

II Tilastollisen aineiston ja analyysin edellytysten tarkistaminen

- Tietojen syöttö
- Karma&Komulainen –aineisto (tutustuminen)
- Muuttujien jakauman tarkistus
- Puuttuva tieto ja sen käsittely
- Muunnokset, uudelleen koodaaminen, summamuuttujien luominen (oma aineisto)
- Kuvaileva tilastoanalyysi vs. tilastollinen päättely

Tilastolliset tunnusluvut (lyhyt kertaus)

Muuttujassa oleva informaatio voidaan tiivistää muutamiaan muuttujaa kuvaavaan tunnuslukuun. Jakauman **sijainnin** tunnuslukuja ovat seuraavat keskiluvut: moodi, mediaani, fraktiilit ja keskiarvot.

Statistics		
matematiikan arvosana		
N	Valid	30
	Missing	0
Mean		7,23
Median		7,00
Mode		7
Percentiles	25	6,75
	50	7,00
	75	8,00

I Jakauman **sijainnin** tunnusluvut (eli Keskiluvut)

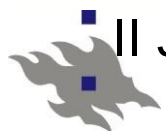
Keskiluvut luonnehtivat numeeristen muuttujien jakauman sijaintia tai painopistettä lukusuoralla.

Moodi (engl. mode) on se muuttujan arvo, jonka frekvenssi jakaumassa on suurin.

Mediaani (engl. median) on se muuttujan arvo, joka jakaa havaintoaineiston kahteen yhtäsuureen osaan, se on "keskimmäinen" havainto. (Vaatii vähintään järjestävän asteikon.)

Keskiarvon (engl. mean) laskemiseen tarvitaan vähintään välimatka-asteikollinen muuttuja.

Fraktiilit: Kvartiilit (engl. quartiles) jakavat aineiston neljään osaa, neljänneksiin. Kvartiilit suuruusjärjestyksessä ovat alakvartiili, mediaani ja yläkvartiili.



■ II Jakauman **muodon** tunnusluvut (eli hajontaluvut)

- **Hajontaluvut:** Hajontalukujen avulla kuvataan aineistossa esiintyvän vaihtelun määrää. Niiden avulla voidaan ilmaista, kuinka voimakkaasti mittaustulokset vaihtelevat havaintoyksiköstä toiseen tai missä määrin ne keskittyvät keskiluvun läheisyyteen.
Hajonnan pienuus → poikemaa vain vähän jakauman keskiluvusta.
Hajonnan suuruus → aineistossa on paljon vaihtelua keskiluvun ympärillä.

Vaihteluväli (engl. range) kertoo muuttujan pienimmän ja suurimman havaintoarvon. Sopii järjestysasteikollisten muuttujien tarkasteluun.

Keskihajonta (engl. standard deviation) on luontevin hajontaluku numeerisille muuttujille. Se kuvaa muuttujan arvojen vaihtelun suuruutta.

Keskihajonnan neliö, varianssi (engl. variance) kuvaa suhteellista vaihtelua

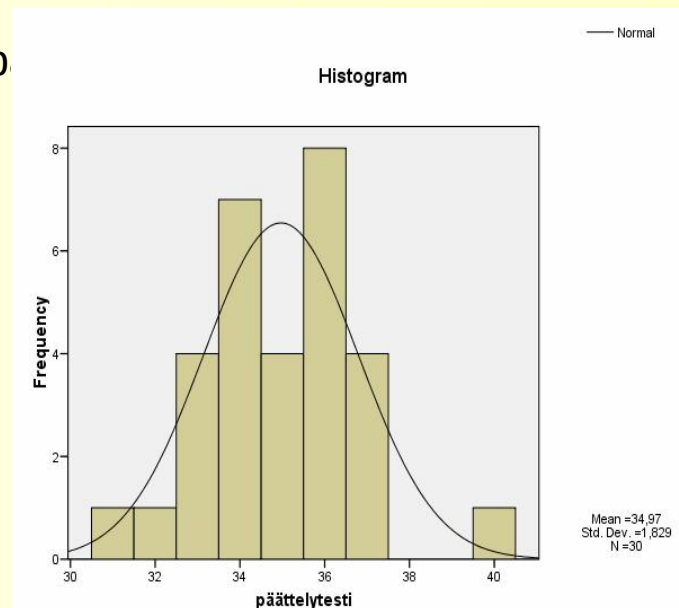
Kvartiilipoikkeama ilmaisee, kuinka pitkällä välillä aineiston keskimmäiset 50% sijaitsevat.

