Topics in Survey Methodology and Survey Analysis

Part 2

Kimmo Vehkalahti

University Lecturer, University of Helsinki
Department of Social Research, Statistics
http://www.helsinki.fi/people/Kimmo.Vehkalahti

fall 2013

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Outline of Part 2

Part 2 focuses heavily on **measurement**, which is one of the sources of **uncertainty** that is always present in survey research.

The following topics will be covered:

- Reliability, validity and measurement errors
- Exploratory and confirmatory analysis
- Data reduction/compression with factor analysis
- Visualization of multidimensional data

As the topics are more or less intertwined, there will not be a strict order of things. Activity of the participants is much appreciated and it will certainly affect the way (and order) we proceed.

The material includes some notes of rather basic statistics and graphs as well. They might be familiar to many participants, depending on everyone's previous studies. **The mathematical formulas represent additional information only.**

Bibliography and other optional material for Part 2

Measurement, reliability, validity

- Alwin, Duane F. (2007). *Margins of Error: A Study of Reliability in Survey Measurement.* Wiley.
- Fowler, Floyd J. (1995). *Improving Survey Questions: Design and Evaluation.* Applied Social Research Methods Series, Volume 38, Sage.
- Payne, Stanley L. (1951). *The Art of Asking Questions.* Princeton University Press.
- Saris, Willem E. & Gallhofer, Irmtraud N. (2007). *Design, Evaluation, and Analysis of Questionnaires for Survey Research.* Wiley.

Factor analysis and other multivariate (survey) methods

- Byrne, Barbara M. (2012). *Structural Equation Modeling with Mplus.* Routledge.
- Cudeck, Robert & MacCallum, Robert C., eds. (2007). *Factor Analysis at 100: Historical Developments and Future.* Lawrence Erlbaum Associates.
- Everitt, Brian (2009). *Multivariable Modeling and Multivariate Analysis for the Behavioral Sciences.* Chapman & Hall/CRC.
- Greenacre, Michael (2007). *Correspondence Analysis in Practice*, Second Edition, Chapman & Hall/CRC.
- Greenacre, Michael & Blasius, Jörg, eds. (2006). *Multiple Correspondence Analysis and Related Methods.* Chapman & Hall/CRC.
- Groves, Robert M.; Fowler Jr., Floyd J.; Couper, Mick P.; Lepkowski, James M.; Singer, Eleanor & Tourangeau, Roger (2004). *Survey Methodology.* Wiley.
- Mulaik, Stanley A. (2009). *Foundations of Factor Analysis,* Second Edition. Chapman & Hall/CRC.

Bibliography etc. for Part 2 (continued)

Visualization of statistical data

- Chen, Chun-houh; Härdle, Wolfgang & Unwin, Antony, eds. (2008). *Handbook of data Visualization.* Springer.
- Cleveland, William S. (1993). *Visualizing Data.* Hobart Press, Summit, NJ.
- Greenacre, Michael (2010). *Biplots in Practice.* Fundación BBVA. `http://www.multivariatestatistics.org`
- Robbins, Naomi B. (2005). *Creating More Effective Graphs.* Wiley.
- Tufte, Edward R. (2001). *The Visual Display of Quantitative Information*, Second Edition. Graphics Press. (Fifth printing, August 2007.)

Some books in Finnish

- Karjalainen, Leila & Karjalainen, Juha (2009). *Tilastojen graafinen esittäminen.* Pii-Kirjat.
- Ketokivi, Mikko (2009). *Tilastollinen päättely ja tieteellinen argumentointi.* Palmenia.
- Kuusela, Vesa (2000). *Tilastografiikan perusteet.* Edita.
- Mustonen, Seppo (1995). *Tilastolliset monimuuttujamenetelmät.* Survo Systems. `http://www.survo.fi/mustonen/monim.pdf`
- Nummenmaa, Tapio; Konttinen, Raimo; Kuusinen, Jorma & Leskinen, Esko (1996). *Tutkimusaineiston analyysi.* WSOY.
- Vehkalahti, Kimmo (2008). *Kyselytutkimuksen mittarit ja menetelmät.* Tammi.

Some studies used in examples

- European Social Survey, Round 5, subset of Finland `http://ess.nsd.uib.no/ess/round5/`
- Economic Freedom `http://www.heritage.org/Index/`
- Prices and Earnings around the Globe `http://www.ubs.com/research` `http://www.macrofocus.com/public/products/infoscope/datasets/pricesandearnings/`

Effects of measurement to (survey) data analysis

Measurement and measures in survey research:

- measurement model: **what** to measure
- measuring instrument: **how** to measure
- instrument in survey research: **questionnaire**
- **pattern:** collection of **items** (questions, statements)

Results are affected by the **measurement quality:**

1. **validity:** are we (really) measuring the right thing?
2. **reliability:** are we measuring accurately enough?

Measurement level sets the limits for the methods:

- **classification — ordering — numeric measurement**
  (*cf.* "nominal", "ordinal", "interval"/"ratio")
- most methods require numeric measurement
- in some methods classification is enough
- the level of ordering is often most problematic

Examples of items – what are their measurement levels?

Source: ESS (European Social Survey), `http://ess.nsd.uib.no/ess/`
Here: modified and abbreviated (DK = Don't Know).

- How interested are you in politics?
  Very interested (1), quite interested (2), hardly interested (3), not at all interested? (4), DK (8)
- Did you vote in the last national election?
  Yes (1), No (2), Not eligible to vote (3), DK (8)
- Have you boycotted certain products?
  Yes (1), No (2), DK (8)
- How satisfied are you with the present state of the economy?
  Extremely dissatisfied 00 01 02 03 04 05 06 07 08 09 10 Extremely satisfied, DK (88)
- The government should take measures to reduce differences in income levels.
  Agree strongly (1), Agree (2), Neither agree nor disagree (3), Disagree (4), Disagree strongly (5), DK (8)
- What is your current situation?
  paid work (1), education (2), unemployed (3), sick or disabled (4), retired (5), housework (6), other (7)
- How many hours do you work weekly: _____
- How much do you use internet: no access (00), never (01), less than once a month (02), once a month (03), several times a month (04), once a week (05), several times a week (06), every day (07), DK (88)
- Any particular religion you have considered yourself as belonging to?
  Roman Catholic (01), Protestant (02), Eastern Orthodox (03), Other Christian denomination (04), Jewish (05), Islamic (06), Eastern religions (07), Other non-Christian religions (08)

General aim: compressing the data

A general aim of **statistical methods** is to **compress** the information in the **data** into a form of **graphs** and **statistics**.

- compressing and other analyses will absolutely require a proper knowledge of the data (and the study in question)
- central for the knowledge: graphs (and statistics) of the distributions
  - empirical distribution: all the measured (and coded) values of one variable in the data
- other ways of compressing the data:
  - combining the variables, e.g., by forming summated variables (meaningful only if summing is reasonable)
  - **multivariate methods**, such as **factor analysis**

Types of variables in the data

- **quantitative variables:**
  - **continuous** variables (such as age, length, weight etc.) (only a few identical values)
  - **discrete** variables (such as scales of opinions, counts) (many identical values, i.e., only a few different ones)

In practice, measuring something and saving it on the computer is possible only on a finite precision ("everything is discrete").

A variable may, however, be *interpreted* as continuous, if it reflects a continuous phenomenon or issue (e.g., age).

- **qualitative variables** (all discrete)**:**
  - **ordered** variables (e.g., education)
  - **classified** variables (e.g., gender)

Quantitative variables may always be transformed to qualitative ones (by classifying) but **not** the other way. Hence, it is worthwhile to **measure** as precisely and accurately as possible! It cannot be redone...

Statistics and their interpretation

When data is compressed into **statistics**:

- some of the information is always lost
- a plain statistic (typically one number!) does not say much

The most general statistics is the **mean** (average):

- the sum of the values divided by the number of observations
- meaningful only if summing is reasonable
- applicable only for quantitative variables
- plain mean is not enough (tells nothing about **variation**)

