

SISÄLTÖ

1 TILASTOJEN KÄYTTÖ	7
MITÄ TILASTOTIEDE ON?	7
TILASTO	7
TILASTOTIEDE	8
HISTORIAA	9
TILASTOTIETEEN NYKYINEN ASEMA	9
TILASTOLLISTEN MENETELMIEN ROOLIT	
ERI TYYPPISET AINEISTOT JA ONGELMAT	10
KYSELYTUTKIMUKSET	11
KOKEELLISET TUTKIMUKSET	11
LAADUNTARKKAILU	12
HAVAINNOINTITUTKIMUKSET	13
TILASTOLLISEN TUTKIMUKSEN VAIHEET	13
KVANTITATIIVISEN TUTKIMUKSEN VAIHEET	15
TUTKIMUKSEN LUOTETTAVUUDEN ARVIOINTI	16
TILASTOAINESTO	17
PERUSKÄSITTEITÄ	17
TILASTOAINESTON LUONNE	19
MITTAAMINEN	20
MITTAREIDEN LUOTETTAVUUS	23
TILASTOTIETOJEN HANKINTA	28
VALMIIT TILASTOT	28
AINEISTON KERÄÄMINEN	30
OTANTA	31
2 TILASTOJEN ESITTÄMINEN	38
TAULUKOINTI	38
JAKAUMATAULUKOT	39
LUOKITTELU	40
KAKSIULOTTEINEN TAULUKKO	42
TAULUKON ULKOASU JA MUOTOILU	44
TAULUKOINTI EXCELILLÄ	46
TAULUKOINTI SPSS-TILASTO-OHJELMALLA	51
GRAAFINEN ESITTÄMINEN	58
KAAVIOTYYPIT	59
KAIKILLE KAAVIOTYYPEILLE YHTEISIÄ OMINAISUUKSIA	60
VISUAALISIA NÄKÖKOHTIA	61
PALKKI- JA PYLVÄSKAAVIOT	64
YMPYRÄKAAVIO	71
VIIVAKAAVIO	72
ALUEKAAVIOT	74
HAJONTAKAAVIO ELI PISTEKAAVIO	75
TEEMAKARTAT	76
KAAVIOIDEN PIIRTÄMINEN EXCEL 2007:LLÄ	77
KAAVIOIDEN PIIRTÄMINEN SPSS-OHJELMISTOLLA	82

3 TUNNUSLUKUJA	87
SIJAINNILUKUJA	87
KESKIARVO	87
MEDIAANI	90
PROSENTTIPISTEET ELI FRAKTIILIT	91
MOODI ELI TYYPIARVO	93
HAJONTALUKUJA	95
VAIHTELUVÄLI	95
KVARTIILIVÄLI	96
KESKIHAJONTA	97
VARIANSSI	99
VARIAATIOKERROIN	99
STANDARDOITU MUUTTUJA	100
MUITA TUNNUSLUKUJA	101
VINOUS	101
HUIPUKKUUS	101
KESKIARVON LUOTTAMUSVÄLI	103
KESKIVIRHE.....	103
TUNNUSLUKUJEN GRAAFISIA ESITYKSIÄ	104
TUNNUSLUKUJEN LASKEMINEN EXCELILLÄ	106
TUNNUSLUKUJEN LASKEMINEN SPSS:LLÄ	112
4 TILASTOLLINEN RIIPPUVUUS	120
KORRELAATIO.....	120
RIIPPUVUUDEN LUONNE	120
RISTIINTAULUKOINTI JA KONTINGENSSIKERROIN	121
SPEARMANIN JÄRJESTYSKORRELAATIOKERROIN	122
HAJONTAKAAVIO JA PEARSONIN KORRELAATIOKERROIN	124
RIIPPUVUUDEN TUTKIMINEN EXCELILLÄ	130
RIIPPUVUUSMITTARIEN LASKEMINEN SPSS:LLÄ	132
REGRESSIO.....	136
LINEAARINEN REGRESSIOMALLI	136
REGRESSIOMALLIN MUODOSTAMINEN EXCELILLÄ	140
REGRESSIOMALLI SPSS:LLÄ	142
5 AIKASARJAT	145
TRENDI	146
AIKASARJAN VAIHTELUKOMPONENTIT	148
TRENDIN ARVIOINTI JA TASOITUS	150
KAUSIVAIHTELUT	153
INDEKSIT	155
YKSINKERTAINEN INDEKSI	155
RYHMÄINDEKSIT	157
AIKASARJAT JA EXCEL	161
6 TODENNÄKÖISYyslASKENTAA	166
KOMBINATORIIKKA	167
TULOPIERIAATE	167
KESKINÄISTEN JÄRJESTYSTEN LUKUMÄÄRÄ ELI PERMUTAATIO	169
VARIAATIO	170

