Faktorianalyysi ja survey-aineiston tiivistäminen

Kimmo Vehkalahti

yliopistonlehtori, soveltavan tilastotieteen dosentti
Helsingin yliopisto, sosiaalitieteiden laitos
http://www.helsinki.fi/people/Kimmo.Vehkalahti/suomeksi.html

Sosiaalitutkimuksen tilastolliset menetelmät, jakso 2
29.–30.1.2013

Jakso 2: tavoitteet ja sisältö

Jakson tavoitteena on oppia analysoimaan kyselytutkimus- eli survey-aineistoa faktorianalyysilla sekä sen perusteella tiivistämään aineistoa jatkoanalyyseja varten.

Tällaiset aineistot ovat tyypillisiä mm. yhteiskuntatieteellisessä tutkimuksessa, jossa mielenkiinto kohdistuu mm. asenteisiin, arvoihin ja mielipiteisiin. Näitä moniulotteisia ilmiöitä mitataan kysely- ja haastattelulomakkeilla.

1. **Faktorianalyysin perusteet**
   - Mittausmalli ja korrelaatiot
   - Faktorianalyysi ja sen tulkinta
2. **Survey-aineiston tiivistäminen**
   - Summamuuttujat
   - Faktoripisteet

1. Faktorianalyysin perusteet

Tutkimuskysymyksistä johdettu mittausmalli ohjaa sekä kyselylomakkeen laatimista että lomakkeen pohjalta kerätyn survey-aineiston analysointia.

- **mittausmalli: mitä** mitataan, **millä** ja **miten**
- mittausinstrumentti: **millä** mitataan ja **miten**
- kysely- ja haastattelututkimuksen instrumentti: **lomake**
- **mittari:** kokoelma saman aihepiirin mittauksia

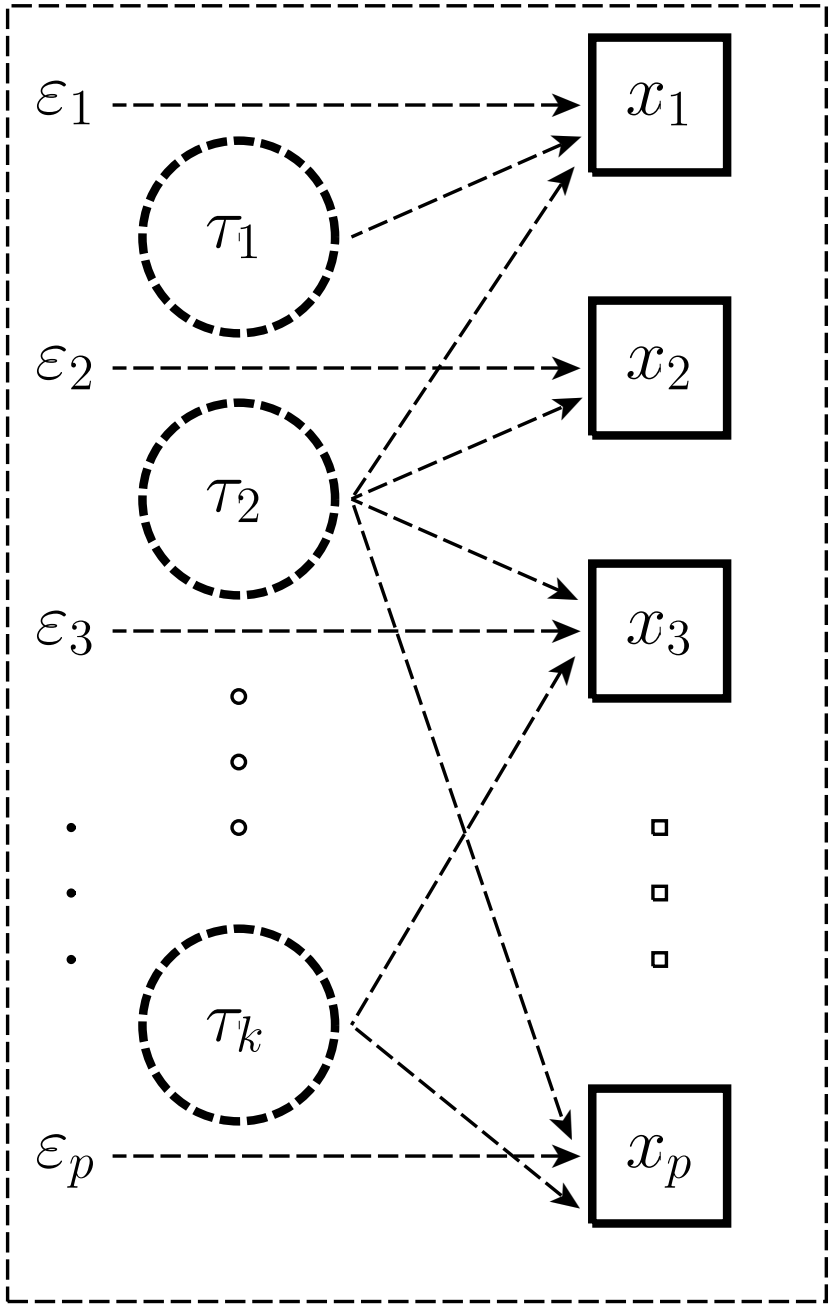
Mittauksen laatu:

1. **validiteetti:** mitataanko sitä mitä pitikin?
2. **reliabiliteetti:** mitataanko riittävän tarkasti?

Mittaustaso vaikuttaa menetelmävalikoimaan:

- luokittelu — järjestäminen — numeerinen mittaus
- useimmat menetelmät edellyttävät numeerista mittausta
- joissakin menetelmissä luokittelukin voi riittää
- järjestystasoinen mittaus usein ongelmallisinta

Mittausmalli (vrt. johdantokurssi, ks. seuraava sivu)



Mittausmalli ja korrelaatiot

Mittausmallissa esiintyy kolmenlaisia käsitteitä:

1. **faktoreita** τ (k kpl), jotka vastaavat ilmiön ulottuvuuksia
2. **muuttujia** x (p kpl), jotka on mitattu (kyselylomakkeella)
3. **mittausvirheitä** ε (p kpl), joita ei voi kokonaan välttää

Paras käsitys faktoreista ja niiden lukumäärästä sekä malliin soveltuvista muuttujista pohjautuu **tutkimuskysymyksiin** ja **tutkimusalan teoriaan**. Lukumäärälle pätee: $k < p$.