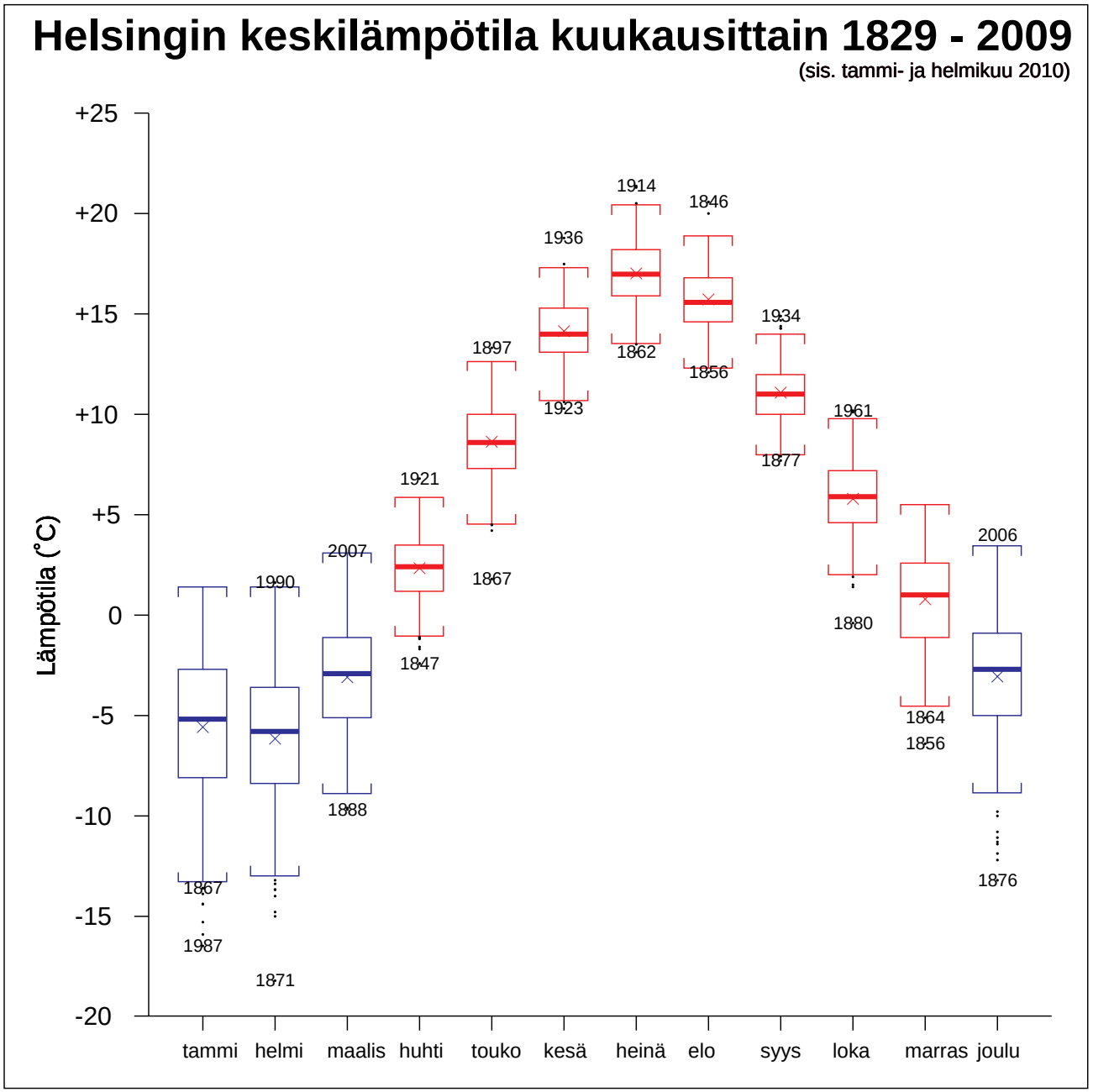
Often a better alternative is offered by the **median**:

- the central value of the ordered variable
- no calculations are employed: also valid for ordering level
- more robust with possible outliers (unlike mean)
- plain median is not enough (tells nothing about **variation**)

A useful collection of **five** (order) **statistics:**

- min, lower quartile (25 %), median (50 %), upper quartile (75 %), max
- graphical representation: box (and whiskers) plot
- note: the "box" will then include half of the observations

Example: average temperatures in Helsinki, 1829–2009



Variation and dependence

One of the key concepts of statistics is **variation:**

- the more variation, the more information
- variable with no variation is *constant* (same value for all)
  - no statistical information (of course, may be interesting)
- measures of variation (*cf.* box plot and its statistics):
  - **range:** [min(imum), max(imum)]
  - **quartile range:** [lower q, upper q]
    (also the lengths of the ranges may be used)
- most typical measure of variation is the **standard deviation:**
  - "the average deviation" of the observations from their mean
  - given in same units with the mean (easy to interpret?)
  - mean $\pm$ 1 std devs covers ca. 68 % of the values
  - mean $\pm$ 2 std devs covers ca. 95 % of the values
    (assuming the distribution is quite *symmetric* and *unimodal*)
  - the square of the std dev is *variance* (more theoretical)

Dependence and correlation

Another key concept is **dependence:** most research questions are somehow related to *dependence of different aspects or issues*.

- the character of the dependence can be evaluated by examining the scatter plot (or a cross tabulation)
- important case: **linear dependence**
- *but:* dependence may often be **nonlinear**

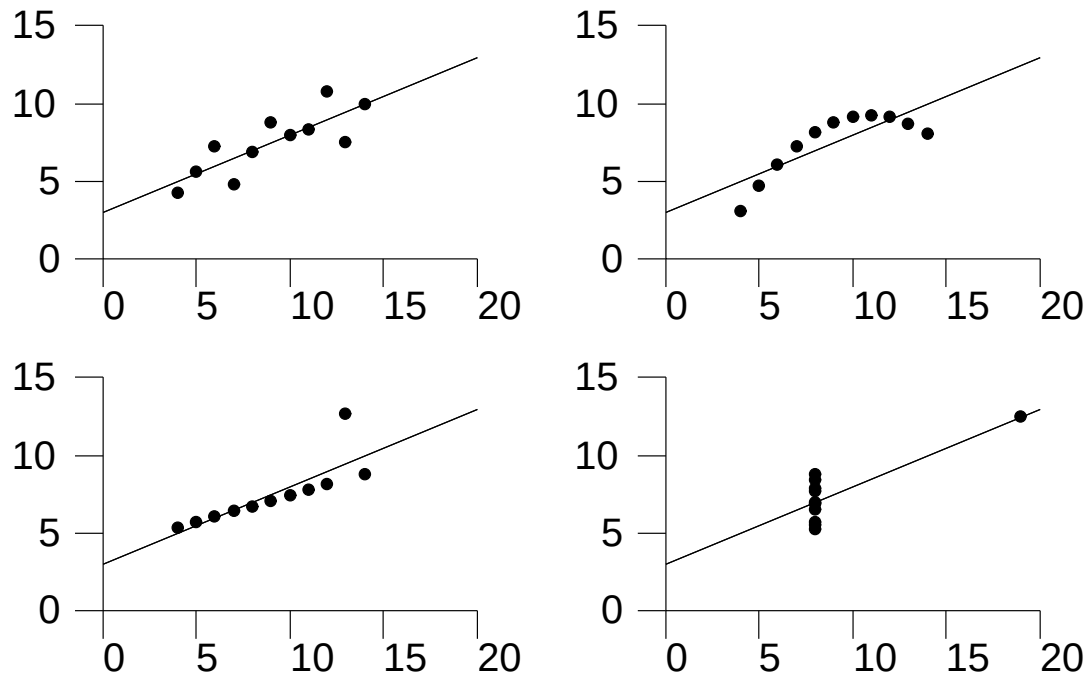
In case of linear dependence the **correlation** might be useful:

- correlation (coefficient) is a statistic of two variables (measure of variation of which is the standard deviation)
- correlation is a number on the interval $[-1, 1]$, e.g., -0.72
- most important: learning to **interpret** the correlation:
  - is the correlation positive of negative?
  - when is the correlation practically zero?
  - what if the correlation is almost $+1$ or -1?
  - *is one statistic again enough for the purpose?*
- correlation describes merely a relation, **not** a causation (causal inference is a subject matter, not a statistical matter!)