KOMBINAATIO	171
TODENNÄKÖISYYS	175
TODENNÄKÖISYYDEN TILASTOLLINEN MÄÄRITTELY	175
TODENNÄKÖISYYDEN KLASSINEN MÄÄRITTELY	176
TODENNÄKÖISYYDEN YLEINEN MÄÄRITTELY	178
LASKUSÄÄNTÖJÄ	178
VASTATAPAHTUMAN TODENNÄKÖISYYS	179
YHTEENLASKUSÄÄNTÖ	180
KERTOLASKUSÄÄNTÖ	181
KOKONAISTODENNÄKÖISYYS JA BAYESIN KAAVA	184
TODENNÄKÖISYYSJAKAUMIA	192
TODENNÄKÖISYYSJAKAUMAN TUNNUSLUKUJA	194
EPÄJATKUVIA TODENNÄKÖISYYSJAKAUMIA	196
JATKUVIA TODENNÄKÖISYYSJAKAUMIA	200
TODENNÄKÖISYYDET EXCELILLÄ	207

7 TILASTOLLINEN PÄÄTTELY214

ESTIMOINTI	215
LUOTTAMUSVÄLI	215
TILASTOLLISET TESTIT	219
TESTAUKSEEN LIITTYVIÄ KÄSITTEITÄ	219
TESTAUKSEN PÄÄVAIHEET	222
JAKAUMAN NORMAALISUUDEN TUTKIMINEN	222
RIIPPUVUUDEN TESTAAMINEN	224
χ^2 -RIIPPUMATTOMUUSTESTI	224
χ^2 -YHTEENSOPIVUUSTESTI	226
KORRELAATIOKERTOIMEN TESTAUS	227
KESKIARVOTESTEJÄ	228
YHDEN OTOKSEN KESKIARVON T-TESTI	228
KAHDEN OTOKSEN KESKIARVOJEN T-TESTI	230
VARIANSSIANALYYSI	232
MUITA TESTEJÄ	234
TESTIN VALINTA	235
TAVALLISIMPIEN TESTIEN VALINTATAULUKKO	236
REGRESSIOMALLIN ARVIOIMINEN	237
TESTIT EXCELILLÄ	239
TESTIT SPSS:LLÄ	240

8 TEHTÄVIEN VASTAUKSIA245

LIITTEET

1	KUNNAT 2010	249
2	OPISKELIJA-AINEISTON (JYU) KYSELYLOMAKE	251
3	ERI MITTA-ASTEIKON MUUTTUIJILLE SOVELTUVAT TUNNUSLUVUT JA RIIPPUVUUSLUVUT.....	254
4	HAKUSANASTO.....	255

Edellisessä esimerkissä työssäkäyvien osuus 43,9 % on **kunnittaisten osuuksien painotettu keskiarvo** painoina väkiluvut. Kun painoja merkitään yleisesti w_i :llä ja lukuja, joiden keskiarvo lasketaan x_i :llä, niin painotettu keskiarvo on

$$\bar{x} = \frac{w_1 \cdot x_1 + w_2 \cdot x_2 + \dots + w_n \cdot x_n}{w_1 + w_2 + \dots + w_n} = \frac{\sum w_i x_i}{\sum w_i}$$