Faktoreiden ja muuttujien välisillä nuolilla osoitetaan niiden välisiä yhteyksiä, joita kuvataan **korrelaatioiden** avulla. Nuolet: faktorin ajatellaan vaikuttavan siihen, miten kysymykseen vastataan.

Kaikki katkoviivoilla esitetty on *hypoteettista*, koska aineisto koostuu vain mitatuista muuttujista. Tavoite on tiivistää aineistoa muodostamalla uusia muuttujia, jotka vastaavat mielenkiinnon kohteena olevia (hypoteettisia) faktoreita.

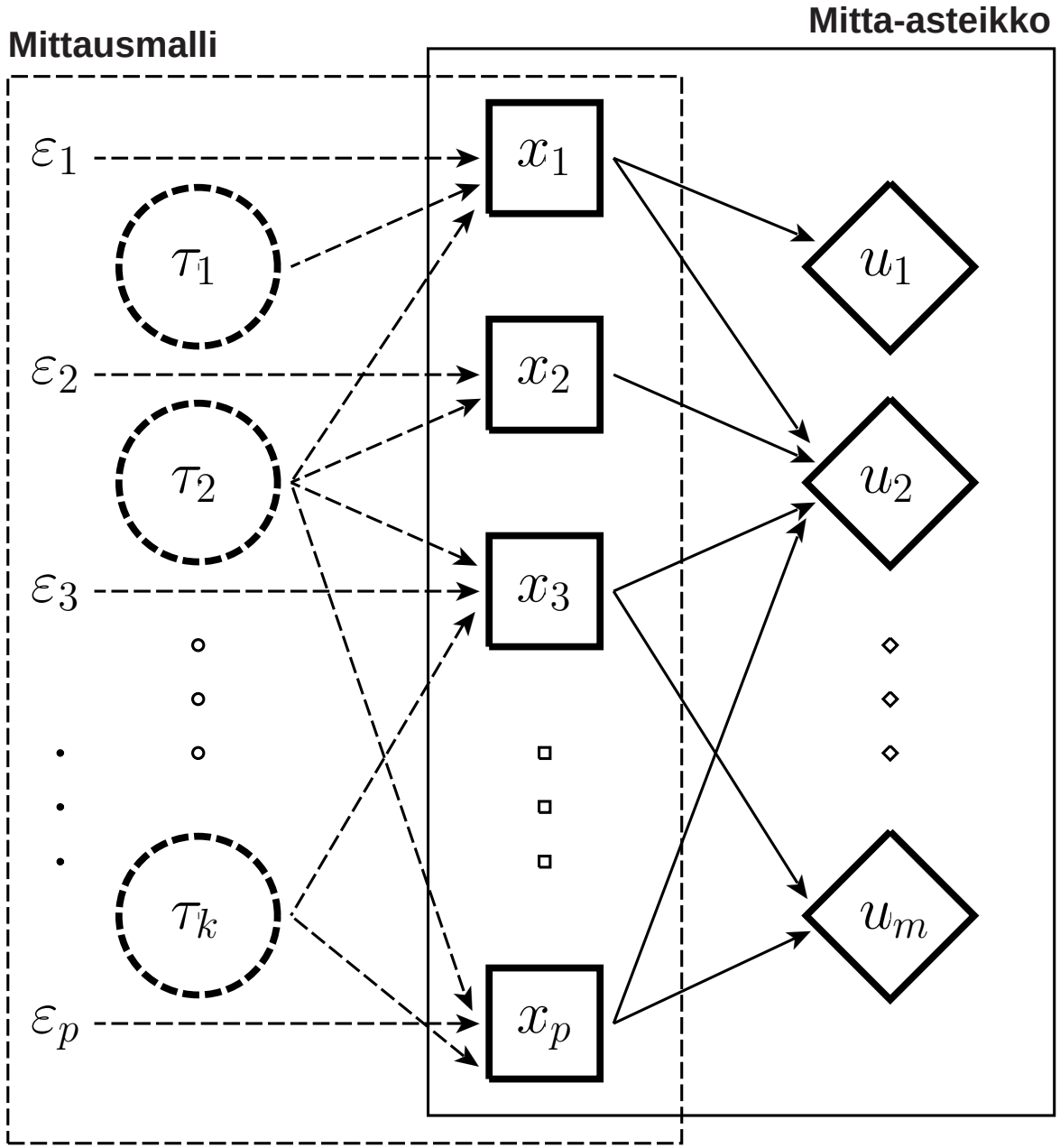
Faktorianalyysin vaiheet

Faktorianalyysi on monimuuttujamenetelmä, ts. siinä käsitellään yhtaikaa useita muuttujia. Muuttujien oletetaan mittaavan samaa, tyypillisesti moniulotteista ilmiötä (kuten asenteet, arvot jne.). Tämänkaltaiset ilmiöt ovat *latentteja*, eli niitä ei voida suoraan mitata, vaan joudutaan käyttämään epäsuoria keinoja, siis useita kysymyksiä tai väitteitä.

Faktorianalyysissa on kaksi vaihetta: **vaiheen 1** tavoitteena on hahmottaa aineiston **taustalla oleva rakenne**, jota voidaan etukäteen kuvata mittausmallilla. Lähestymistapoja ovat *eksploratiivinen* eli aineistoperustainen ja *konfirmatorinen* eli malliperustainen faktorianalyysi. Käytännön analyysit ovat näiden väliltä eli niissä on tyypillisesti piirteitä molemmista.

Faktorianalyysin **vaiheen 2** myötä **aineistoa tiivistetään** vaiheen 1 tulosten ja tutkimusalan teorian tuella.