Visualization of variation and dependence: scatter diagram

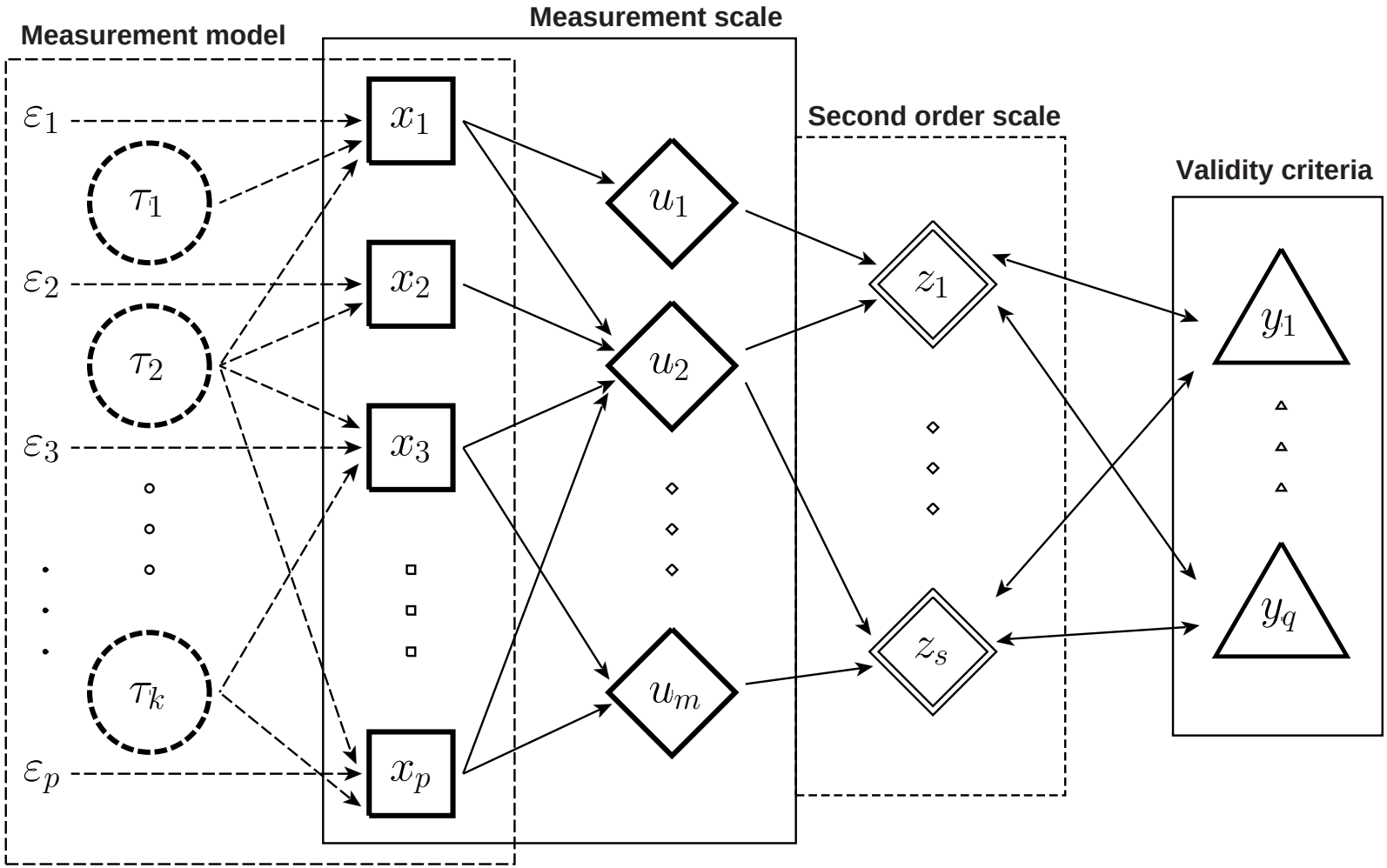
A **scatter diagram** is an excellent way to visualize and analyse **variation** and **dependence** simultaneously:

- basically a graph of two continuous variables
- numerous variations for different settings
- capable of presenting a large amount of information
- also reveals surprises such as outliers



Anscombe, F. (1973). Graphs in statistical analysis, *American Statistician*, **27**, 17-21.

Measurement framework: "guide from plans to analyses"



- a basis for assessing the measurement quality
- **essential statistical method: factor analysis**
- includes other multivariate methods as well

Four parts of the measurement framework

Measurement model

- What is the *phenomenon* under study?
- How many *dimensions* it might consist of?
- How could those dimensions be *measured*?
- **factor analysis** (exploratory—confirmatory)

Measurement scale

- combination of measures items
- examples: factor scores, summated scales, indices
- **compressing the data**

Second order scale

- result of e.g., regression or discriminant analyses
- connects measurements with other multivariate methods

Validity criteria

- a criteria defined outside of the measurement model
- for comparisons, orderings, classifications etc. of respondents

Theory behind the framework: Measurement model

In this material, the mathematical formulas represent additional information only.

Tarkkonen, L. & Vehkalahti, K. (2005). Measurement errors in multivariate measurement scales, *Journal of Multivariate Analysis*, **96**, 172–189.

Let $\boldsymbol{x} = (x_1, \ldots, x_p)'$ measure k (**important here:** $k < p$)
unobservable **true scores** $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_k)'$ with
unobservable **measurement errors** $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)'$.

Assume $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\mathrm{cov}(\boldsymbol{\tau}, \boldsymbol{\varepsilon}) = \mathbf{0}$. The measurement model is

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{\tau} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{B} \in \mathbb{R}^{p \times k}$ specifies the relationship between $\boldsymbol{x}$ and $\boldsymbol{\tau}$.

Denoting $\mathrm{cov}(\boldsymbol{\tau}) = \boldsymbol{\Phi}$ and $\mathrm{cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi}$ we have

$$\mathrm{cov}(\boldsymbol{x}) = \boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{\Phi}\boldsymbol{B}' + \boldsymbol{\Psi}, \tag{2}$$

- where it is assumed that $\boldsymbol{\Sigma} > \mathbf{0}$ and $\boldsymbol{B}$ has full column rank.

Theory behind the framework: Estimation of parameters

In this material, the mathematical formulas represent additional information only.

The **parameters** are the $pk + k(k+1)/2 + p(p+1)/2$ (unique) elements of the matrices $\boldsymbol{B}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Psi}$. In general, there are too many, since $\boldsymbol{\Sigma}$ has only $p(p+1)/2$ elements.

- Identifiability is obtained by imposing assumptions on the true scores and the measurement errors.
- **Typical:** assume that $\text{cov}(\boldsymbol{\tau}) = \boldsymbol{I}_k$, an identity matrix of order k, and $\text{cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi}_d = \text{diag}(\psi_1^2, \ldots, \psi_p^2)$.
- With these the model conforms with the orthogonal factor analysis model where the *common factors are directly associated with the true scores* and the *specific factors are interpreted as measurement errors*.

Assuming **multinormality** the parameters can be estimated using e.g., **the maximum likelihood** method of factor analysis.

All this is closely related to Structural Equation Models (SEM).

Theory behind the framework: Structural validity

In this material, the mathematical formulas represent additional information only.

Structural validity is a property of the measurement model.

- Important, as the model forms the core of the framework and hence affects the quality of all scales created.
- Lack of structural validity can be revealed by testing
  - hypotheses on the dimension of $\boldsymbol{\tau}$
  - hypotheses on the effects of $\boldsymbol{\tau}$ on $\boldsymbol{x}$ (matrix $\boldsymbol{B}$)
- The whole approach could be called *semi-confirmatory*.
- Residuals of the model obtained by estimation of var($\boldsymbol{\varepsilon}$).
- Dimension of $\boldsymbol{\tau}$ will make the reliabilities identified.
- Appropriate (e.g. **graphical**) factor rotation is essential.

Similarly with other questions of validity, knowledge of the theory and practice of the application is crucial.

Also these topics are related to Structural Equation Models (SEM).

Theory behind the framework: Measurement scale

In this material, the mathematical formulas represent additional information only.

In further analyses, the variables $\boldsymbol{x}$ are best used by creating **multivariate measurement scales** $\boldsymbol{u} = \boldsymbol{A}'\boldsymbol{x}$, where $\boldsymbol{A} \in \mathbb{R}^{p \times m}$ is a matrix of the weights. Using (2) we obtain

$$\text{cov}(\boldsymbol{u}) = \boldsymbol{A}'\boldsymbol{\Sigma}\boldsymbol{A} = \boldsymbol{A}'\boldsymbol{B}\boldsymbol{\Phi}\boldsymbol{B}'\boldsymbol{A} + \boldsymbol{A}'\boldsymbol{\Psi}\boldsymbol{A}, \tag{3}$$

the (co)variances generated by the **true scores** and the (co)variances generated by the **measurement errors**.

Examples of measurement scales include factor scores, psychological test scales, or any other linear combinations of the observed variables. The weights of the scale may also be predetermined values according to a theory.

Theory behind the framework: Predictive validity

In this material, the mathematical formulas represent additional information only.

Predictive validity is a property of the measurement scale.

- Assessed by the correlation(s) between the (second order) scale and an *external criterion*.
- In general, a second order scale is denoted by $\boldsymbol{z} = \boldsymbol{W}'\boldsymbol{u} = \boldsymbol{W}'\boldsymbol{A}'\boldsymbol{x}$, where $\boldsymbol{W} \in \mathbb{R}^{m \times s}$ is a weight matrix and a criterion is denoted by $\boldsymbol{y} = (y_1, \ldots, y_q)'$.
- Often, these scales are produced by regression analysis, discriminant analysis, or other multivariate statistical methods.

In the most general case, the predictive validity would be assessed by the **canonical correlations** between $\boldsymbol{z}$ and $\boldsymbol{y}$.

Theory behind the framework: Predictive validity

In this material, the mathematical formulas represent additional information only.

Example: consider the regression model $y = \beta_0 + \boldsymbol{\beta}'\boldsymbol{u} + \delta$, where y is the response variable, β_0 is the intercept, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)'$ is the vector of the regression coefficients, $\boldsymbol{u}$ is the vector of the predictors (e.g., factor scores), and δ is a model error.

Now, the criterion y is a scalar, and the second order scale is given by the prediction scale $z = \hat{\boldsymbol{\beta}}'\boldsymbol{u}$, where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_m)'$.
Hence the predictive validity is equal to ρ_{zy}, the multiple correlation of the regression model.

Monte Carlo simulations indicate that the factor scores offer the most stable method for predictor selection in the regression model.
Details are found in

Vehkalahti, K., Puntanen, S. & Tarkkonen, L. (2007). Effects of measurement errors in predictor selection of linear regression model, *Computational Statistics & Data Analysis*, **52**(2), 1183–1195.

