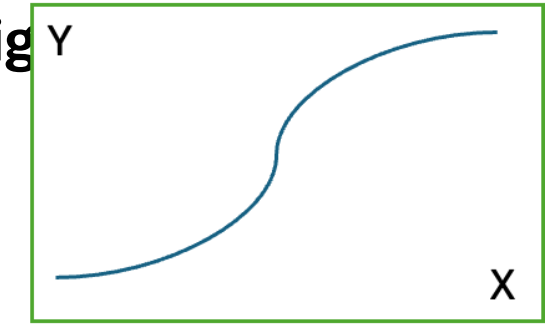
Lineaarinen yhteys

Pearsons korrelation

- **Idé**: Tar hänsyn till både rangordning och avstånd mellan värden
- **Förutsättning**: Normalfördelade intervallvariabler, även dikotoma variabler
- Vid **linjära samband** utan starka avvikande värden är Pearson starkast
  Koefficienten $\mathbf{r^2}$ (R^2) = **förklaringsgrad** av samvariationen

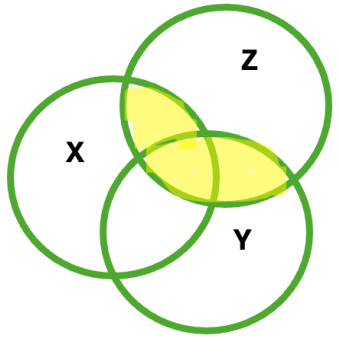
Spearmans rangkorrelation

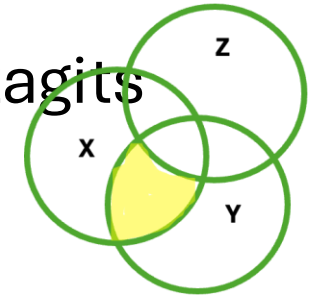
- **Idé**: Tar endast hänsyn till rangordning
- **Fungerar** för: Intervallvariabler, Ordinalvariabler, Dikotoma kategorivariabler
- Inte automatiskt bäst för någon av dessa, men mycket **mångsidig**
- Används ofta i korrelationsmatriser med blandade skalnivåer
  Pearson är starkare vid linjära samband, Spearman vid icke-linjära (t.ex. monotona)



Ns. monotoninen yhteys

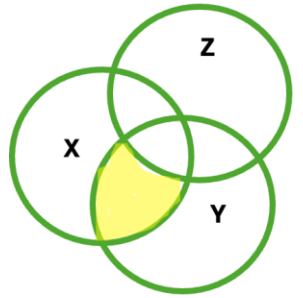
Andra korrelationstyper





- **Multipel korrelation**: Hur mycket flera x-variabler tillsammans påverkar en z-variabel
- **Partiell korrelation**: Hur mycket x påverkar y när effekten av z har tagits bort
- **Specialfall**
  - **Ordinalvariabler**: ***polychoric correlation*** – kraftfull men sällan använd
  - **Äkta dikotom + kontinuerlig variabel**: ***punkt-biserial correlation*** – idealisk, men Pearson ger identiskt resultat om dikotom kodad som 0/1
  - **Dikotomiserad + kontinuerlig variabel**: ***biserial correlation*** – men konstgjord dikotomisering rekommenderas inte!
  - **Nominalvariabel med ≥3 klasser**: kan inte korreleras direkt, om inte klasserna har tydlig ordning → då är det en ordinalvariabel!

Mer om partiell korrelation



- **Partiell korrelation** visar hur mycket x påverkar y när z:s effekt har tagits bort. Exempel: två variabler korrelerar delvis för att båda korrelerar med ålder
- Kan beräknas **med Pearson eller Spearman** (finns t.ex. i jamovi)
- **Användbar i många situationer** där vanliga korrelationer är svårtolkade, då flera variabler korrelerar starkt med varandra
  - Spearmans partiella korrelation kan användas när linjär regression inte är möjlig (t.ex. ordinalskala, liten n, sned fördelning, outliers)
- **Flera kontrollvariabler** (z) kan användas
- Utan kontrollvariabler är resultatet identiskt med vanlig korrelation!
- Med kontrollvariabel: r blir högst lika stor, oftast mindre
- Om z korrelerar starkt med både x och y → r minskar kraftigt
  Om z korrelerar svagt → r minskar lite