Mittausmallista aineiston tiivistämiseen (faktorianalyysi)



Mittausmalli ja mitta-asteikko

Mittausmalli (luento 1)

- Mitä **ilmiötä** tutkitaan? Montako **ulottuvuutta** siinä on?
- Millä ilmiötä **mitataan** — mahdollisimman hyvin?
- Yhteyksien ja ulottuvuuksien tutkiminen: **faktorianalyysi**
- Esimerkki: ESS (European Social Survey)

Mitta-asteikko (luento 2)

- Osioiden eli mitattujen muuttujien yhdistelmä
- Yhdistelytapoja: **faktoripisteet**, **summamuuttujat**
- Yleinen tavoite: **aineiston tiivistäminen**
- Esimerkki: ESS (European Social Survey)

Esimerkki: faktorianalyysi (ESS)

ESS (European Social Survey), round 5, 2010, Suomi (n=1878)

Tarkastellaan tässä vain seuraavia muuttujia:

- How interested in politics (**1=very**, ..., 4=not at all)

Seuraavissa asteikko 0=no trust at all, ..., 10=complete trust:

- Trust in country's parliament
- Trust in the legal system
- Trust in the police
- Trust in politicians
- Trust in political parties
- Trust in the European Parliament
- Trust in the United Nations

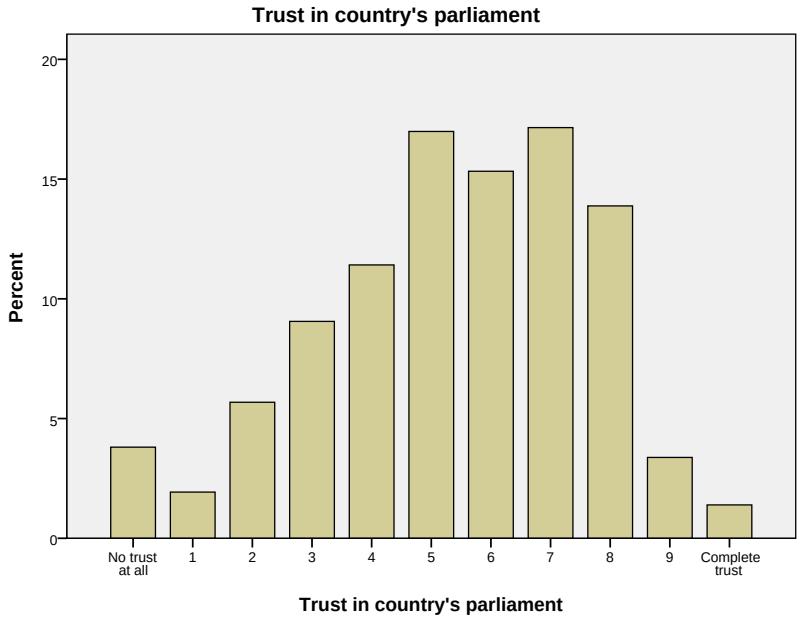
Seuraavissa asteikko 0–10 eri sanamuodoin (10=positiivisin):

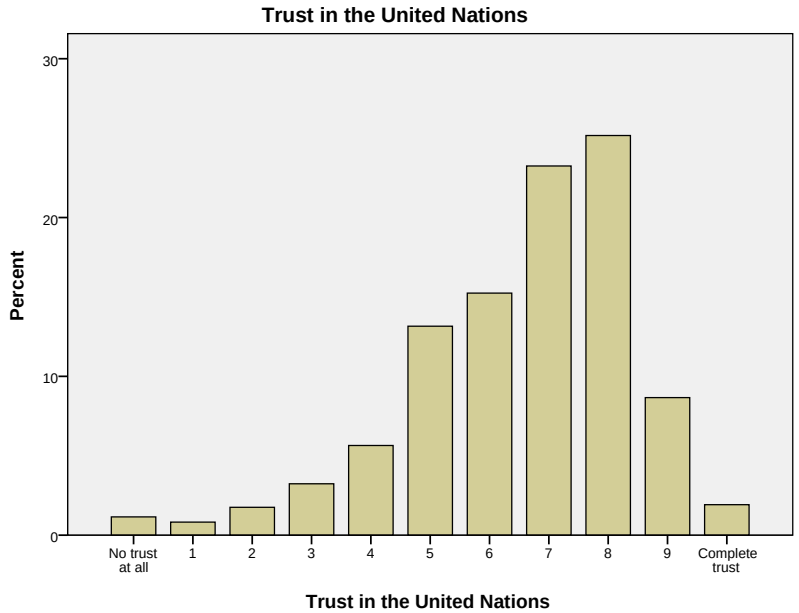
- Most people can be trusted or you can't be too careful
- Most people try to take advantage of you, or try to be fair
- Most of the time people are helpful or mostly looking out for themselves

(Kysymysten tarkemmat sanamuodot on tässä sivuutettu.)

Esimerkki: faktorianalyysin oletukset

Muuttujien jakaumat vaihtelevat, olennaista on mittaustaso:





Oletetaan nyt, että olisi hahmoteltu **3 faktorin** mittausmalli (ulottuvuuksina mm. luottamus poliittisiin toimijoihin ja ihmisiin).

Tehdään seuraavaksi tältä pohjalta faktorianalyysi SPSS:llä.

(Esimerkkitilanne on tarkoituksella valittu suppeaksi ja osin keinotekoiseksi, koska fokus on enemmän menetelmässä ja sen tulkinnoissa. Aineisto mahdollistaa sisällöllisesti paljon monipuolisempiakin asetelmia!)

Esimerkki: faktorianalyysin suoritus SPSS:llä

Faktorianalyysi tiivistää tiedot muuttujien välisistä korrelaatioista annetuksi määräksi faktoreita.