Theory behind the framework: Reliability (Tarkkonen's rho)

In this material, the mathematical formulas represent additional information only.

According to the definition of reliability, Tarkkonen's rho is obtained as a **ratio of the variances**, i.e., the diagonal elements of the matrices in (3). Hence we have

$$\boldsymbol{\rho_u} = \text{diag}\left(\frac{\boldsymbol{a}_1'\boldsymbol{B\Phi B}'\boldsymbol{a}_1}{\boldsymbol{a}_1'\boldsymbol{\Sigma a}_1}, \ldots, \frac{\boldsymbol{a}_m'\boldsymbol{B\Phi B}'\boldsymbol{a}_m}{\boldsymbol{a}_m'\boldsymbol{\Sigma a}_m}\right)$$

$$= (\boldsymbol{A}'\boldsymbol{B\Phi B}'\boldsymbol{A})_d \times [(\boldsymbol{A}'\boldsymbol{\Sigma A})_d]^{-1}$$

or, in a form where the matrix $\boldsymbol{\Psi}$ is explicitly present:

$$\boldsymbol{\rho_u} = \text{diag}\left(\left[1 + \frac{\boldsymbol{a}_1'\boldsymbol{\Psi a}_1}{\boldsymbol{a}_1'\boldsymbol{B\Phi B}'\boldsymbol{a}_1}\right]^{-1}, \ldots, \left[1 + \frac{\boldsymbol{a}_m'\boldsymbol{\Psi a}_m}{\boldsymbol{a}_m'\boldsymbol{B\Phi B}'\boldsymbol{a}_m}\right]^{-1}\right)$$

$$= \{\boldsymbol{I}_m + (\boldsymbol{A}'\boldsymbol{\Psi A})_d \times [(\boldsymbol{A}'\boldsymbol{B\Phi B}'\boldsymbol{A})_d]^{-1}\}^{-1}$$

Theory behind the framework: Reliability (special cases)

In this material, the mathematical formulas represent additional information only.

Many models, scales, and reliability coefficients established in the test theory of psychometrics are special cases of the framework.

Example: $\boldsymbol{x} = \boldsymbol{\mu} + \mathbf{1}\tau + \boldsymbol{\varepsilon}$ and $u = \mathbf{1}'\boldsymbol{x}$ (unweighted sum).

Now, $\boldsymbol{\Sigma} = \sigma_\tau^2 \mathbf{1}\mathbf{1}' + \boldsymbol{\Psi}_d$ and $\sigma_u^2 = \mathbf{1}'\boldsymbol{\Sigma}\mathbf{1} = p^2\sigma_\tau^2 + \mathrm{tr}(\boldsymbol{\Psi}_d)$.

$$\rho_{uu} = \frac{p^2\sigma_\tau^2}{\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}} = \frac{p}{p-1}\left(\frac{p^2\sigma_\tau^2 - p\sigma_\tau^2}{\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}}\right)$$

$$= \frac{p}{p-1}\left(\frac{\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1} - \mathrm{tr}(\boldsymbol{\Psi}_d) - \mathrm{tr}(\boldsymbol{\Sigma}) + \mathrm{tr}(\boldsymbol{\Psi}_d)}{\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}}\right)$$

$$= \frac{p}{p-1}\left(1 - \frac{\mathrm{tr}(\boldsymbol{\Sigma})}{\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}}\right) = \frac{p}{p-1}\left(1 - \frac{\sum_{i=1}^p \sigma_{x_i}^2}{\sigma_u^2}\right),$$

which is the original form of Cronbach's alpha, given in Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests, *Psychometrika*, **16**, 297–334.

Factor analysis and its interpretations

Factor analysis (FA) is a multivariate method, where the variables are assumed to measure a multidimensional *latent* construct. It means that the phenomena under study (such as attitudes, values etc.) can not be measured directly. Instead, it must be measured indirectly, by using several questions or items.

There are two steps or phases in FA. First, the aim is to find the **latent structure** using the observed data. The structure can be described a priori by a measurement model. Two different approaches of FA are *explorative* (data-based) and *confirmatory* (model-based) analysis *(cf. SEM)*. In practice, most analyses are somewhere in between: when we know more about the phenomena, the measurement model has a bigger role, whereas exploring new phenomena we have less advance information and must rely on our data and measurements (more uncertainty!).

Factor analysis and its interpretations

The second step of FA is the **compression of the data** based on the factor structure found in the first phase. Then we will "get rid of" a large number of variables, that are difficult to analyse as such, but they are definitely needed to measure the phenomenon.

A successful compression requires that we have the right **number of factors** (the number of dimensions of the phenomenon). The measurement model plays a key role, as there are too many uncertainties, if the number of factors is found solely based on the data. It is the task of the researcher to decide the right number.

Another requisite is that the factors can be **named** and **interpreted** understandably. Knowing the phenomenon is again crucial. If the interpretation does not work, the compression of the data remains artificial and it does not help the subsequent analyses or visualizations.

- Let us consider FA and its interpretation with a small example.

Example: factor analysis (ESS)

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

We will only use the following variables here:

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

These have a scale 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

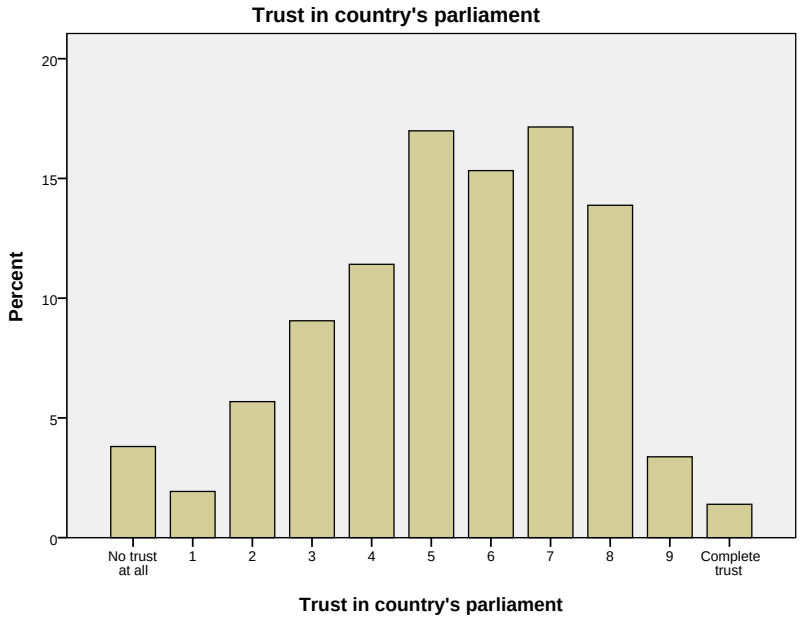
These have a scale 0–10 with various wordings (10=most positive):

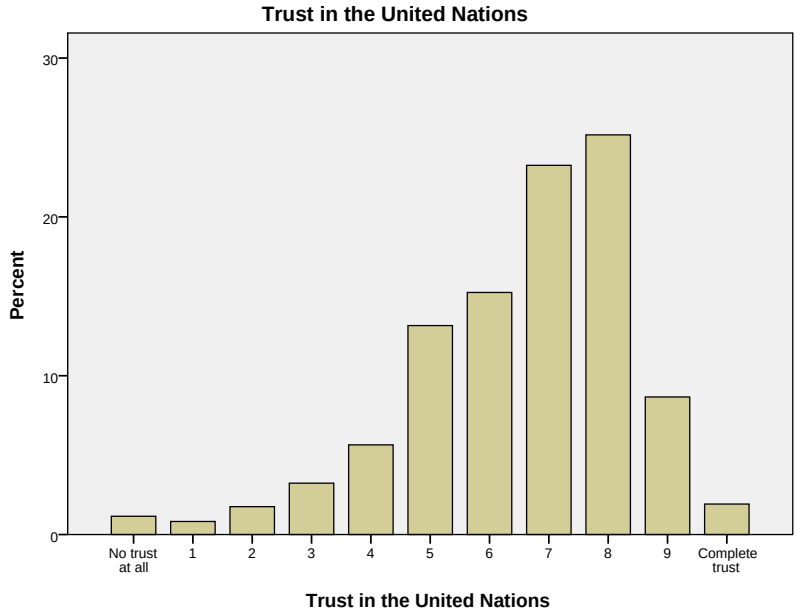
- Most people can be trusted or you can't be too careful
- Most people try to take advantage of you, or try to be fair
- Most of the time people are helpful or mostly looking out for themselves

(Details can be found from the codebook of the ESS.)

Example: factor analysis (ESS); assumptions

The distributions vary, but the measurement level is essential:





We assume that we have a measurement model of **2–3 factors** (dimensions including trust to political actors and human beings).

We conduct a factor analysis using SPSS *(during the lecture)*.

(The example is a bit artificial, the focus being on the method and its interpretations. These data allow much richer settings!)

Example: factor analysis (SPSS syntax)

Factor analysis (FA) will compress the information on the linear relations of the variables into a given number of factors.