MEDIAANI

Mediaani, M_d , on **suuruusjärjestykseen** järjestettyjen havaintoarvojen **keskimmäinen** arvo, jos arvoja on pariton määrä. Jos arvoja on parillinen määrä, mediaanina pidetään yleensä kahden lähinnä keskimmäisen arvon keskiarvoa, joskus toista kahdesta lähinnä keskimmäisestä arvosta. **Mediaani on siis arvo, jota pienempiä tai yhtä suuria on 50 %** havainnoista. Vastaavasti puolet havainnoista on suurempia (tai yhtä suuria) kuin mediaani. Mediaani voidaan määrittää, mikäli muuttujan arvot voidaan panna suuruusjärjestykseen eli muuttuja on **vähintään järjestysasteikon muuttuja**.

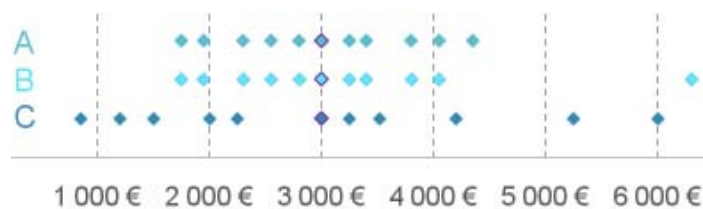
Esim. 3.3 Määritetään tammikuun 2010 viimeisen viikon lämpötilojen (esim. 3.1) mediaani.

Lämpötilat (C) ovat järjestyksessä pienimmästä suurimpaan

$$-19,3, -17,2, -16,9, -13,6, -10,6, -8,8, -3,7$$

Havaintoja on kaikkiaan seitsemän ja keskimäinen arvo on $-13,6$, joten lämpötilan mediaani on $-13,6$ C.

Alla on kuvattu kolmen ryhmän palkkoja. Siitä käy ilmi, että mediaani ei huomioi sitä, miten suuria tai pieniä sen eri puolille sijoittuvat havaintoarvot ovat. Ryhmiä A ja B verrattaessa näkyy, miten yksi havainto vaikuttaa keskiarvoon, mutta ei mediaaniin.