Regressionsanalys – allmänt

- Regressionsanalys modellerar hur **x**-variabler (**förklarande** ***engl. explanatory***) påverkar **y**-variabeln (**respons,** ***engl. response, output***)
- Viktigt att **skilja mellan två olika typer av ”förklaring”:**
- **I modellen**: x förklarar y **matematiskt**
- **I verkligheten**: x och y är endast **associerade** – kausalitet kräver särskild studiedesign (t.ex. RCT)
- **Användning**: förklaring, prediktion, modellering, klassificering
- *Endast ytlig översikt här – metoderna är många, med olika optioner*

Typer av regressionsanalys

- **Enkel (*engl. simple*) eller multipel regression**: beroende på antal x-variabler
- **Linjär regression**: linjärt samband mellan x och kontinuerlig y – grundform
- **Binär logistisk regression**: samband mellan x och dikotom y (vid ≥3 klasser: multinomial) – resultat som *Odds Ratio (OR)*
  Andra former: (Se också s. 73-74!)
  - *General Linear Model (GLM)* – alternativ till linjär regression
  - *Ordinal regression* – y är ordinal; ofta bryts antagandet om ”parallel slopes”
  - *Cox regression* – y är överlevnadstid
  - Generalized linear model, Linear/Generalized mixed model https://gamlj.github.io/book/booklet.html
- **Modellbygge** kräver ämneskunskap så att de förklarande variablerna som ingår är relevanta och inte bara matematiskt lämpliga och sådana som ger ett ”bra p-värde”
- Kontinuerliga variabler bör **inte dikotomiseras** utan goda skäl
- **Interaktioner** (x1 × x2) kan inkluderas, men gör modellen komplex och svårtolkad

Linjär regressionsanalys – allmänt

- **Modell**: $y = a + bx$
  - Med flera x: $y = a + b_1x_1 + b_2x_2 + \cdots$
  - **Klassvariabler** dummy-kodas (0/1) – en klass som referens
  - **Residualer** = variation som modellen inte förklarar
- **Antaganden**
  1. **Linjära** samband
  2. **Ingen multikollinaritet** (hög korrelation mellan x-variabler)
  3. **Minst (10-) 20–50 observationer per x-variabel**
  4. **Inga** starka **outliers**
  5. **Residualanalys** centralt!:
     - (A) normalfördelade
     - (B) homogen varians
     - (C) oberoende (inga mönster i residualplot)

Linjär regression i jamovi – exempel

- **Välj analysmodul**
- **Ange variabler:**
  - y → ***Response*** *variable*
  - kontinuerliga x → ***Covariates***
  - kategoriska x → ***Factors*** (dummy-kodas automatiskt)
- **Ytterligare inställningar**:
  - Referensnivåer för faktorer
  - Kontroll av antaganden
  - R^2, p-värden för modell och koefficienter
  - Konfidensintervall, standardiserade koefficienter
- **Rapportera**:
  - Modellens signifikans ($p < 0,05$?)
  - Förklaringsgrad (justerad R^2)
  - Koefficienter och deras p-värden
  - Kollinearitet
  - Residualanalys

Linjär regressionsanalys – genomförande i jamovi

Förklaringsgrad (R^2)

Beroende variabel (y)

Koefficient (b): hur mycket y förändras när x ökar med en enhet

p-värde för x: testar om x har signifikant effekt på y

Standardiserade koefficienter: gör effekterna jämförbara

Outliers: värden med påverkan > 1

DW): mäter om residualerna korrelerar; idealvärde ≈ 2

Multikollinearitet: tolerans < 0,2 tyder på problem; idealvärde ≈ 1

Intercept

Covariate

Factor, dikotom

Factor, 2 dummy+ref)

Factor, 2 dummy+ref)