* faktorianalyysi 3-ulotteisen mittausmallin pohjalta (SPSS):

```
FACTOR
/VARIABLES polintr trstprl trstlgl trstplc trstplt trstprt trstep
trstun ppltrst pplfair pplhlp
/MISSING LISTWISE
/ANALYSIS polintr trstprl trstlgl trstplc trstplt trstprt trstep trstun
ppltrst pplfair pplhlp
/PRINT INITIAL EXTRACTION ROTATION
/FORMAT SORT
/CRITERIA FACTORS(3) ITERATE(25)
/EXTRACTION ML
/CRITERIA ITERATE(25)
/ROTATION VARIMAX.
```


Esimerkki: faktorimatriisi ja faktorilataukset

Faktorianalyysin olennaisin tuloste on ns. rotatoitu faktorimatriisi:

Rotated Factor Matrix[a]

	Factor		
	1	2	3
Trust in politicians	,895	,249	,140
Trust in political parties	,876	,237	,135
Trust in country's parliament	,742	,205	,301
Trust in the European Parliament	,703	,187	,213
Trust in the United Nations	,452	,210	,368
How interested in politics	-,212	-,005	-,088
Most people can be trusted or you can't be too careful	,160	,720	,183
Most people try to take advantage of you, or try to be fair	,099	,686	,153
Most of the time people helpful or mostly looking out for themselves	,195	,582	,104
Trust in the legal system	,378	,214	,745
Trust in the police	,177	,206	,651

Extraction Method: Maximum Likelihood.
Rotation Method: Varimax with Kaiser Normalization.

(järjestettynä faktorilatausten perusteella faktoreittain)

Esimerkki: kommunaliteetit

On syytä tarkastella myös muuttujien kommunaliteetteja:

Communalities

	Initial	Extraction
How interested in politics	,055	,053
Trust in country's parliament	,656	,683
Trust in the legal system	,538	,743
Trust in the police	,392	,498
Trust in politicians	,801	,883
Trust in political parties	,776	,841
Trust in the European Parliament	,619	,575
Trust in the United Nations	,459	,383
Most people can be trusted or you can't be too careful	,397	,577
Most people try to take advantage of you, or try to be fair	,345	,504
Most of the time people helpful or mostly looking out for themselves	,299	,387

Extraction Method: Maximum Likelihood.

(valitettavasti SPSS antaa ne eri järjestyksessä eri taulukkoon!)

Esimerkki: tulosten esittäminen ja tulkinta

Factor structure of ESS (Finland), $n = 1878$	F1	F2	F3	h^2
F1: **Trust in Political Actors**				
Trust in politicians	**0.90**	0.25	0.14	0.88
Trust in political parties	**0.88**	0.24	0.14	0.84
Trust in country's parliament	**0.74**	0.21	*0.30*	0.68
Trust in the European Parliament	**0.70**	0.19	0.21	0.58
Trust in the United Nations	*0.45*	0.21	*0.37*	0.38
How interested in politics	-0.21	-0.01	-0.09	0.05
F2: **Trust in People**				
Most people can be trusted	0.16	**0.72**	0.18	0.58
Most people try to take advantage of you	0.10	**0.69**	0.15	0.50
Most of the time people are helpful	0.20	**0.58**	0.10	0.39
F3: **Trust in Authorities**				
Trust in the legal system	*0.38*	0.21	**0.75**	0.74
Trust in the police	0.18	0.21	**0.65**	0.50
Sum of squares	3.11	1.66	1.36	6.13
Variance explained %	28.3	15.0	12.4	55.7

h^2 = communalities

Esimerkki: faktorianalyysin tulkinnoista

(Taulukoita ja tulkintoja katsastellaan tarkemmin luennolla.)

Edeltä tulostaulukoista ilmenee, että kolme faktoria ”selittää” yhteensä 55.7 % kyseisten muuttujien välisestä vaihtelusta. Olennaisempia ovat faktorien ja muuttujien väliset yhteydet (korrelaatiot), joita tulkitaan faktorilatausten avulla *(vrt. nuolet)*.

Ensimmäinen faktori (”luottamus poliittisiin toimijoihin”?) on voimakkain. Toinen voisi olla nimeltään ”luottamus ihmisiin” ja kolmas ”luottamus auktoriteetteihin”. Mitä enemmän nojataan tutkimusalan teoriaan, sitä helpompi on nimetä faktorit. (Teoria saattaa myös tukea *korreloivia faktoreita*; tässä ne on oletettu korreloimattomiksi, mikä on ainakin alkuvaiheessa selkeämpää.)