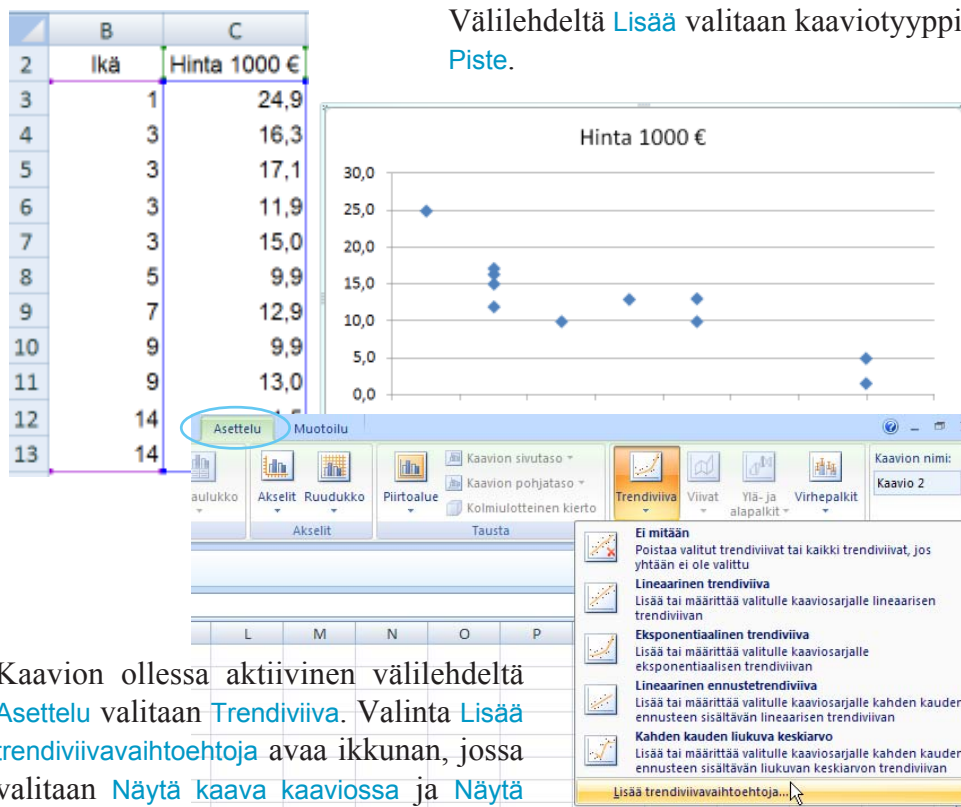


A:	$M_d = 3\ 000$	$\bar{x} \approx 3\ 018$
B:	$M_d = 3\ 000$	$\bar{x} \approx 3\ 195$
C:	$M_d = 3\ 000$	$\bar{x} \approx 3\ 002$

REGRESSIOMALLIN MUODOSTAMINEN EXCELILLÄ

Regressiosuoran kertoimet voidaan laskea eri tavoin. Yhden selittävän muuttujan malli on havainnollisinta ja helpointa muodostaa hajontakaavion (**Pistekaavio**) yhteyteen. Tällöin kuvioon saadaan suoran kuvaaja, sen yhtälö ja selitysaste. Hajontakaaviossa (**Piste-kaaviossa**) selittävän muuttujan tulee sijaita x-akselilla ja selitettävän (riippuvan). muuttujan y-akselilla.

Esim. 4.14 Muodostetaan käytettyjen autojen hinnan riippuvuutta iästä kuvaavan regressiosuoran yhtälö hajontakaavioon. Kaaviossa selittävä muuttuja auton ikä on x-akselilla ja selitettävä eli riippuva muuttuja hinta y-akselilla. Valmiiseen kaavioon lisätään regressiosuora ja sen yhtälö.



Kaavion ollessa aktiivinen välilehdeltä **Asettelu** valitaan **Trendiviiva**. Valinta **Lisää trendiivivaihtoehtoja** avaa ikkunan, jossa valitaan **Näytä kaava kaaviossa** ja **Näytä korrelaatiokertoimen arvo kaaviossa**.

Aseta leikkauspiste = 0,0

Näytä kaava kaaviossa

Näytä korrelaatiokertoimen arvo kaaviossa

Antaa nimityksestä huolimatta selitysasteen r^2 .



- 3-26** Suomalaisien arvoja mittaavassa otantatutkimuksessa kysyttiin mm. miten hyväksyttävänä he pitivät eri asioita (asteikko 1–10, 1 = ei koskaan hyväksyttävä, ..., 10 = aina hyväksyttävä). Aineistosta saatiin seuraavat keskiarvot ja luottamusvälit (Lähde: Yhteiskuntatieteellinen tietoaarkisto):

	Keskiarvo	95 %:n luottamusväli
Abortti	5,11	keskiarvo \pm 0,19
Avioero	6,52	keskiarvo \pm 0,17
Eutanasia	6,04	keskiarvo \pm 0,20
Homoseksuaalisuus	4,54	keskiarvo \pm 0,22
Prostituutio	3,54	keskiarvo \pm 0,18