Covariate

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

Data Summary

Cook's Distance

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

Assumption Checks

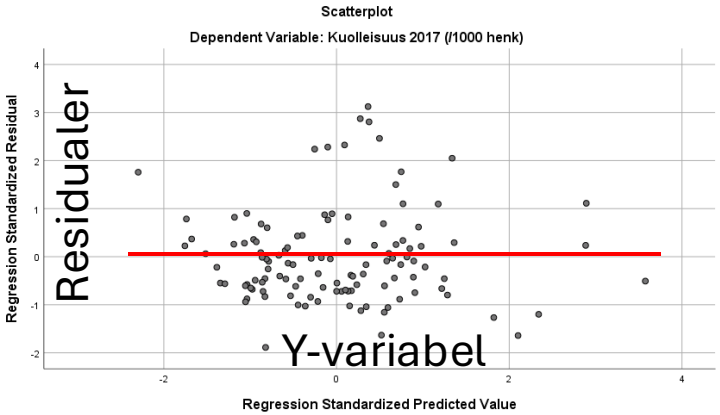
Durbin–Watson Test for Autocorrelation

Autocorrelation	DW Statistic	p
-0.028	2.056	0.06600

Collinearity Statistics

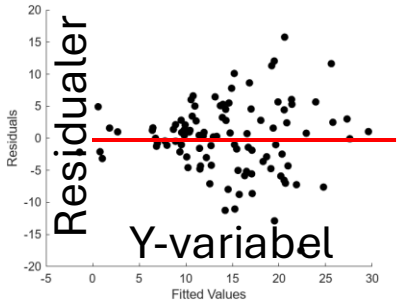
	VIF	Tolerance
Sum_14_utför_extrov	1.154	0.867
modersmål4_dummy_Sv1_FiRef	1.012	0.988
utbild_Lev4_dum	1.030	0.971
utbild_Lev5_dum	1.046	0.956
kännsfys27a_dummy_Yngre1_LikaRef	1.077	0.928
kännsfys27a_dummy_Äldre1_LikaRef	1.062	0.941
ålder	1.076	0.930

Grafisk tolkning av residualer i linjär regression



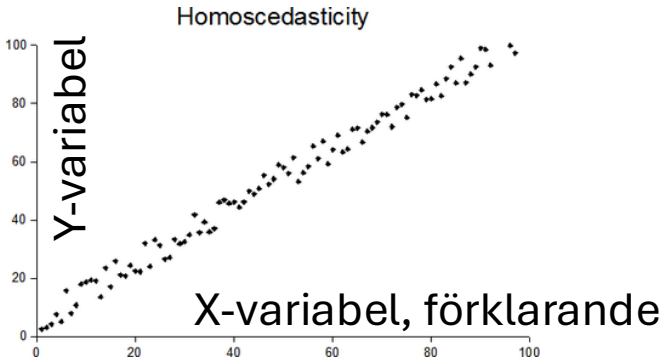
Normalt fördelade residualer: tätast kring 0

Källa: Tietoarkisto, Tfors univ., andra bilder: Wikipedia

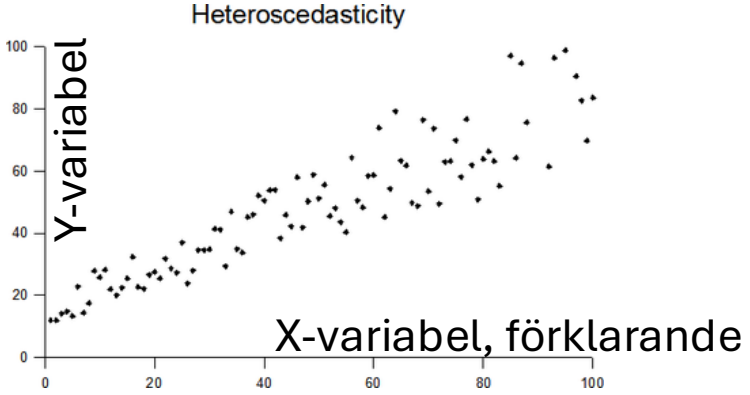


Normalt fördelade residualer: tätast kring 0

Heterogen varians: ”molnet” breddas åt höger → SD ökar



Homogen varians: jämn spridning över hela x-området



Icke-homogen varians: spridning ökar med x

→ problem med modellens antaganden

Binär logistisk regressionsanalys – grundidé

- **När** används: När **y är dikotom** (två kategorier, t.ex. 0/1)
- **Modellens logik:**
  - Linjär regression: y=a+bxy=a+bx
  - Logistisk regression: samma högra sida (a+bxa+bx), men vä. sidan (y) kan bara vara 0 eller 1
- **Lösning till den problematiska vä sidan**:
  - Ersätt y med Odds: **p/(1-p)**
  - Sedan med log(Odds): **ln (p/1-p)**
    → vänstra sidan kan anta alla värden → modellen fungerar!
- Priset för en matematiskt fungerande men komplex modell är att den är svårare att förstå konkret och att sambandet mellan x och y inte är linjärt, utan S-format
- **Resultaten** redovisas inte med koefficienterna för de förklarande variablerna, utan med hjälp av : ”När värdet på den förklarande variabeln ökar med en enhet, då är Odds-kvoten =**Odds-kvoten** t.ex. 1,5, att värdet på responsvariabeln är 1 istället för 0” (om Odds ratio/kvoten se bild 26!).