* FA, based on a 3-dimensional measurement model (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.
```


Example: FA (ESS); phase 1: factoring & rotation

The most essential output of FA is the rotated factor matrix:

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.

(hierarchical ordering based on the factor loadings)

Example: factor analysis (ESS); output

It is also good to study the communalities of the variables:

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.

Unfortunately, SPSS gives them in another table, in another order!
(In addition, the "Initial" values above are completely useless.)

Example: factor analysis (ESS); interpretation

(Tables and interpretations are studied in detail in the class.)

From other tables we can see that three factors "explain" 55.7 % of the variance. However, we are more interested in the factor loadings, the dependencies between the variables and the factors *(cf. the arrows of the measurement model)*.

Factor 1 ("trust to political actors") is the most powerful one. Factor 2 could be "trust to human beings" and Factor 3 "trust to authorities". The more we have support from the subtantial theory, the easier it is to name the factors. (The theory may also support *correlating factors*; here we have assumed that they are uncorrelated, which makes things a bit simpler.)

"How interested in politics" has a negative loading (note the direction of its scale!), but it is not very large on any of the factors. The same is reflected by its communality, which is about zero.

Example: Presenting and interpreting the results of FA

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

Example: factor analysis (SPSS syntax 2)

In phase 2 of FA, the active part of the data is compressed into new factor score variables, reflecting the 3 hypothetical factors.

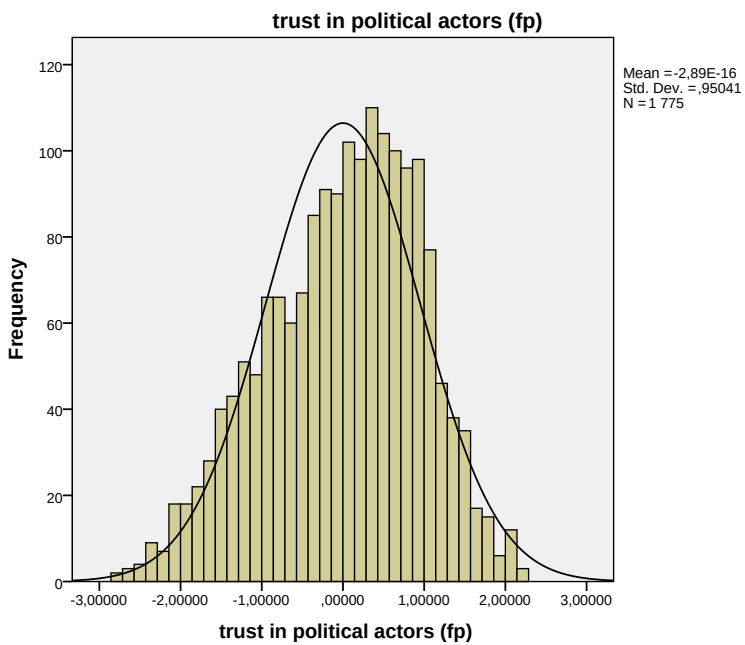
```
* FA, based on a 3-dimensional measurement model (SPSS):

