

Kandidaatintutkielman aineistonhankinta ja analyysi

Anna-Kaisa Ylitalo

M 315, anna-kaisa.ylitalo@jyu.fi
Musiikin, taiteen ja kulttuurin tutkimuksen laitos
Jyväskylän yliopisto

2018

Havaintomatriisi

Havaintomatriisi sisältää kaikki tutkittavat havaintoyksiköt ja muuttujat

Henkilö	Sukupuoli	Ikä	Bruttopalkka (€ / kk)	Siviilisäät
1	mies	25	2700	naimisissa
2	nainen	43	2100	naimisissa
3	nainen	63	2350	leski
4	nainen	32	3560	naimaton
5	mies	45	4320	naimisissa
6	mies	18	0	naimaton

Havaintoyksikkö

Havainto

Muuttuja

Aineiston kuvailu

Tilastolliset tunnusluvut

Tutkimusaineistoa pyritään yleensä kuvailemaan ja tiivistämään *tilastollisten tunnuslukujen* avulla.

Tarkoituksena on kuvailla mitattujen muuttujien keskeiset ominaisuudet mahdollisimman selkeässä muodossa.

Eri mitta-asteikolla mitattuja muuttujia voidaan kuvata erilaisten tunnuslukujen avulla.

Tarkasteltaessa yhden muuttujan jakaumaa, oleelliset tunnusluvut ovat **keskiluvut** ja **hajontaluvut**.

Keskiluvut

Keskiluvut ilmoittavat muuttujan keskimääräisen, tyypillisen tai yleisimmän arvon.

Tavoitteena on siis määritellä muuttujan jakauman keskikohta.

Keskilukuja ovat

- ▶ Moodi (*mode*)
- ▶ Mediaani (*median*)
- ▶ Keskiarvo (*mean*)

Hajontaluvut

Hajontaluvut kuvaavat sitä, kuinka laajalle tai suppealle välille havainnot sijoittuvat.

Hajontalukuja ovat mm.

- ▶ Vaihteluväli ja kvartiiliväli (*range, interquartile range*)
 - ▶ Järjestysasteikollisille muuttujille
- ▶ Keskihajonta ja varianssi (*standard deviation, variance*)
 - ▶ Välimatka-asteikollisille muuttujille

Frekvenssitaulukko (luokitelluille muuttujille)

Frekvenssijakauma (*frequency distribution*) voidaan esittää **frekvenssitaulukkona**.

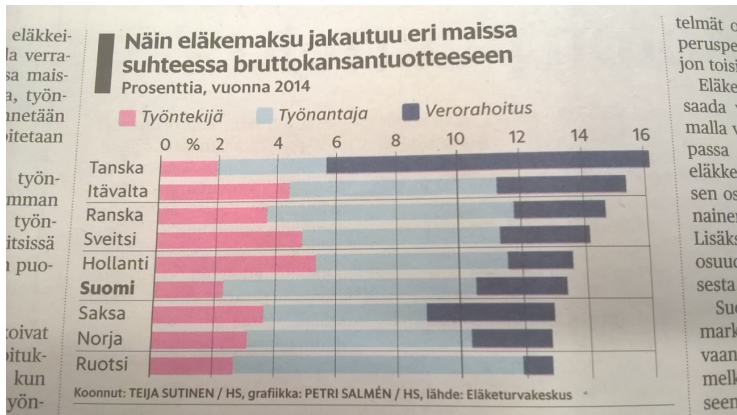
Frekvenssitaulukossa esitetään usein myös **prosentuaaliset frekvenssit** (kuinka monta prosenttia havainnoista kuuluu kyseiseen luokkaan).

Pääaine	frekv	%
Musiikkikasvatus	10	38
Musiikkitiede	15	58
Muu	1	4
Yhteensä	26	100

Graafinen kuvaaminen

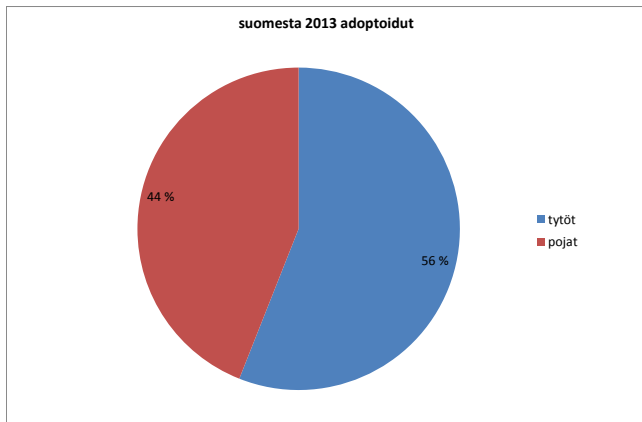
Pylväsdiagrammi (*bar graph*)

Pylväsdiagrammi (tai pylväskuvaaja) kuvaa laadullisen muuttujan arvojen frekvenssijakaumaa (joskus myös epäjatkuvan määrällisen muuttujan).