Binär logistisk regression – antaganden och genomförande

Antaganden:

- Ingen stark multikollinearitet, Få outliers, Alla kombinationer av x bör finnas i datan, Residualer: normalfördelade, homogen varians, oberoende
- Datamängd: minst dubbelt så stor som vid linjär regression, Alternativt: antal x ≤ 1/3 (eller 1/10) av den mindre y-gruppens storlek
- Ej krav på normalfördelning eller linjäritet

Genomförande i jamovi:

- **Dependent**: dikotom y-variabel
- **Covariates**: kontinuerliga x-variabler
- **Factors**: kategoriska x-variabler
- Välj referensnivåer för faktorer
- Begär:
  - Estimat, Odds Ratio (OR) och dess konfidensintervall
  - Granskningar av förutsättningar att använda metoden, p-värden för modell och variabler, pseudo-R^2
  - prediktionsmått (sensitivitet, specificitet)

Binär logistisk regression – Genomförande i jamovi

Ovanför **bilden** (jamovi) finns ett område där du anger **variabler** som i linjär regression

- Beroende – 2-klass responsvariabel
- Kovariater – kontinuerliga variabler
- Faktorer – kategorivariabler

Från avsnittet **Referensnivåer** väljer du en referens från de olika kategorierna av kategorivariabler, som de andra jämförs med. Detta gäller både förklarande och responsvariabler

Sammanfattning om användning av regressionsanalyser

- Fler x → fler korrelationer, interaktioner, kausalkedjor, bias, mätfel, risk för överanpassning
- → **Använd enkla modeller med relevanta x-variabler**
  - Välj inte enbart utifrån p-värden
  - Undvik p-hacking
- Resultat bör ses som **indikativa**, inte absoluta
- Kausalitet kräver särskild studiedesign – annars bara association
- Testa gärna flera regressionsmetoder – om resultaten är samstämmiga, ökar trovärdigheten
- Vid små datamängder: linjär regression (om antaganden uppfylls)
- Om regression inte är möjlig: använd korrelationer, subgruppsanalyser, grafik

Explorativ faktoranalys (EFA) – allmänt

- Här ges endast en översiktlig presentation av EFA
- Förutom matematik kräver EFA **omfattande tolkningar och val baserade på ämneskunskap**
- Syftet är **att identifiera underliggande faktorer** som förklarar variationen i ett antal observerade variabler
- Dessa ”**korrelationskluster**” av variabler **representerar latenta faktorer – teoretiska konstruktioner** som inte kan mätas direkt (t.ex. intelligens mäts via flera verbala, spatiala, logiska, matematiska tester)
- Exempel: en allmän kompetensfaktor för studenter, där varje delområde (gruppfaktor) mäts med flera frågor som korrelerar inom området men svagare med totalpoängen
- I **konfirmatorisk faktoranalys (CFA)** har forskaren en förhandsuppfattning om faktorstrukturen, som testas mot data – behandlas ej här

Terminologi i EFA

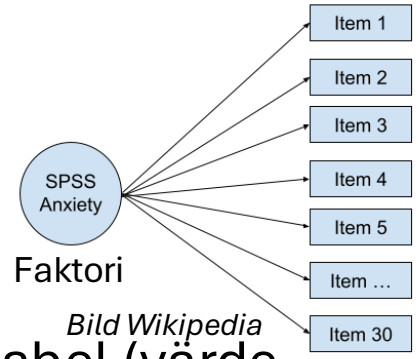


Bild Wikipedia

- **Faktorladdning (factor loading)** = korrelation mellan faktor och variabel (värde mellan -1 och +1)
- **Eigenvalue (egenvärde)** = summan av kvadrerade faktorladdningar
- **Förklaringsgrad per faktor** = eigenvalue / antal variabler (värde mellan 0–1)
- **Kommunalitet** = hur väl faktorerna förklarar variationen i en enskild variabel (summa av kvadrerade laddningar)
- **Unikhet (uniqueness)** = den del av variationen som inte förklaras av faktorerna = 1 – kommunalitet