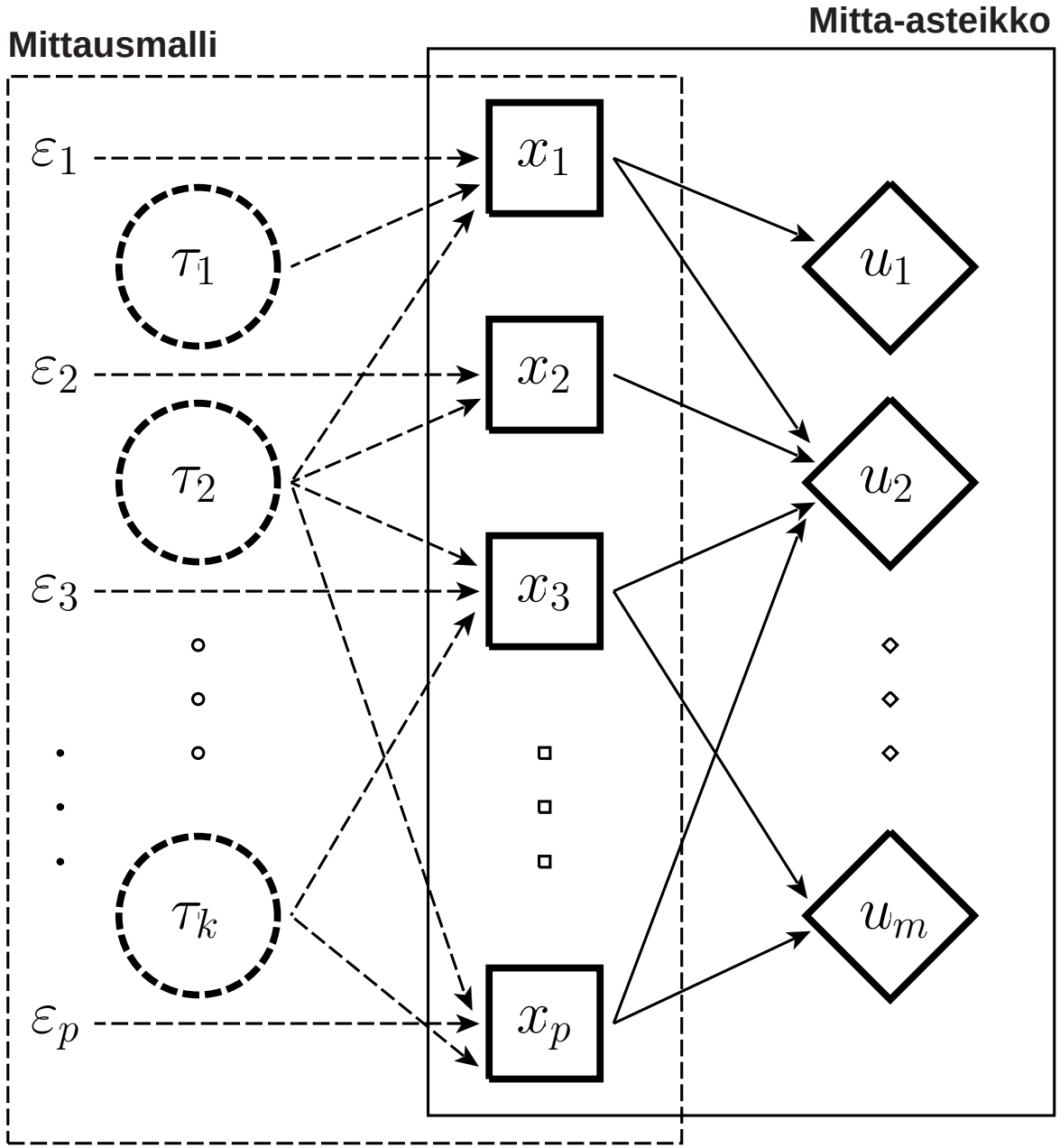
Kiinnostus politiikkaan, jolla on negatiivinen lataus (muista asteikon suunta!), ei nouse juurikaan esiin millään faktorilla. Samaa kuvastaa myös sen kommunaliteetti, joka on käytännössä nolla.

2. Survey-aineiston tiivistäminen

Tilastollisten menetelmien yleinen tavoite on **tiivistää** aineistoon sisältyvää informaatiota kuviksi, tunnusluvuiksi, taulukoiksi yms.

- tiivistäminen (ja muu analysointi) edellyttää ehdottomasti kunnollista aineistoon **tutustumista**
- tutustumisen ja tiivistämisen kannalta keskeistä: **jakaumien kuvaileminen** (graafisesti ja tunnusluvuin)
  - (empiirinen) **jakauma**: aineiston yhden muuttujan **kaikki** (mitatut ja koodatut) arvot
- tiivistäminen edellyttää kuitenkin yleensä enemmän:
  - kyselyaineistoissa on paljon (”liikaa”) muuttujia
  - muuttujien **yhdistely** tutkimusalan teorian perusteella
  - **faktorianalyysi**: faktoripisteet, summamuuttujat
  - muut monimuuttuja- ym. analyysimenetelmät

Mittausmallista aineiston tiivistämiseen (faktorianalyysi)



Mittausmalli ja mitta-asteikko

Mittausmalli (luento 1)

- Mitä **ilmiötä** tutkitaan? Montako **ulottuvuutta** siinä on?
- Millä ilmiötä **mitataan** — mahdollisimman hyvin?
- Yhteyksien ja ulottuvuuksien tutkiminen: **faktorianalyysi**
- Esimerkki: ESS (European Social Survey)

Mitta-asteikko (luento 2)

- Osioiden eli mitattujen muuttujien yhdistelmä
- Yhdistelytapoja: **faktoripisteet**, **summamuuttujat**
- Yleinen tavoite: **aineiston tiivistäminen**
- Esimerkki: ESS (European Social Survey)

Faktorianalyysin tulkinnasta aineiston tiivistämiseen

Faktorianalyysin toinen vaihe on aineiston **tiivistäminen** löydetyn faktorirakenteen perusteella. Siinä ”päästään eroon” suuresta määrästä muuttujia, joita on tarvittu ilmiön mittaamiseen, mutta joista on vaikea sellaisenaan saada kunnollista kokonaiskäsitystä.

Tiivistämisen onnistuminen edellyttää, että on löydetty oikea **faktorilukumäärä**, joka siis vastaa ilmiön ulottuvuuksien määrää. Tässä ennalta pohdittu mittausmalli on avainasemassa, sillä jos lukumäärä hahmotetaan vain aineiston pohjalta, jää turhan paljon epävarmuuksia. On tutkijan tehtävä päättää oikea lukumäärä.

Toinen edellytys on, että faktorit on pystytty **nimeämään** ja **tulkitsemaan** ymmärrettävästi. Myös tässä ilmiön tuntemus on ratkaisevan tärkeää. Ellei tulkinta onnistu, jää aineiston tiivistäminen keinotekoiseksi eikä siitä ole jatkossa vastaavaa hyötyä muiden analyysien tai visualisointien kannalta.

Esimerkki: faktorianalyysi (ESS), jatkoa aiempaan

ESS (European Social Survey), round 5, 2010, Suomi (n=1878)

Tarkastellaan tässä vain seuraavia muuttujia:

- How interested in politics (**1=very**, ..., 4=not at all)

Seuraavissa asteikko 0=no trust at all, ..., 10=complete trust:

- Trust in country's parliament
- Trust in the legal system
- Trust in the police
- Trust in politicians
- Trust in political parties
- Trust in the European Parliament
- Trust in the United Nations

Seuraavissa asteikko 0–10 eri sanamuodoin (10=positiivisin):

- Most people can be trusted or you can't be too careful
- Most people try to take advantage of you, or try to be fair
- Most of the time people are helpful or mostly looking out for themselves

(Kysymysten tarkemmat sanamuodot on tässä sivuutettu.)