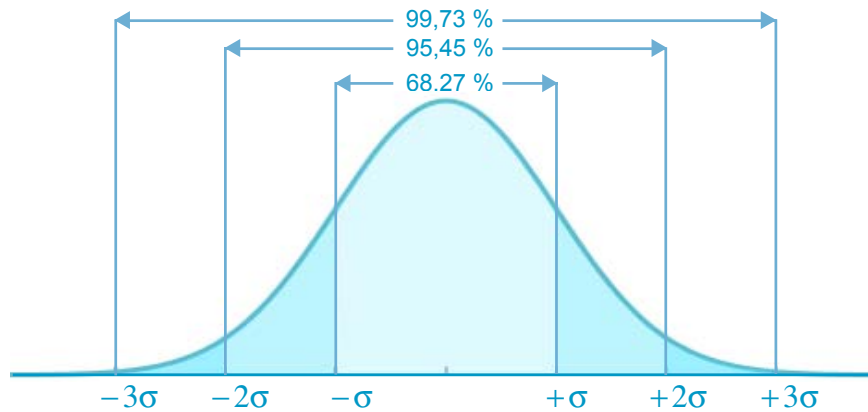
Havainnollista keskiarvot luottamusväleineen samassa kaaviossa.

- 3-27** Kuvaile tyypillinen Jyu:n opiskelija.
- 3-28** Laske Opiskelija-aineiston (Jyu) muuttujan Opintotuki
a) kvartiiliväli
b) kvartiilipoikkeama
ja selitä niiden antama informaatio.
- 3-29** Tee tunnuslukujen yhteenveto (tunnuslukutiivistelmä) Opiskelija-aineiston (Jyu) muuttujista Opintotuki ja Ansiotulot
a) ottamalla mukaan kaikki havainnot
c) ottamalla mukaan vai nollasta poikkeavat havainnot.
Vertaile jakaumia.
- 3-30** Laske Opiskelija-aineiston (Jyu) muuttujan Nukkuminen keskiarvot ja keskihajonnat
a) sukupuolittain
b) sukupuolittain parisuhteittain
- 3-31** Laske Opiskelija-aineiston (Jyu) muuttujien Opiskelu ja Liikunta keskiarvot, keskihajonnat ja variaatiokertoimet
a) sukupuolittain
b) tiedekunnittain.
Tulkitse tunnuslukujen antama informaatio.
- 3-32** Havainnollista graafisesti Opiskelija-aineiston (Jyu) muuttujan Opiskelu keskiarvot tiedekunnittain. Piirrä kaavioon myös 95 %:n luottamusvälit.
- 3-33** Millaisia eroja mies- ja naisopiskelijoiden välillä näyttäisi olevan Opiskelija-aineiston (Jyu) perusteella
a) välivuosien lukumäärissä
b) suhtautumisessa hyviin harrastusmahdollisuuksiin?

Koska tiheysfunktion kuvaaja on symmetrinen odotusarvon suhteen, niin

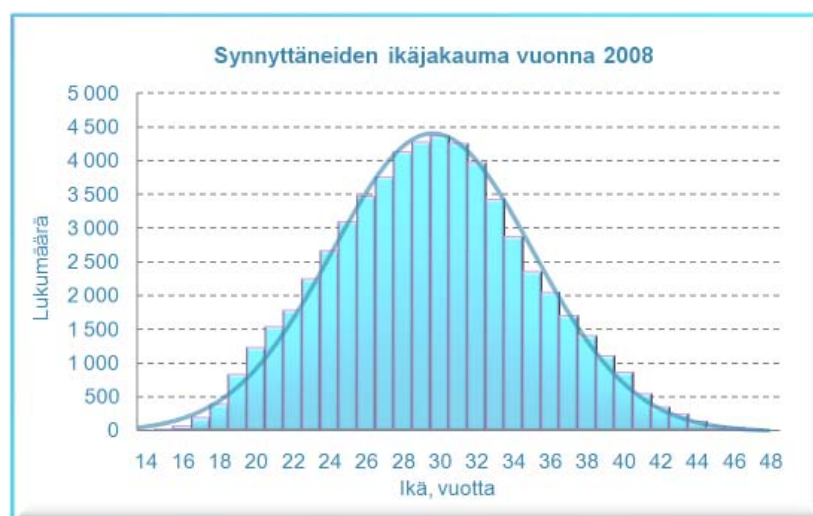
$$P(\underline{x} \leq \mu) = 0,5 \quad \text{ja} \quad P(\underline{x} \geq \mu) = 0,5$$