Antaganden och förutsättningar

- **Sphericity**: Bartlett's test $p < 0,05$
- **Sampling adequacy**: Kaiser-Meyer-Olkin (KMO) index ≥ 0,5
- **Urvalsstorlek**:100 = svagt, 300 = bra, 1000 = utmärkt
  - Nummenmaa: ≥200, Field: ≥10–15 observationer per variabel

Val av faktorkombination

Kriterier – ofta motstridiga:

- Hög förklaringsgrad
- Få faktorer, helst ≥3 variabler per faktor
- Stora eller små laddningar (undvik medelstora)
- Faktorerna ska vara meningsfulla att tolka

Praktiska urvalskriterier:

- **Eigenvalue > 1**
- **Scree plot** – brytpunkt
- **Parallel analysis** – antal extraherbara faktorer (±1 noggrannhet)
  - Rekommenderas, men EFA är inte strikt regelstyrt

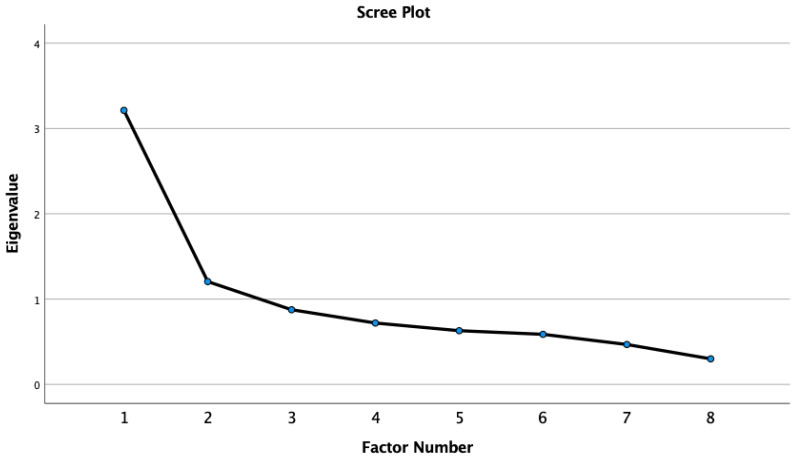


Bild Wikipedia

Övriga kriterier:

- Signifikant faktorladdning: > 0,3 (vid n ≈ 300) eller > 0,5 (vid n ≈ 100)
- Signifikant kommunalitet: > 0,3–0,5

Genomförande av EFA

- Välj variabler och mata in i programmet
- Kontrollera antaganden
- **Extraktion**: generera faktorer och laddningar
  - Laddningar visar korrelation mellan faktorer och variabler
  - Extraktionsmetoder i jamovi: Minimum residuals, Principal axis, Maximum likelihood (för stora, normalfördelade dataset)
- **Rotation**: omstrukturera laddningar
  - Maximerar laddning till en faktor, minimerar till andra
  - Ortogonala metoder: Varimax, Quartimax
  - Oblique metoder: Oblimin, Promax
  - Vanligtvis används oblique metoder om faktorer korrelerar > 0,3

Överlevnadsanalyser (Survival analysis)

- **Används** för att analysera tid till händelse – både negativa och positiva **Exempel**: sjukdomsrecidiv, flytt, byte av arbetsplats, institutionsvård, tillfrisknande från long covid, äktenskap
- **Under uppföljning**: Vissa upplever **händelsen**, vissa **censureras** (försvinner utan händelse)
- **Vissa händelser är kombinerade**: t.ex. progression-free survival kan sluta med recidiv eller död
- Om till exempel återfall och icke-återfallsmortalitet (NRM) analyseras separat, bör helst en mer noggrann **competitive events-model** användas, som skiljer mellan till exempel förlust av uppföljning genom NRM och annan förlust av uppföljning, d.v.s. det finns tre grupper av händelser vid uppföljningens slut (1. händelse under studie, 2. annan händelse och 3. censurerad). Detta kräver R-språksanalyser och diskuteras inte mer i detalj här, se t.ex. https://www.ebmt.org/sites/default/files/2018-03/Suggestions%20on%20the%20use%20of%20statistical%20methodologies%20in%20studies%20of%20the%20EBMT.pdf