FACTOR
/VARIABLES polintr trstprl ...
... (similarly as before)
/SAVE REG(ALL).
* rename the new variables according to the interpreted factors:
RENAME VARIABLES
(FAC1_1=Trust1)
(FAC2_1=Trust2)
(FAC3_1=Trust3).
VARIABLE LABELS
Trust1 'trust in political actors (fp)'
Trust2 'trust in people (fp)'
Trust3 'trust in authorities (fp)'.
```

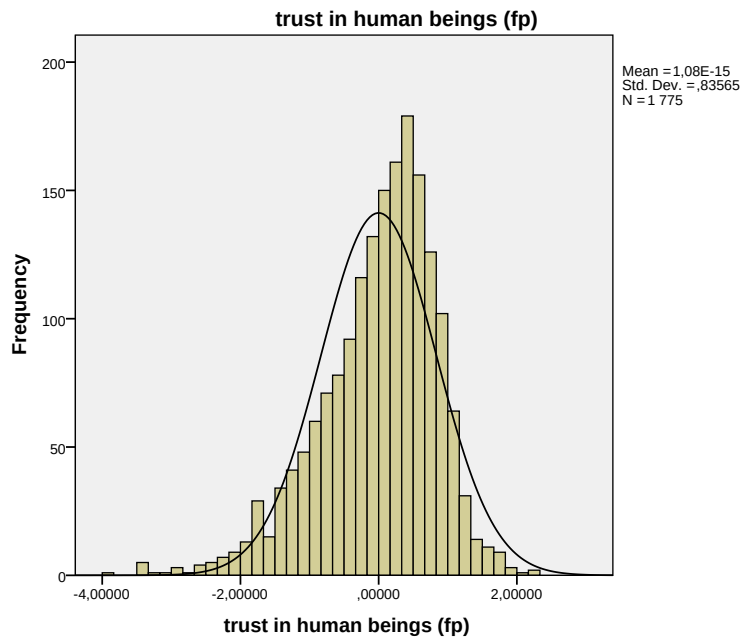

Example: FA (ESS); phase 2: compression

After the factors have been named and interpreted, it is time to move back to the individual level of the data, by computing new *factor score* variables. They will indicate the points on the dimensions, where each respondent would be located.

An alternative method is to compute **sums** (or means) **of the items**, but they do not take the full use of the FA, e.g., *how well* the items do measure each factor or dimension.

Let us visualize the distributions of two factor scores:

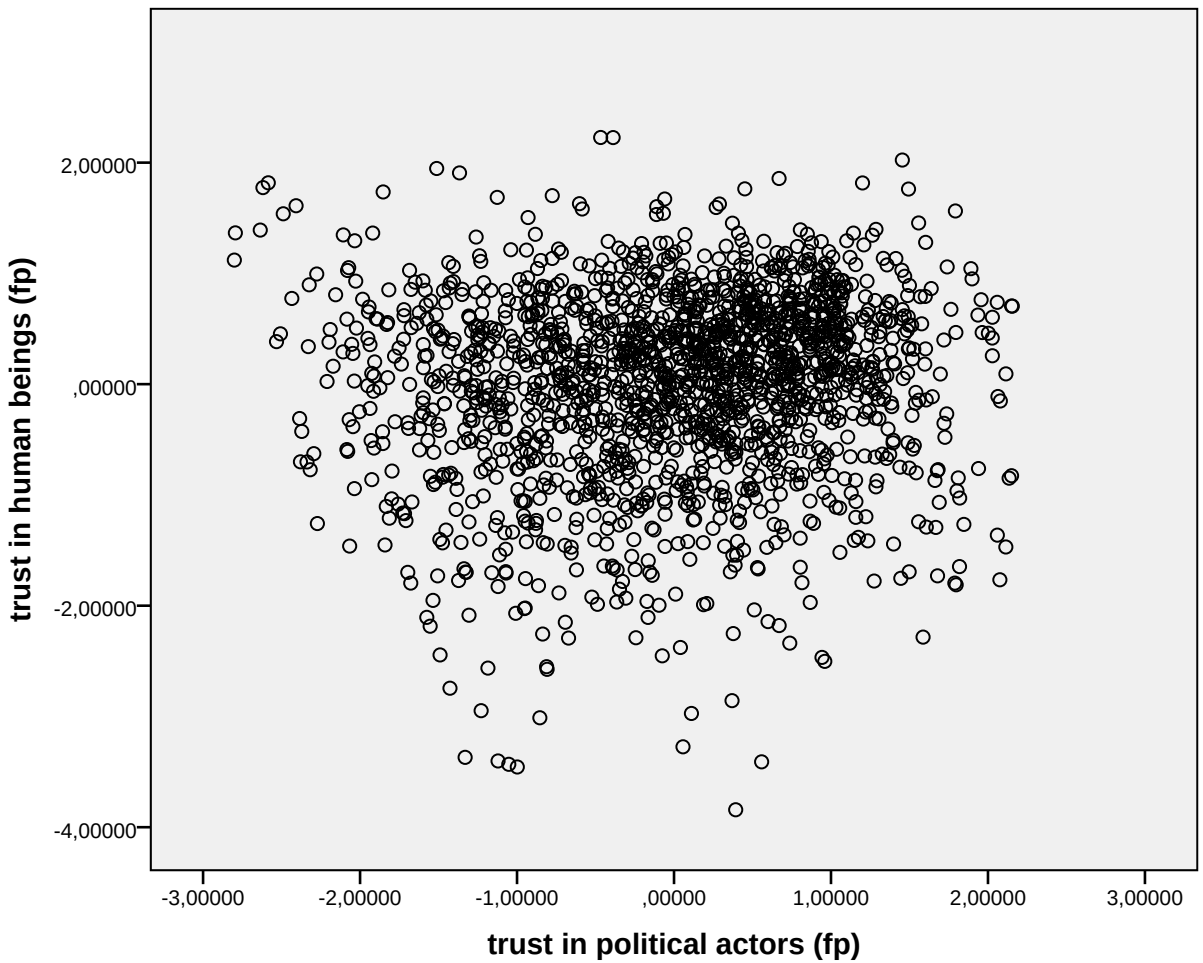




Example: factor analysis (ESS); factor scores

The factor scores are more continuous than the original items (of course, some skewness is still present).

Let us visualize these scores simultaneously by a scatter plot:



Do you see the combinations of the "trust/mistrust"? Interpret!

Example: factor analysis (ESS); factor scores

We may further compress the factor scores, e.g., by using the lower and upper quartiles (<25 %, 25–75 %, >75 %) to three categories: 1="mistrust", 2="between", 3="trust".

Here we have a cross table of two such variables:

trust in political actors (123) * trust in human beings (123) Crosstabulation

Count

		trust in human beings (123)			
		1 mistrust (H)	2 between (H)	3 trust (H)	Total
trust in political actors (123)	1 mistrust (P)	139	205	100	444
	2 between (P)	218	461	208	887
	3 trust (P)	88	220	136	444
Total		445	886	444	1775