Muita tunnuslukuja: (Jakauman vinoutta kuvaava vinousmitta (engl. skewness). Jakauman terävyyttä kuvaava huipukkuusluku (engl. kurtosis).



Normaalijakauma

- Muuttujan jakauman normaalisuus on monien tilastollisten testien oletus
- Normaalijakauma on symmetrinen, sen sijainti ja muoto riippuvat keskiarvosta ja hajonnasta
- Havaintojen jakautumista keskiarvon (mean) ympärille kuvataan keskihajonnalla (standard deviation)
- Normaalijakauman havainnoista 95 % sijoittuu lähemmäs kuin kahden keskihajonnan päähän keskiarvosta.





Jakauman tarkastelua: *vinous ja huipukkuus*

- Normaalijakauman vinous (skewness) ja huipukkuus (kurtosis) ovat nollia.
- Jakauman vinouden ja huipukkuuden nolasta eroavuutta voi testata jakamalla saatu arvo sitä vastaavalla keskivirheellä (Standard Error). Jos näin saatu luku on $< |2|$, voidaan jakauman vinous ja huipukkuus hyväksyä vielä normaaliseksi ja siten tarkasteltava jakauma on riittävän normaalin tilastollisiin testeihin (vrt. Jos taas on $> |2|$ niin vinous/huipukkuuskerroin eroaa tilastollisesti merkitsevästi nolasta)
- Jakauman vinous ja huipukkuus pulmallisia perinteisissä tilastoanalyyseissä.



Vinous (vrt. normaalijakauman vinous = 0)

→ Skewness = vinouskerroin (vk) ilmoittaa, mihin suuntaan jakauma on vino; Jos $vk > 0$, positiivisesti vino eli oikealle vino

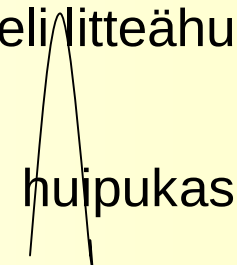
=> Jos $vk < 0$, negatiivisesti vino eli vasemmalle vino

■ Huipukkuus (vrt. normaalijakauman huipukkuus = 0)

=> Kurtosis (kur) = huipukkuuskerroin ilmoittaa, kuinka terävähuippuinen jakauma on

=> Jos $Kur > 0$, jakauma on terävähuippuinen

=> Jos $Kur < 0$, jakauma on huiputon eli litteähuippuinen





Aineiston tarkistus ja poikkeavat arvot (outlier:it)

Yksittäiset äärimmäisen suuret tai pienet arvot voivat tuottaa pulmia analyyseissa

- Kannattaa tarkastella myös graafisesti (esim. boxplot)
- Jo ennen keräystä on hyvä miettiä muuttujien vaikutukset analyyseihin:

Esim. Kuntien asukkaiden keskitulovertailuissa

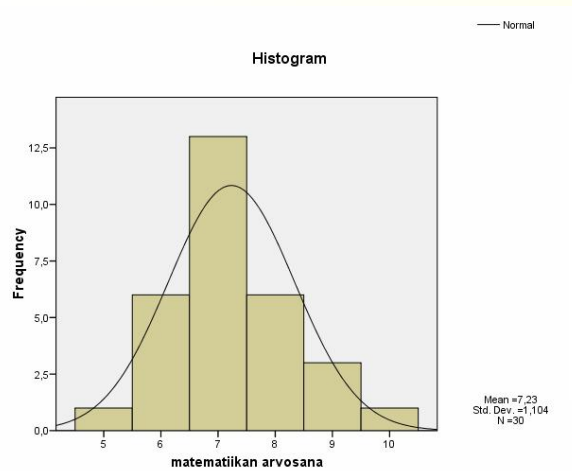
- Kauniainen vääristää kuntien vertailua,
 - Optiomiljönääri vanhempien palkkojen vertailua, jne....
- HUOM: Esim. kuvailutasolla palkkatietoja raportoidessa on järkevämpää käyttää keskilukuna mediaania keskiarvon sijasta