Metoder inom överlevnadsanalys

Livstidsdata (life table):

- **Registrera**: antal levande vid periodens början, antal döda under perioden, antal censurerade
- **Beräkna**: riskpopulation, dödsrisk, sannolikhet att överleva, kumulativ överlevnad

Kaplan-Meier-kurva:

- Visar **överlevnad** över tid
- Jämförelse mellan grupper med **Log-rank-test**

Cox regressionsanalys:

- Liknar andra regressionsmetoder
- Kräver: tidsvariabel, statusvariabel, förklarande variabler

Meta-analys

Allmänt

- Detta är endast **en ytlig översikt** – ämnet är mycket omfattande
- Metaanalys är en metod inom vetenskaplig forskning där man systematiskt samlar och kombinerar kvantitativa studier
- Vanligtvis inkluderas RCT-studier eller andra experimentella eller observationella studier med jämförelsegrupper (kohort, fall–kontroll). Pre–post-design är sämre
- Metaanalyser står högst i många evidenshierarkier:
  - *Metaanalys > systematisk översikt > experimentell studie > observationsstudie*
- De citeras ofta, särskilt inom medicin, t.ex. vid riktlinjer och evidensbaserade rekommendationer
- Kan vara separat analys eller en del av en mera omfattande systematisk översikt
- Att lära sig metaanalys kräver mer tid än vanliga statistiska metoder – kräver även många omvandlingsformler
- Vid första metaanalysen rekommenderas stöd från en erfaren forskare!

Steg i meta-analys
(ref: Muka T. et al)

- Forskningsfråga > Sökstrategi > Inklusionskriterier > Databassökningar > Insamling av referenser och abstrakt > Stegvis screening och granskning > Urval av artiklar > Datainsamling och bearbetning > Kvalitetsbedömning > Filskapande > Deskriptiv syntes och flödesschema > Kvantitativ syntes (själva metaanalysen) > Analys av heterogenitet och publikationsbias > Bedömning av evidensens kvalitet > Rapportering

Teknisk genomföring av metaanalys

- **Fixed-effect eller random-effects**-meta-analys?
  - *Fixed-effect*: samma effekt antas i alla studier
  - *Random-effects*: olika studier antas mäta olika effekter
- **Data måste matas in i standardiserat format**
  - **Ofta krävs transformation eller ett selktivt urval** av endast delar av resultat
  - Dessa är ofta mycket mödosamma steg
  - **Dikotoma utfall**: riskkvot, riskdifferens, odds ratio
    - Variabler: antal händelser och urvalsstorlek i interventions- och kontrollgrupp
  - **Kontinuerliga utfall**: medelvärdesskillnad, standardiserad skillnad
    - Variabler: medelvärde, SD och urvalsstorlek
  - **Effektstorlek**: t.ex. Cohen's d och standardfel (eller varians)
- Program: **jamovi** med ***MAJOR*-modul** (efter initial inlärning lättanvänd)
  - Alternativ: STATA, R (paket som *meta*, *metafor*)

Resultat i metaanalys

- **Forest plot**: visar varje studies effektstorlek och CI samt samlad effekt
- **Funnel plot**: används för att bedöma publikationsbias
- **Heterogenitet**: variation mellan studier
  - Finns den? → *tau*^{2}, *Q-test*
  - Hur mycket? → I^2*-statistik*
- **Avvikande** värden (influence)
- **Metaregressioner** eller **subgruppsanalyser**

Litteratur och resurser om meta-analys

- Cochrane Handbook for Systematic Reviews of Interventions, särskilt Part 2: 6. Effect measures 10. Meta-analyses https://training.cochrane.org/handbook/current
- Jouko Miettunen, Presentations (2 föredrag om dessa): 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
- Dessutom olika programpaket (bl.a.. meta, metafor) har sina egna webbplatser

Vanliga orsaker till att MAJOR-modulen inte fungerar i jamovi:

- Tomma celler i data
- En tom rad i slutet av importerad CSV-fil (överraskande t.ex. av någon teknisk orsak)

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

En grupp mångsidiga modeller som kan anpassas till ett brett spektrum av analyser