These types of compressions are typical in social sciences. Often, the continuous variables serve only as an intermediate form. Much information is lost, but it may be much easier to comprehend the whole. Quite often we may end up in two categories only, i.e., dichotomous variables. Compression, indeed!

Factor scores vs sums of variables

The factor scores and the straight sums (means) of items can both be justified. Let us compare them a bit:

Factor scores are computed based on the factor analysis. They will then correspond to the factors as well as possible. The best items will get more weight and vice versa. The items may be measured with different scales and different directions. If the factors are assumed uncorrelated, also the factor scores will be uncorrelated.

Sums (means) of items are computed using the best items (found by FA), but weighting them equally. The substantial theory may dictate the items (which means that FA would not be needed at all). It is still wise at least to check the situation with the current data (and not to compute the sums blindly). In any case, the items must be measured with similar scales and similar directions. Usually the sums will correlate with each other.

Compressing the data: conclusions

Compressing the data is important and necessary when working with survey data. Measurement is quite impossible with a small number of variables, but in order to comprehend the whole there are "too many" variables. A successful compression should provide with a good ground to continue the analyses. It is easier to proceed when the number of variables to be simultaneously worked with is first reduced.

Several interesting analyses will just begin from here!

For subsequent analyses, it is most essential to succeed with FA: to find the right number of factors as well as to give them good names and interpretations.

Methods for classifying and clustering the respondents

After compressing the data we can dig deeper in it, and ask:

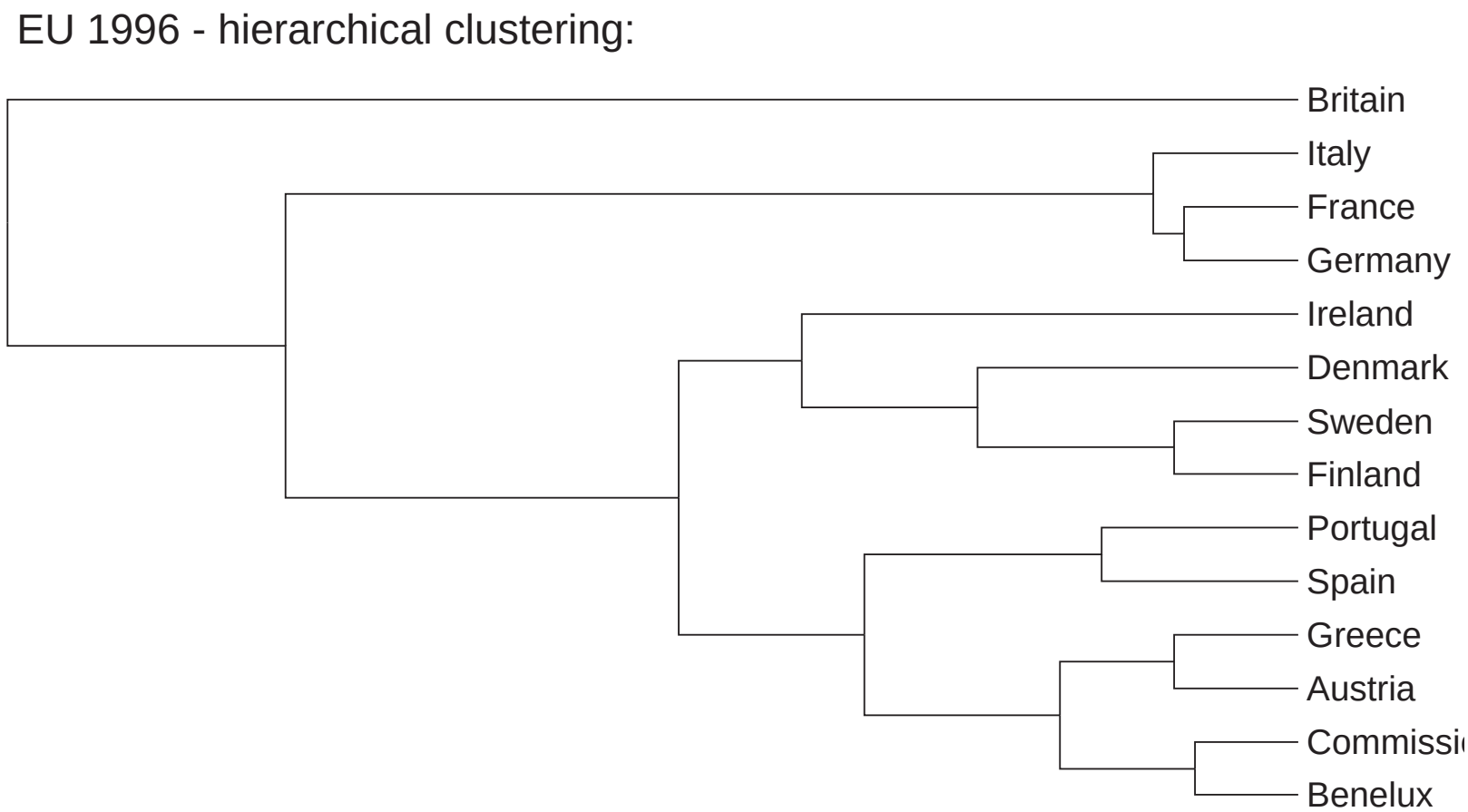
- "what type of (respondent) groups could be found in data?" and "how should we interpret and label them?"