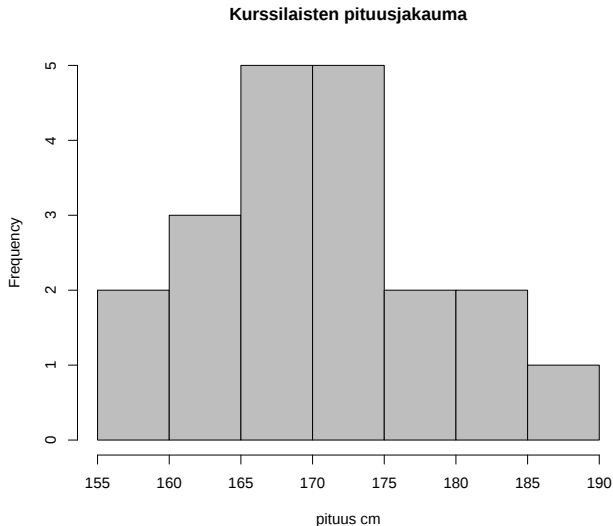
Piirakkakuvio (*pie chart*)

Pylväsdiagrammin sijaan voi joskus olla parempi käyttää *piirakkakuvio*a (ympyrädiagrammi), esimerkiksi kun halutaan havainnollistaa jonkin kokonaisuuden jakautumista osiin.



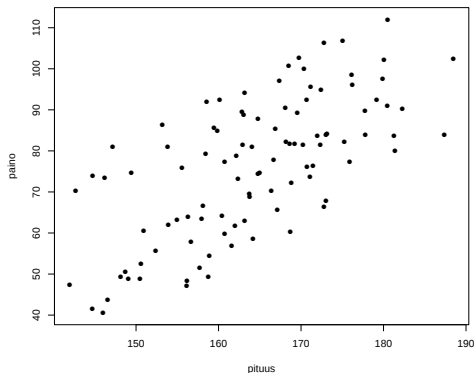
Histogrammi (*histogram*)

Histogrammilla kuvataan määrällisen **jatkuvan** muuttujan frekvenssijakaumaa.



Hajontakuvio (*scatter plot*)

Hajontakuvio (tai *sirontakuvio*) on koordinaatistoon sijoitettu pistekuvio, jonka avulla havainnollistetaan kahden muuttujan välistä riippuvuutta.



Analyysimenetelmistä

Ristiintaulukko (*crosstabulation*)

Kahden muuttujan välisiä riippuvuuksia voidaan tutkia *ristiintaulukon* avulla.

Ristiintaulukko (eli kontingenssitaulu) on siis kaksiulotteinen frekvenssitaulu.

Muuttujien täytyy tässäkin tapauksessa olla luokiteltuja.

Yhdessä solussa olevaa havaintojen lukumäärää kutsutaan **solufrekvenssiksi**.

Ristiintaulukko muuttujille sukupuoli ja opintopisteet

		sukupuoli		
		mies	nainen	yht.
op	0-19	3	0	3
	20-39	4	2	6
	40-59	9	5	14
	60-79	2	4	6
	80-99	0	1	1
yht.		18	12	30

Sukupuoli-muuttujan frekvenssijakauma (reuna- tai marginaalijakauma)

Opintopiste-muuttujan frekvenssijakauma (reuna- tai marginaalijakauma)

Suhteelliset jakaumat

Muuttujien välistä riippuvutta voi tarkastella myös suhteellisten (tai prosentuaalisten) frekvenssien kautta, tällöin ryhmien kokoerot eivät häiritse.

Lasketaan siis frekvenssien prosentuaaliset osuudet joko sarakkeittain (*sarakeprosentit*) tai riveittäin *riviprocentit*.

Jos muuttujat riippuvat toisistaan, tämä näkyy erilaisina sarakeprosentteina (tai riviprocentteina).

Mikäli muuttujat eivät riipu toisistaan, sarakeprosentit (tai riviprocentit) eivät eroa juurikaan toisistaan.