Klassificering av modeller

Oberoende variabler	Beroende variabel: Normalfördelad & kontinuerlig	Beroende variabel: Icke-normalfördelad eller kategorisk
Oberoende observationer	General Linear Model (GML)	Generalized Linear Models
Grupperade observationer	Mixed Model	Generalized Mixed Models

- En grupp mångsidiga modeller där analyser och rapportering av resultat har systematiserats
  - **GLM** kan ersätta linjär regression, ANOVA och kovariansanalys (ANCOVA) osv.
  - **Generalized Linear Model**: Den beroende variabeln omvandlas till en kontinuerlig variabel med hjälp av en länkfunktion, och dess avvikande fördelning beaktas för att erhålla följande modeller: binomial och multinomial logistisk, Poisson, ordinal etc.
  - **Mixed models**: Variablerna är inte helt oberoende av varandra om de tillhör en av flera grupper (**multilevel**modeller, t.ex. skolelever och skolor), i vilket fall de som tillhör samma grupp är mer lika varandra och konventionella tester ger något felaktiga resultat. Det finns ett liknande beroende i tidsseriemätningar.
- Se mera t.ex. https://gamlj.github.io/book/gzlm.html om också är en guide till tilläggsmodulerna **gamlj** och **GAMLj3** som används i dessa analyser.

Multipel imputation

för att fylla i saknade värden

Allmänt om multipel imputation

(MI, multiple imputation)

- **Ersättning av saknad data med endast ett värde är problematiskt**
  - SD och därmed SE (standardfel) för litet, vilket gör statistiska slutsatser "för bra"
  - I medelvärde-imputation är ersättningen i en viss kolumn/rad fast, inte baserad på prediktion
  - Multipel imputation undviker ovanstående problem
- **Beskrivning av MI-metoden och processen**
  1. **Flera imputerade versioner av den kompletta datan skapas** (standardvärde = 5, men krävs t.ex. 20 eller till och med 100) från källdatan, där ersättningen av saknade värden baseras på en matematisk prediktion, till vilken en lämplig slumpmässig variation också läggs till. Standardvärdet för **iterationer** (repetitioner) = 5, men det kan krävas t.ex. 20-40 st. tills den så kallade konvergensen uppnås.
  2. **Parametrarna av intresse analyseras från varje skapad fil** (deskriptivt eller t.ex. genom regressionsanalys). Resultaten av dessa skiljer sig åt beroende på mängden osäkerhet relaterad till imputationerna.
  3. **Resultaten kombineras** ("pooling"). I detta fall beaktas, utöver variationen i källdatan, mängden osäkerhet relaterad till imputationerna. På detta sätt finns det ingen bias i MCAR- och MAR-data och de statistiska egenskaperna är lämpliga.
- **MI kan utföras** med R-programmet (*mice*-paketet) eller även Stata, SAS, SPSS (inte i studentversionen).

Annat om att producera MI

- **Predictor matrix** är en tabell som anger vilken variabel som förutsäger vilken variabel.
  - Meddelas även t. ex. vilket tal som är matematiskt helt definierat av ett annat (t.ex. indikatorns poäng = summan av poängen i dess frågor).
- **Imputationsmetoder** för saknade värden. Antagande t.ex. kontinuerliga variabler: PMM = Prediktiv medelvärdesmatchning och dikotom: logistisk regr.
  - Dessa kan ändras vid behov. Även så kallad **post-processing**, vilket begränsar värdena för de imputerade värdena till endast de möjliga värdena för variabeln i fråga.
- **Visuell undersökning** av mean och SD för de imputerade variablerna: Målet är att ”fläta samman” **konvergenskurvorna** med varandra, utan trender i senare iterationer.
- Att använda metoden i R kräver kunskap om språkets grunder och kräver något mer inlärning än grundläggande statistiska metoder. Rekommenderade källor:
  - Vignetter 1-4 i Mice README https://cran.r-project.org/web/packages/mice/readme/README.html
  - Valda avsnitt Vanbuuren: Flexibel imputering av saknade data https://stefvanbuuren.name/fimd/

6. Hjälp från nätet – kalkylatorer

Kalkylatorer - allmänt

- Data bör förstås i regel analyseras av forskargruppen/statistiker
- Ibland behövs enkel extraberäkning från redan sammanställda data
- **OBS:** mata aldrig in personidentifierbar data online!
- **Användningsområden:**
  - Urvalsstorleksberäkning för t-test
  - P-värde från korstabeller (χ^2 eller Fisher)
  - Konfidensintervall för andelar/procent