Esimerkki faktorianalyysin tulosten esittämisestä

Factor structure of ESS (Finland), $n = 1878$	F1	F2	F3	h^2
F1: **Trust in Political Actors**				
Trust in politicians	**0.90**	0.25	0.14	0.88
Trust in political parties	**0.88**	0.24	0.14	0.84
Trust in country's parliament	**0.74**	0.21	*0.30*	0.68
Trust in the European Parliament	**0.70**	0.19	0.21	0.58
Trust in the United Nations	*0.45*	0.21	*0.37*	0.38
How interested in politics	-0.21	-0.01	-0.09	0.05
F2: **Trust in People**				
Most people can be trusted	0.16	**0.72**	0.18	0.58
Most people try to take advantage of you	0.10	**0.69**	0.15	0.50
Most of the time people are helpful	0.20	**0.58**	0.10	0.39
F3: **Trust in Authorities**				
Trust in the legal system	*0.38*	0.21	**0.75**	0.74
Trust in the police	0.18	0.21	**0.65**	0.50
Sum of squares	3.11	1.66	1.36	6.13
Variance explained %	28.3	15.0	12.4	55.7

h^2 = communalities

Esimerkki: faktorianalyysin suoritus (jatkoa)

Faktorianalyysin toisessa vaiheessa tiivistetään aineiston tiedot uusiksi faktoripistemuuttujiksi.

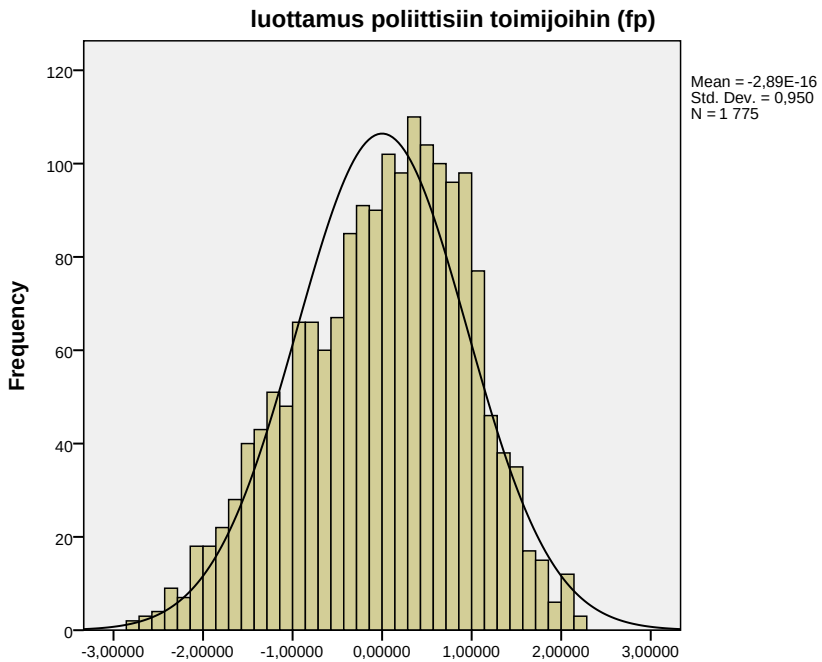
* faktorianalyysi 3-ulotteisen mittausmallin pohjalta (SPSS):

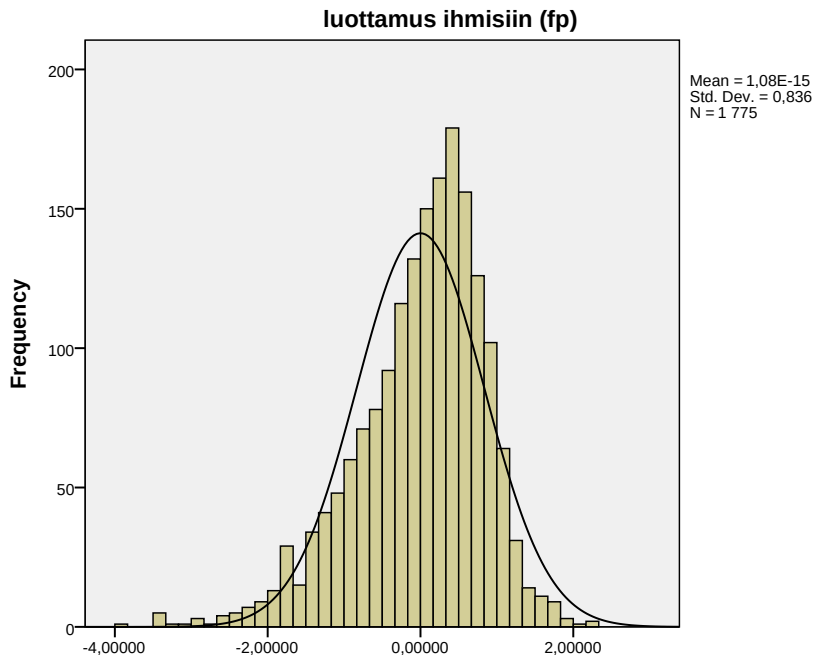
```
FACTOR
/VARIABLES polintr trstprl ...
... (samoin kuin aiemmin)
/SAVE REG(ALL).
* nimetään uudet muuttujat faktoreiden mukaisesti:
RENAME VARIABLES
(FAC1_1=Trust1)
(FAC2_1=Trust2)
(FAC3_1=Trust3).
VARIABLE LABELS
Trust1 'luottamus poliittisiin toimijoihin (fp)'
Trust2 'luottamus ihmisiin (fp)'
Trust3 'luottamus auktoriteetteihin (fp)'.
```


Esimerkki: faktoripisteet ja niiden jakaumat

Kun faktorit on tulkittu ja nimetty, on aika siirtyä takaisin aineiston havaintoyksikkötasolle. Muodostetut uudet faktoripistemuuttuja kertovat, mihin kohtaan mitäkin ulottuvuutta kuvaavaa jatkumoa kukin vastaaja sijoittuu.