Esimerkki, sarakeprosentit

	sukupuoli		yht.
	mies	nainen	
0-19	$\frac{3}{18} * 100 = \mathbf{17}$	$\mathbf{0}$	$\frac{3}{30} * 100 = \mathbf{10}$
20-39	$\frac{4}{18} * 100 = \mathbf{22}$	$\frac{2}{12} * 100 = \mathbf{17}$	$\frac{6}{30} * 100 = \mathbf{20}$
op 40-59	$\frac{9}{18} * 100 = \mathbf{50}$	$\frac{5}{12} * 100 = \mathbf{42}$	$\frac{14}{30} * 100 = \mathbf{47}$
60-79	$\frac{2}{18} * 100 = \mathbf{11}$	$\frac{4}{12} * 100 = \mathbf{33}$	$\frac{6}{30} * 100 = \mathbf{20}$
80-99	$\mathbf{0}$	$\frac{1}{12} * 100 = \mathbf{8}$	$\frac{6}{30} * 100 = \mathbf{3}$
yht.	$\mathbf{100}$	$\mathbf{100}$	$\mathbf{100}$

χ^2 -riippumattomuustesti (*Chi-squared test, Fisher's exact test*)

χ^2 -riippumattomuustesti (khiin neliö -testi) testataa muuttujien välistä riippuvuutta

- ▶ Kahden luokitteluasteikollisen muuttujan testi
- ▶ vrt. ristiintaulukko

Korrelaatio (*correlation*)

Korrelaatio mittaa kahden muuttujan välisen riippuvuuden voimakkuutta (vrt. hajontakuvio).

- ▶ Pearsonin korrelaatiokerroin
 - ▶ Kaksi välimatka-asteikollista muuttujaa
 - ▶ Soveltuu lineaarisen riippuvuuden voimakkuuden mittaamiseen
- ▶ Spearmanin järjestyskorrelaatiokerroin
 - ▶ Kaksi järjestysasteikollista muuttujaa
 - ▶ Soveltuu lineaarisen ja monotonisen riippuvuuden mittaamiseen

HUOM! Muuttujien välinen korrelaatio ei välttämättä tarkoita sitä, että muuttujien välillä olisi syy-seuraus-suhde.

T-testit

T-testeillä voidaan verrata kahden ryhmän keskiarvoja (populaatioiden odotusarvoja).

- ▶ Vaatii välimatka-asteikollisen muuttujan, joka on normaalisti jakautunut (populaatiossa)
- ▶ Riippumattomien otosten t-testi
 - ▶ Ryhmät ovat keskenään riippumattomia
- ▶ Riippuvien otosten t-testi
 - ▶ Ryhmät riippuvat toisistaan (esim. 2 mittauskertaa samoille henkilöille)

Varianssianalyysi (ANOVA)

Varianssianalyysillä voidaan tutkia eroavatko ryhmien (>2) keskiarvot tilastollisesti merkitsevästi toisistaan.

- ▶ Vastemuuttujan oltava välimatka-asteikollinen (ja normaalisti jakautunut kussakin perusjoukossa)
- ▶ Selittävät muuttujat (faktorit) luokitteluasteikollisia
- ▶ Havainnot toisistaan riippumattomia (sekä ryhmien välillä että niiden sisällä)
- ▶ Jos eroja löytyy, täytyy tehdä parittaiset vertailut (post hoc)
- ▶ kts. myös toistomittausten varianssianalyysi

Parametrittomat testit

Mikäli testien oletukset (esim. normalisuus, välimatka-asteikollisuus) eivät ole voimassa tai havaintoja on vähän, voi käyttää parametrittomia testejä.

- ▶ Riippumattomien otosten t-testi -> Mann-Whitneyn testi
- ▶ Riippuvien otosten t-testi -> Wilcoxonin testi
- ▶ ANOVA -> Kruskal-Wallis test
- ▶ toistomittausten ANOVA -> Friedmanin testi

Selecting Statistical Tests

Measurement Scale of the Dependent Variable	One Independent Variable				Two Independent Variables	
	Two Levels		More than 2 Levels		Factorial Designs	
	Two Independent Groups	Two Dependent Groups	Multiple Independent Groups	Multiple Dependent Groups	Independent Groups	Dependent Groups
Interval or Ratio	Independent t-test	Dependent t-test	One-Way ANOVA	Repeated Measures ANOVA	Two -Factor ANOVA	Two-Factor ANOVA Repeated Measures
Ordinal	Mann-Whitney U	Wilcoxon	Kruskal-Wallis	Friedman		
Nominal	Chi-Square		Chi-Square		Chi-Square	