Exempel på nätkalkylatorer

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

- Proportion /procent CI: https://www.statskingdom.com/proportion-confidence-interval-calculator.html
- Exempel: 3 av 12 → 25,0 % (95 % CI: 8,9–53,2 %)
  - Tryck på Calculate, titta nedan
  - Vi får: 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) rutten:
  Frequency data>For a 2x2 Table of Cross-Categorized Frequency Data>Version 1
- Genom att skriva in siffrorna i de gula rutorna och trycka på Calculate får du Chi-kvadrat-värdena för Yates och Pearson samt den lägsta tvåsidiga Fisher-metoden (vilket i princip alltid är den mest exakta!)

Hjälp från nätet

Results - APA Style (Wilson score interval)

95% CI [0.089, 0.53]

The estimated probability of success is $\hat{p} = 0.25$.

Normal approximation - good for students!

Students are often expected to utilize the normal approximation as it offers a more straightforward calculation method compared to other techniques.
Confidence interval: [0.005005 , 0.495].
Alternatively: 0.25 ± 0.245
The margin of error (MOE) or estimated bound on proportion (EBP) is 0.245 or 24.4995%
As a rule of thumb, the normal approximation method requires a large sample size of at leas 30.

Wilson score interval ✔ recommended!

The Wilson score interval is usually the recommended method for calculating confidence intervals for proportions.
Confidence interval: [0.08894 , 0.5323].
Alternatively: 0.3106 ± 0.2217
The margin of error (MOE) or estimated bound on proportion (EBP) is 0.2217 or 22.1682%

Data Entry

		X = 0	X = 1	Totals
Y	1	20	28	48
	0	20	12	32
Totals		40	40	80

Expected Cell Frequencies per Null Hypothesis

24	24
16	16

Calculate Reset

Phi	Chi-Square: Yates	Chi-Square: Pearson
+0.2	2.55	3.33
P	0.110294	0.068027

Chi-square is calculated only if all expected cell frequencies are equal to or greater than 5. The Yates value is corrected for continuity; the Pearson value is not. Both probability estimates are non-directional.

Fisher Exact Probability Test:

P		
	one-tailed	0.05475426209327384
	two-tailed	0.10950852418654773

Home | Click this link **only** if you did not arrive here via the VassarStats main page.

Bra och breda webbresurser inom statistik

- **Kvantitatiivisen tutkimuksen verkkokäsikirja (Tietoarkisto)** https://www.fsd.tuni.fi/fi/palvelut/menetelmaopetus/kvanti/
  - En gedigen guide för forskare
- **Learning statistics with jamovi (Navarro&Foxcroft)** https://blog.jamovi.org/2018/10/25/learning-statistics-with-jamovi.html
  - En utmärkt guide till jamov och allmän statistik. Det finns också en onlineversion, men du kan förmodligen ladda ner den till din PC här!
- **Tilastokunto (Aleksi Reito)** https://www.tilastokunto.fi
  - Mångsidiga sidor av en finsk docent i ortopedi
- **Statistics notes (The BMJ)** https://www.bmj.com/specialties/statistics-notes
- **Presentations- Jouko Miettunen** http://www.joukomiettunen.net/presentations/
  - En stor uppsättning bra, tydliga presentationer från en epidemiologiprofessor från Uleåborg
- **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/
  - Jag har tagit fram dessa länkar, sidan innehåller även länkar till statistiska delområden och ett Excel-utbildningsdokument som jag har förberett.

7. Rapportering

Rapportering

Metoder

- Beskrivs med sådan noggrannhet att en kunnig läsare kan bedöma deras lämplighet
- Sammanfogning av variabelklasser och andra bearbetningar inför analys ska motiveras vid behov

Resultat

- Följ tidskriftens instruktioner noggrant:
  - n, %, antal decimaler, CI-format, exakt p-värde eller ***
- Alla metoder som används i resultatdelen ska vara beskrivna i metodavsnittet

Diskussion

- Använd exakta termer:
  - Skriv inte *correlates* om inga korrelationer gjorts
  - Använd *significant* endast för statistisk signifikans
  - Undvik *explains, causes* om inte kausalitet är bevisad
  - Använd istället: *association, connection, relation, correlation*