Todennäköisyysmassa on keskittynyt odotusarvon ympärille seuraavan kuvion mukaisesti:



Kuvion mukaan 68,27 % muuttujan arvoista poikkeaa odotusarvosta korkeintaan keskihajonnan verran suuntaan tai toiseen ja 99,73 % muuttujan arvoista on korkeintaan 3 keskihajonnan mitan päässä odotusarvosta.

Esim. 6.35 Synnyttäjien ikäjakauma vuonna 2008 (lähde: THL/ Syntymärekisteri) oli oheisen kaavion mukainen. Jakauma noudattaa likimain normaalijakaumaa siten, että odotusarvo on 30 vuotta ja keskihajonta 5,4 vuotta.



Histogrammiin on lisätty normaalijakauman kuvaaja.

Ensiksi tutkitaan varianssien yhtäsuuruutta. SPSS-tilasto-ohjelma antaa t-testin yhteydessä Levenen varianssien yhtäsuuruustestin tuloksen. Tässäkin testissä nollahypoteesina on, että varianssit ovat yhtäsuuret.

		Levene's Test for Equality of Variances	
		F	Sig.
Internet	Equal variances assumed	3,780	,052
	Equal variances not assumed		

Koska p-arvo (Sig.) on 0,052, niin nollahypoteesi jää voimaan (mikäli rajana pidetään 5 %:a) ja varianssit voidaan siis olettaa yhtä suuriksi (tällöin t-testin tulos luetaan ylempältä riviltä).

T-testin tulosteet SPSS-tilasto-ohjelmalla:

t-test for Equality of Means						
t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
					Lower	Upper
2,565	473	,011	,4715	,1838	,1103	,8326
2,377	252,915	,018	,4715	,1983	,0809	,8620

Excelillä:

Kahden otoksen t-testi olettaen varianssit yhtäsuuriksi		
	Mies	Nainen
Keskiarvo	3,006	2,535
Varianssi	4,649	2,986
Havainnot	155	320
t Tunnusluvut	2,565	
P(T<=t) kaksisuuntainen	0,011	
t-kriittinen kaksisuuntainen	1,965	

Koska p-arvo on 0,011, niin nollahypoteesi hylätään. Nais- ja miesopiskelijoiden internetin käytön keskiarvoissa esiintyvä ero on todellinen eikä ilmeisestikään johdu sattumasta. Koko opiskelijajoukossa miehet näyttävät käyttävän enemmän aikaa internetin selailuun kuin naiset.