Muuttujan normaalijakauman testaus

■ Kolmogorov-Smirnov -testi

■ SPSS: Analyze /Descriptive statistics/Explore – normality plots with tests

→ Konservatiivinen testi eli hylkää helposti normaalijakaumaoletuksen... (vrt. Shapiro-Wilk)



Tests of Normality ^a						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
päätelytesti	,147	30	,095	,952	30	,191
verbaalinen testi	,149	30	,088	,962	30	,355
kielten keskiarvo	,075	30	,200*	,974	30	,665
matematiikan arvosana	,250	30	,000	,916	30	,021
opintomenestys	,140	30	,135	,936	30	,069



Aineiston tarkistaminen & poikkeavat havainnot

Tarkistus #1: Descriptives –komento

SPSS/Analyze/Descriptive Statistics/Descriptives

-valinnat mean, std.devation, minimum ja maximum

■ Joskus äärimmäiset arvot johtuvat laite-, mittaus-, koodaus tms. virheistä → **Arvojen korjaus oikeaksi, jos se on mahdollista** (esim. tarkistus alkuperäisistä lomakkeista)

■ Poistetaan datasta, raportoidaan

■ Havaintojen käyttäminen sellaisenaan, jos niille on järkevä selitys, mutta huomioidaan niiden vaikutus tuloksiin

■ Korvataan keskiarvolla

-kannattaa pohtia löytyykö juuri kyseistä puuttuvaa arvoa kuvaava keskiarvo SPSS/Analyze/Compare Means/Means



...mikä on tarpeeksi normaalia?

Testien rinnalla jakaumaa kannattaa tarkastella myös graafisesti. Testien oletukseksi riittää, kunhan jakauma on noin 'suurin piirtein' normaalin.

→ skewness & kurtosis tarkastelu

TOISAALTA: Vaikka "riittävä" normalisuus täyttyisikin tilastollisista oletuksista on hyvä olla tietoinen ja pohdiskella niiden mahdollista vaikutusta tuloksiin...

- ONKO REALISTA OLETTAA, ETTÄ KAIKKI PSYKOLOGISET ILMIÖT KUTEN ESIM. AHDISTUS, ONGELMAKÄYTTÄYTYMINEN, OLISIVAT KUTA KUINKIN NORMAALISTI JAKAUTUNEITA?



▪ Uudelleen koodaaminen (recode)

SPSS: Transform/recode

- Voidaan tehdä joko alkuperäiseen tai uuteen muuttujaan (mieluummin uuteen)
- Tarvitaan esim. muuttujan "suunnan" vaihtoon
- Tarvitaan myös jatkuvan muuttujan luokittelussa jne...

Esim. kengännumero luokitellaan kolmeen ryhmään

- *numero < 36: uusi arvo 1*
- *numero 37 - 39: uusi arvo 2*
- *numero > 39: uusi arvo 3*



SUMMAMUUTTUJIEN LUOMINEN

- SPSS: Transform / Compute:

$SUMMA = (KYS1 + KYS2 + KYS3 + KYS4 + KYS5)/5$

$SUMMA = MEAN(KYS1, KYS2, KYS3, KYS4, KYS5)$

- Suoran summamuuttujan muodostaminen edellyttää skaalan riittävää reliabiliteettia (esim. Cronbachin alfa > .70)
- Muista: Kysymysten (item) kääntäminen, niin että kaikki summamuuttujan osakysymykset ovat samansuuntaisia siten, että mitä suurempi arvo sitä enemmän se mittaa mitattavaa muuttujaa