Visualisoidaan kahden faktoripistemuuttujan jakaumia:

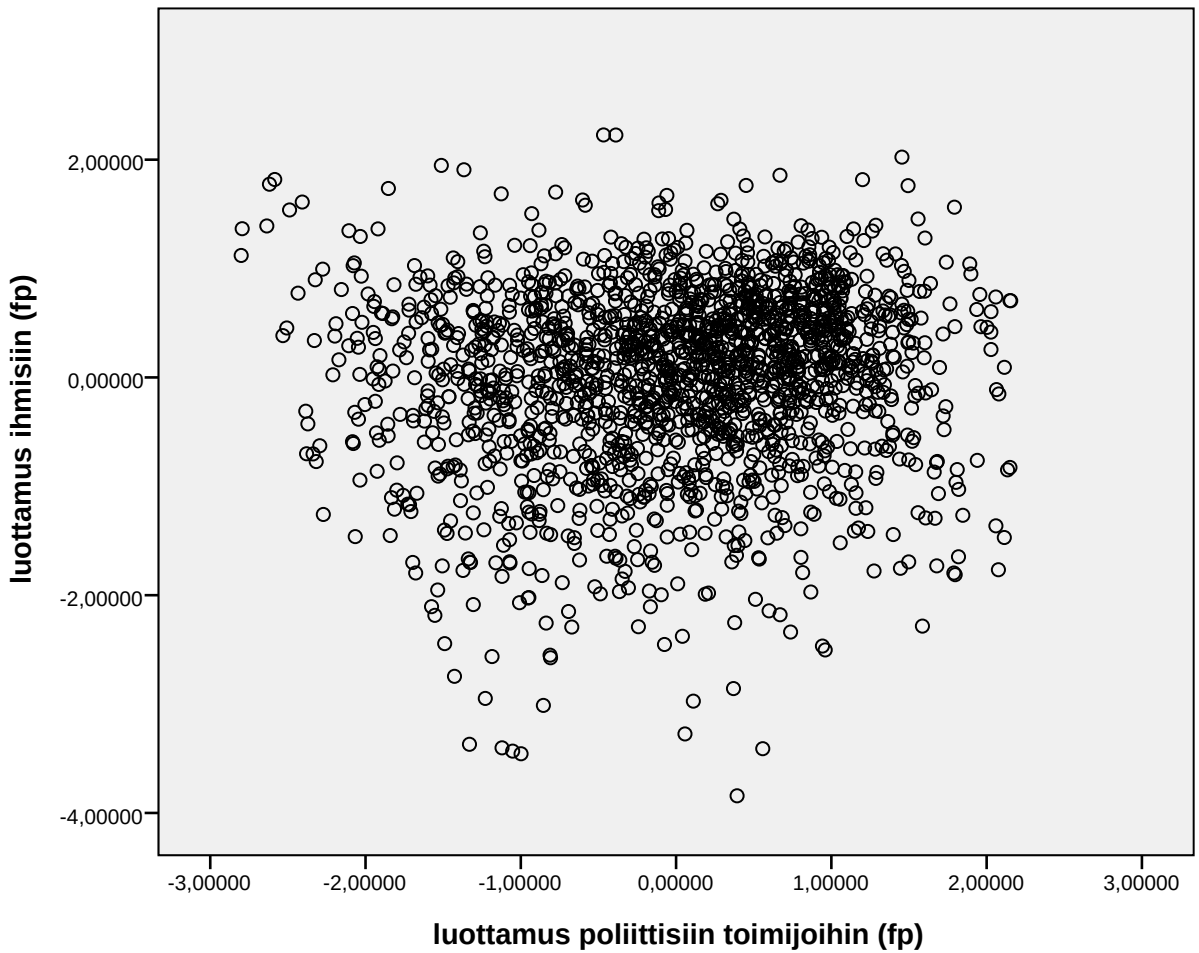




- Faktoripisteet ovat alkuperäisiä jatkuvampia (vinoutta on yhä).

Esimerkki: faktoripisteiden väliset yhteydet

Visualisoidaan näiden samojen muuttujien yhteisjakaumaa:

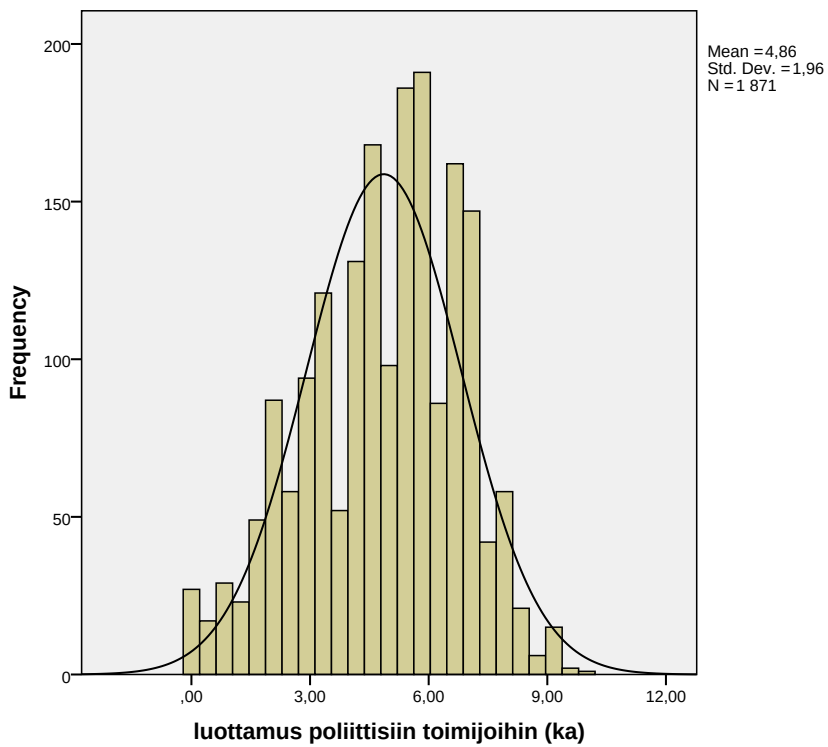


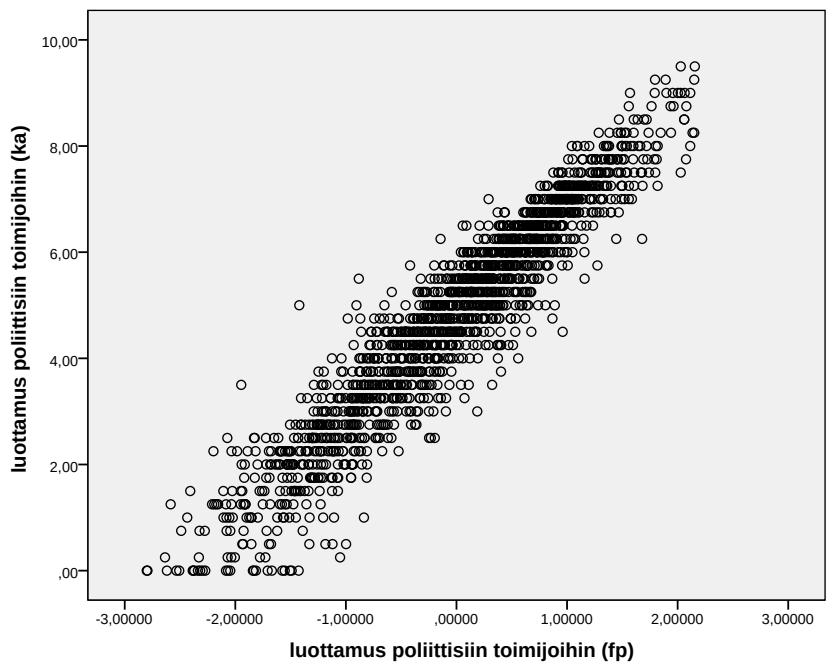
- Hahmotatko ”luottamus/epäluottamus”-nelikentän? Tulkitse!

Esimerkki: summamuuttujien muodostaminen

Vaihtoehtoinen tapa on tehdä **summamuuttujia**, mutta ne eivät hyödynnä faktorianalyysista saatua tietoa niin hyvin (esim. sitä, *miten hyvin* eri muuttujat mittaavat kutakin ulottuvuutta).

Histogrammi neljän ensimmäisen luottamusmuuttujan summamuuttujasta (keskiarvo) sekä hajontakuva siitä ja vastaavasta faktoripistemuuttujasta:





Esimerkki: faktoripisteiden luokittelu

Faktoripisteet voidaan edelleen tiivistää esim. ala- ja yläkvartiilien avulla (<25 %, 25–75 %, >75 %) kolmeen luokkaan: 1="epäluottamus", 2="siltä väliltä", 3="luottamus".

Kahden tällaisen muuttujan välinen ristiintaulukointi:

luottamus poliittisiin toimijoihin (123) * luottamus ihmisiin (123) Crosstabulation

Count

		luottamus ihmisiin (123)			
		1 epäluottamus	2 siltä väliltä	3 luottamus	Total
luottamus poliittisiin toimijoihin (123)	1 epäluottamus	139	205	100	444
	2 siltä väliltä	218	461	208	887
	3 luottamus	88	220	136	444
Total		445	886	444	1775

Tämän tyyppiset tiivistykset ovat tyypillisiä yhteiskuntatieteissä. Jatkuvat muuttujat toimivat edellä vain välivaiheena. Informaatiota hukkuu (paljon), mutta kokonaisuus voi olla helpompi hahmottaa. (Monesti saatetaan tiivistää vain kahteen luokkaan!)

Faktoripiste- ja summamuuttujat

Edellä on mainittu summamuuttujat faktoripisteiden vaihtoehtona. Kummankin käyttöön on perusteensa. Vertaillaan vähän:

Faktoripisteet muodostetaan faktorianalyysin perusteella, jolloin ne pyrkivät vastaamaan mahdollisimman hyvin aikaansaatuja faktoreita. Paremmat muuttujat huomioidaan suuremmilla painoilla kuin huonommat. Muuttujat voivat olla erilaisilla asteikoilla mitattuja sekä eri suuntaisia. Jos faktorit oletetaan keskenään korreloimattomiksi, faktoripisteetkään eivät korreloi keskenään.

Summamuuttujat muodostetaan valitsemalla vain parhaita muuttujia, mutta painottamalla niitä keskenään samanarvoisesti. Tutkimusalan teoria voi "sanella" valittavat muuttujat, jolloin ei faktorianalyysia periaatteessa tarvita lainkaan. On kuitenkin hyvä tarkistaa, miltä tilanne aineistossa näyttää. Muuttujien on joka tapauksessa oltava vertailukelpoisilla asteikoilla mitattuja sekä samansuuntaisia. Summamuuttujat korreloivat yleensä keskenään.

Aineiston tiivistäminen: johtopäätöksiä

Aineiston tiivistäminen on tärkeä ja välttämätön vaihe, kun työskennellään kyselytutkimusaineiston parissa. Mittaus ei onnistu kovin vähillä muuttujilla, mutta kokonaiskäsitysten saamiseksi muuttujia on aluksi ”liikaa”. Tiivistämisen pohjalta pitäisi saada hyvät mahdollisuudet jatkaa eteenpäin. On paljon helpompi jatkaa, kun yhtaikaa käsiteltävien muuttujien määrä on pienempi. *Monenlaiset kiinnostavat analyysit voivat alkaa vasta tästä!*

Faktoripisteet (tai summamuuttujat) voivat toimia esim. regressioanalyysin tai varianssianalyysin selitettävinä muuttujina (tai selittäjinä).

Jatkon kannalta olennaisen tärkeää on faktorianalyysin toteutus huolellisesti: hahmotetaan faktoreiden oikea (sopiva) lukumäärä sekä nimetään ja tulkitaan faktorit sisällöllisesti mielekkäästi.