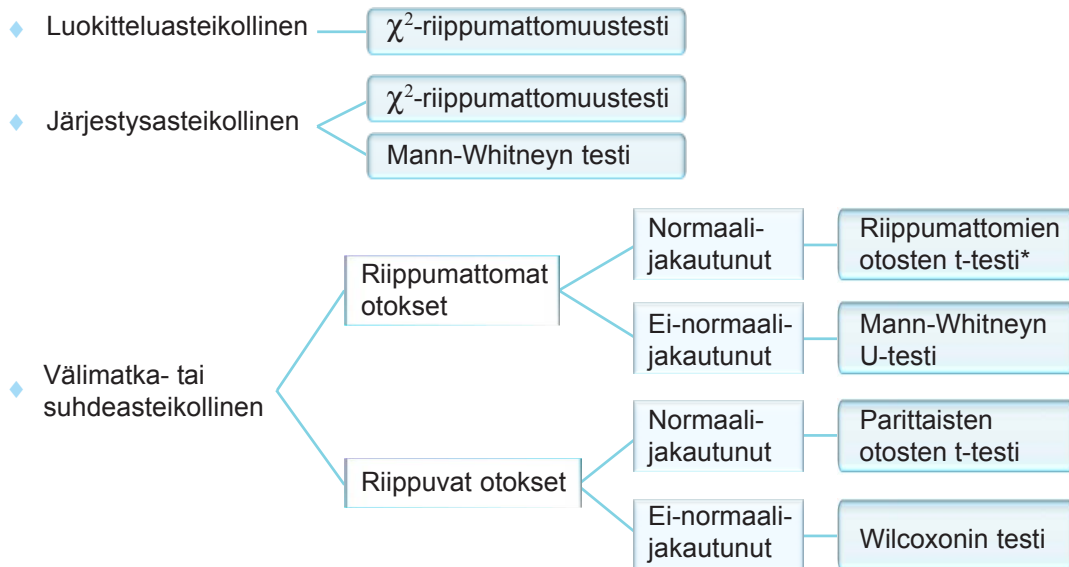
Taulukosta käy lisäksi ilmi, että miesten ja naisten keskiarvojen erotus on noin 0,47 (Mean Difference). Erotuksen 95 %:n luottamusvälin alaraja on 0,11 ja yläraja 0,83, joten perusjoukossa miesten ja naisten keskiarvojen erotus on 95 %:n varmuudella välillä (0,11, 0,83).

TAVALLISIMPIEN TESTIEN VALINTATAULUKKO

VERRATAAN RYHMIEN SAMANLAISUUTTA

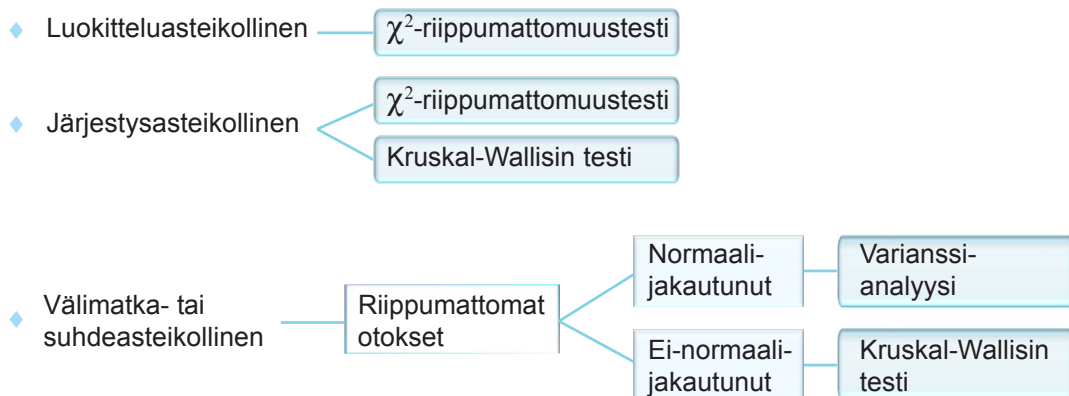
Ryhmittelvä muuttuja luokitteluasteikollinen, 2 ryhmää

Verrattava muuttuja



Ryhmittelvä muuttuja luokitteluasteikollinen, ryhmiä 3 tai enemmän

Verrattava muuttuja



TARKASTEELLAAN KAHDEN MUUTTUJAN RIIPPUVUUTTA/KORRELAATIOITA

- ◆ Ainakin toinen muuttuja luokitteluasteikollinen — χ^2 -riippumattomuustesti
- ◆ Toinen muuttuja järjestysasteikollinen, toinen järjestys-, välimatka- tai suhdeast. — (Spearmanin) järjestyskorrelaatiokertoimen merkitsevyydestesti
- ◆ Molemmat muuttujat välimatka- tai suhdeast. — Pearsonin korrelaatiokertoimen merkitsevyydestesti

* t-testi sen mukaan ovatko varianssit yhtä suuret vai eri suuret