  - **clustering methods**
- "what makes the difference between the (known) groups?" and "into which group we would classify a new respondent?"
  - **discriminant analysis**
- "how do the respondents/groups settle and relate to each other in respect of the background variables and other classifications?"
  - **correspondence analysis**

The results of multivariate methods are best explained by visualizing them — e.g. using variations of the scatter plot.

Tools for immediate visualization of multidimensional data may be used for small data, aggregated subsets, clusters etc.

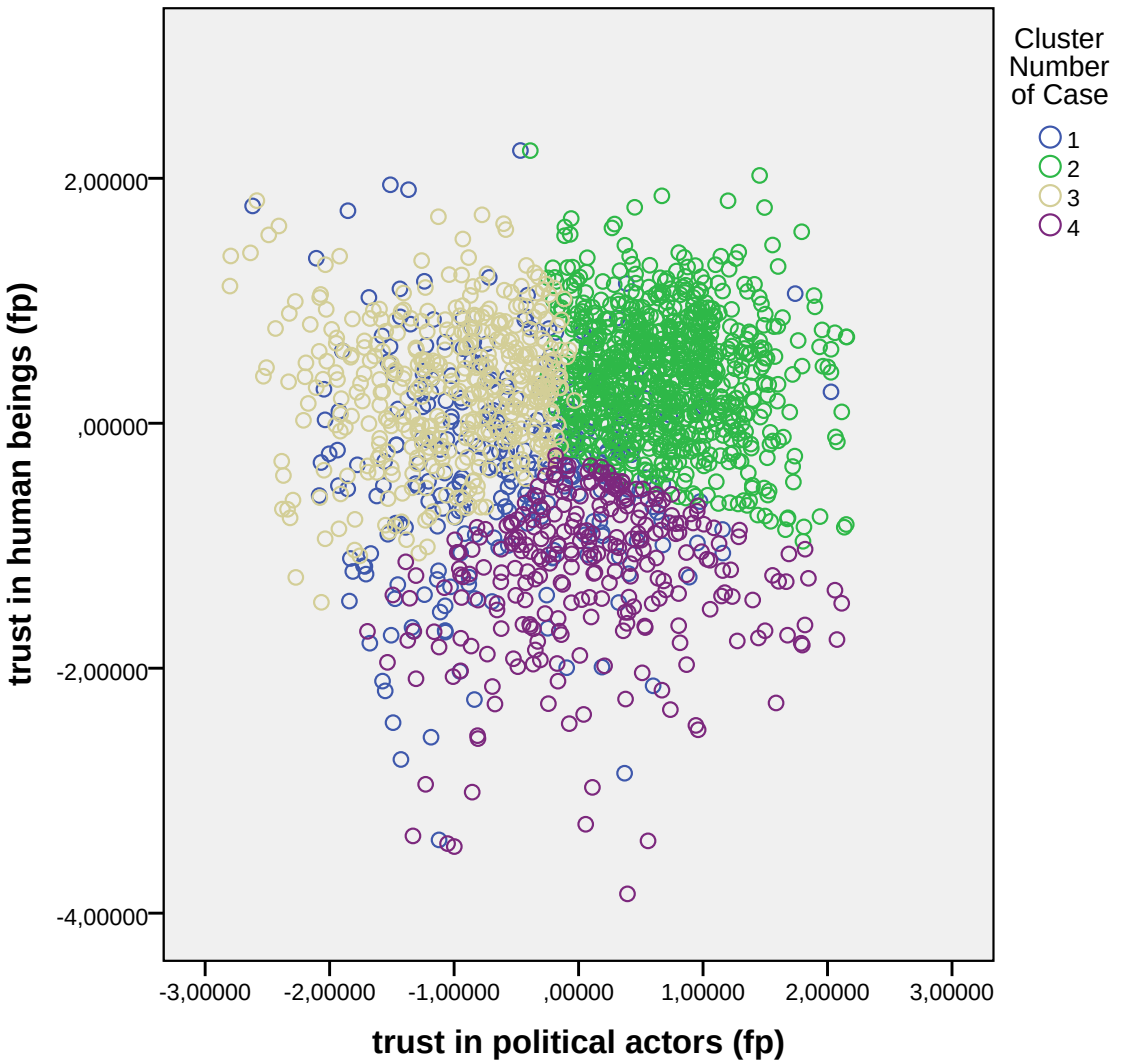
Example: hierarchical clustering

The aim is to find similar profiles of groups/individuals and form new groups by some suitable criteria. Better suited for smaller (subsets of) data, where the observations have names:



Example: clustering (ESS)

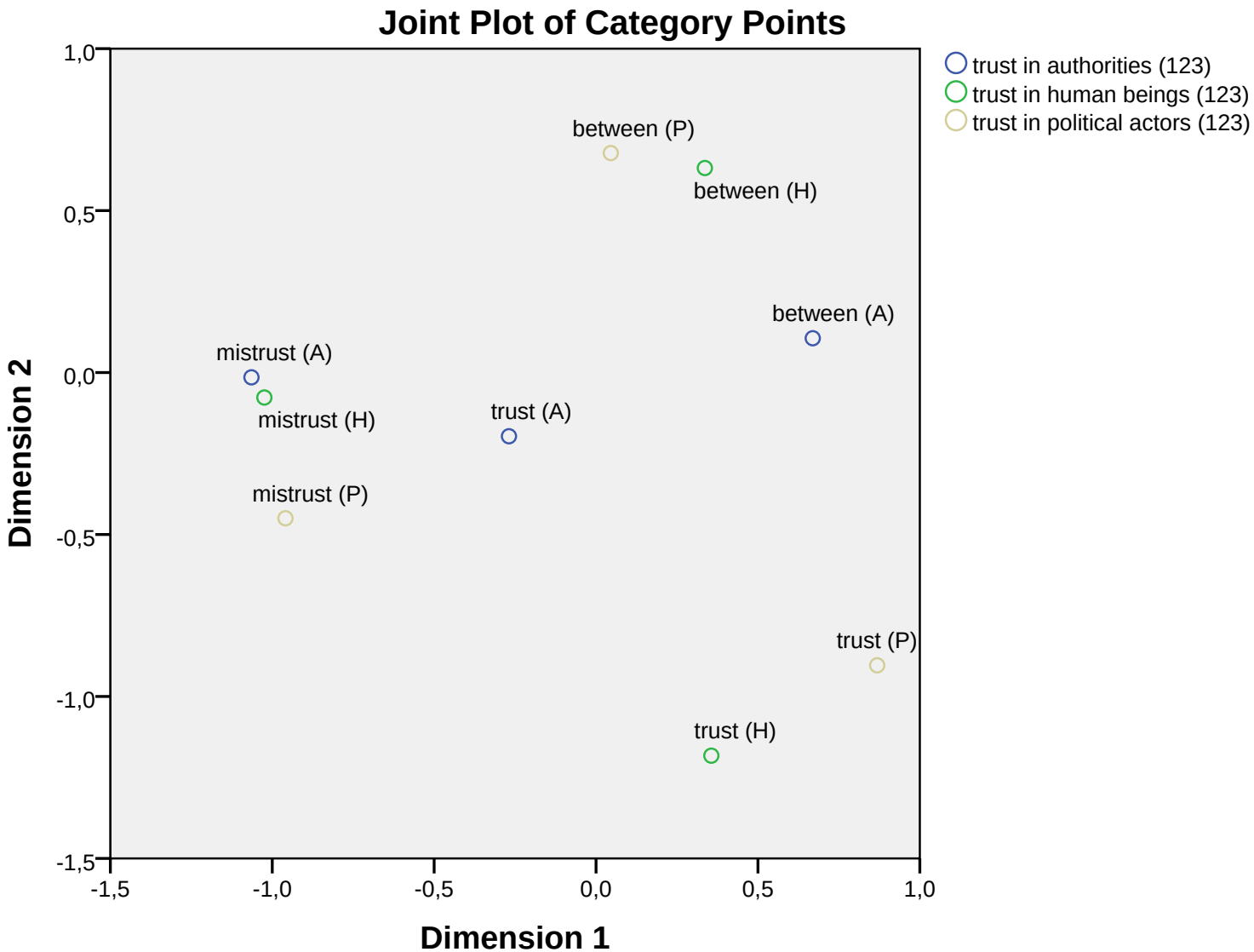
Visualizing so called *k-means* clustering using factor scores:



(clusters indicated on the previous scatter plot)

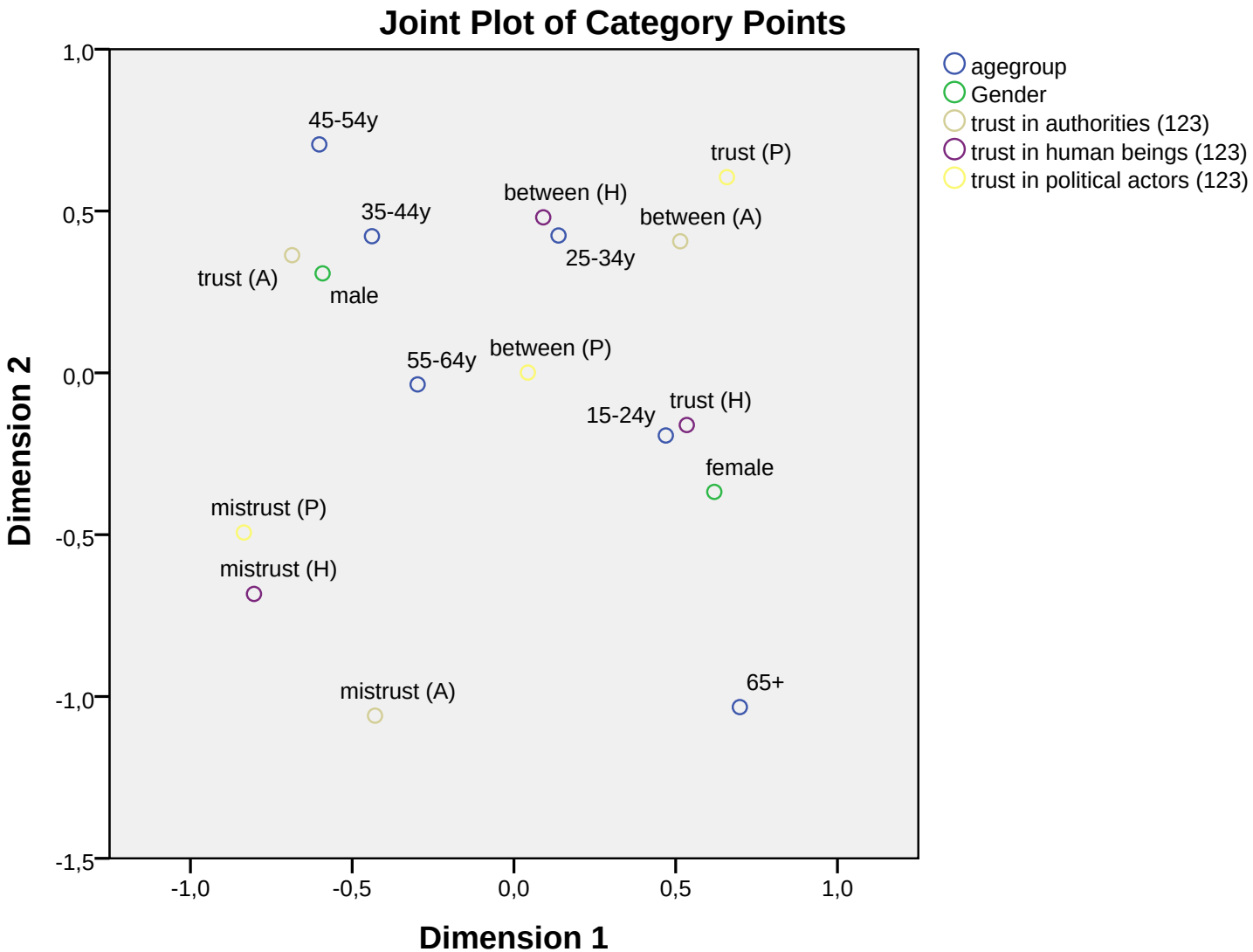
Example: correspondence analysis (ESS)

Three factor score variables, all categorized as shown earlier:



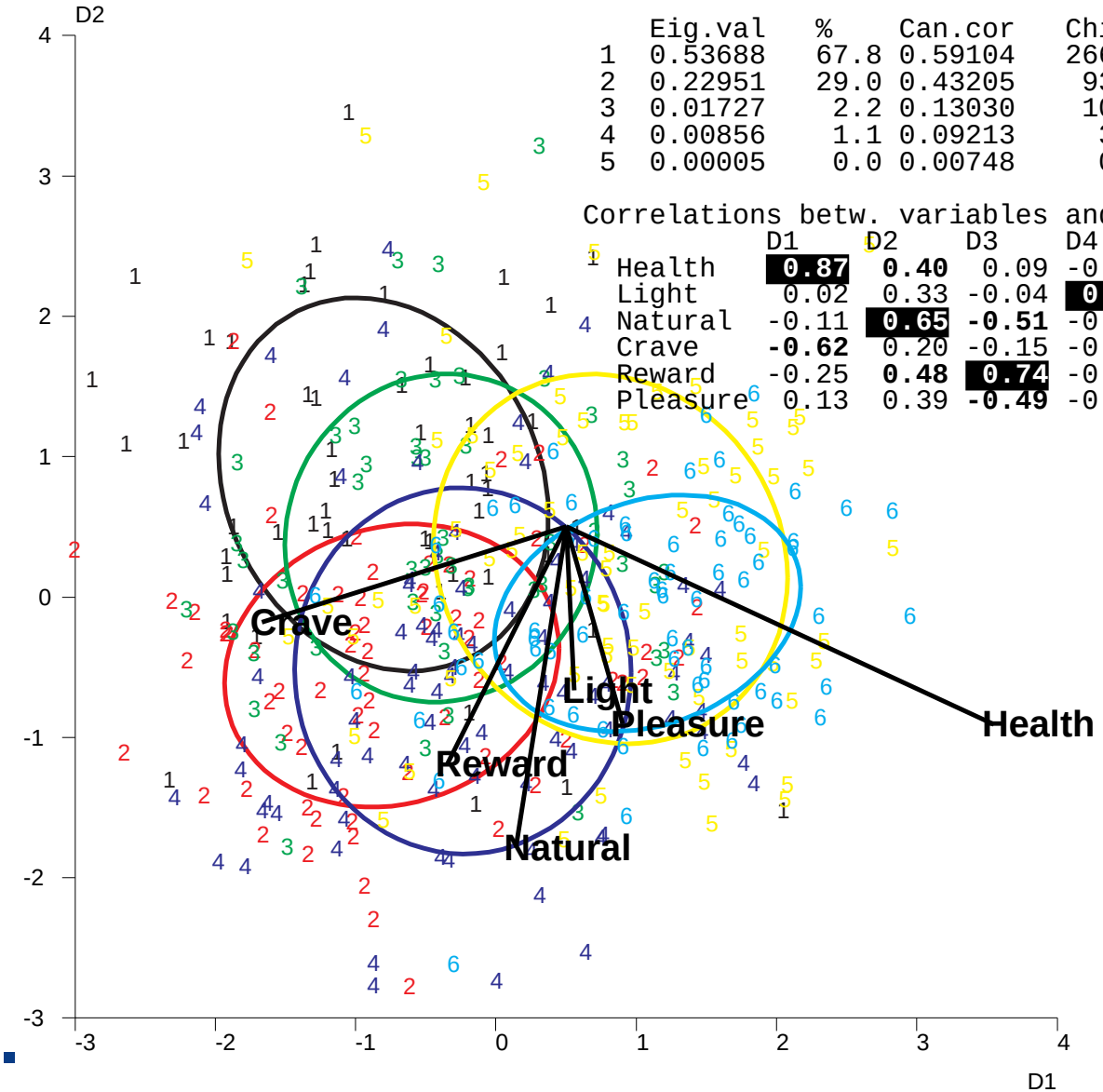
Example: correspondence analysis (ESS); background

In addition, two background variables (gender and age):



Example: visualizing the results of multivariate methods

Discriminant analysis (factor dimensions, age and gender):



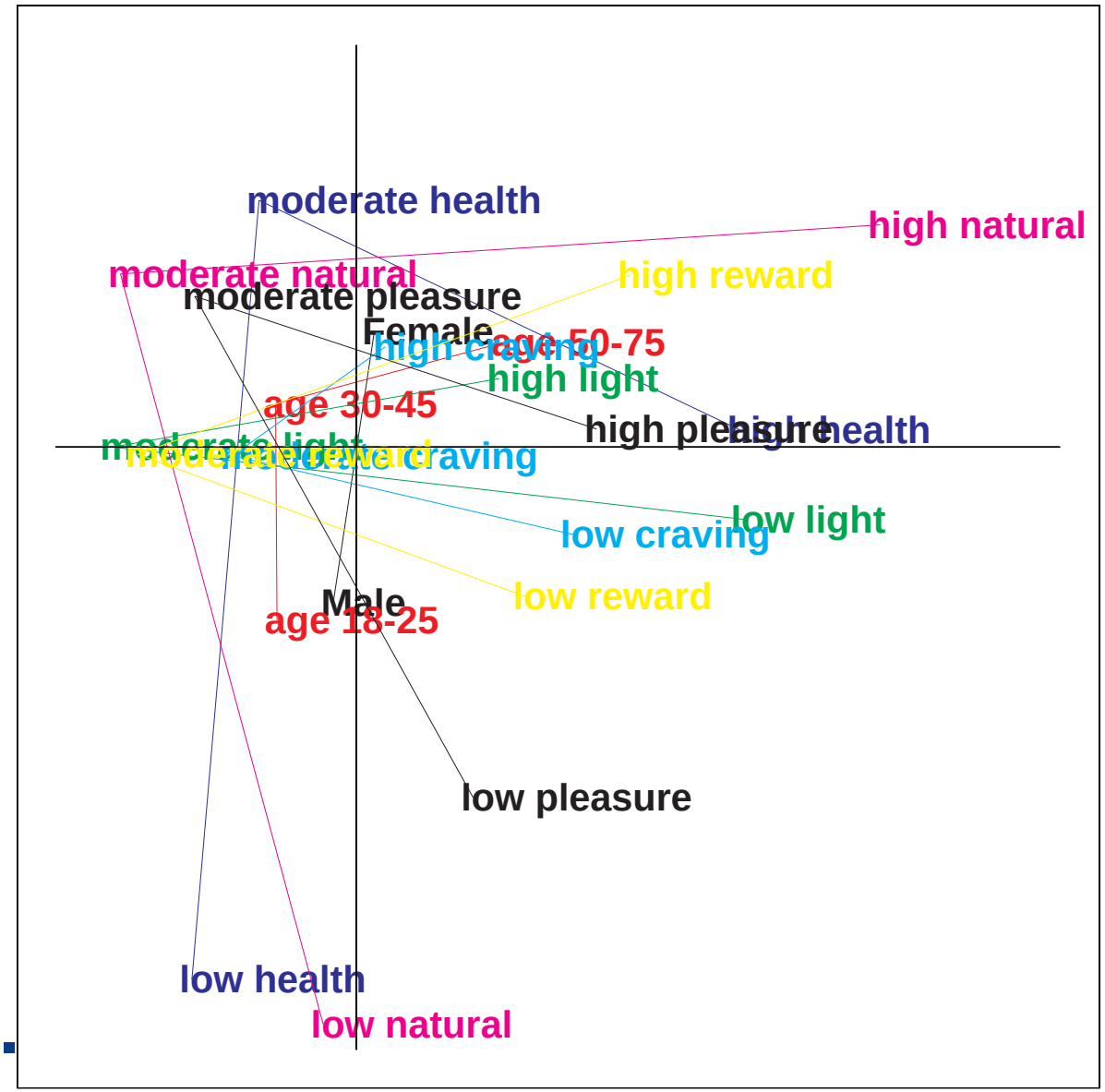
	Eig.val	%	Can.cor	Chi^2	df	P
1	0.53688	67.8	0.59104	266.15	30	0.999
2	0.22951	29.0	0.43205	93.39	20	0.999
3	0.01727	2.2	0.13030	10.33	12	0.413
4	0.00856	1.1	0.09213	3.44	6	0.249
5	0.00005	0.0	0.00748	0.02	2	0.011

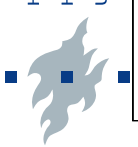
Correlations betw. variables and discriminators

	D1	D2	D3	D4	D5
Health	**0.87**	**0.40**	0.09	-0.18	0.12
Light	0.02	0.33	-0.04	**0.90**	0.13
Natural	-0.11	**0.65**	**-0.51**	-0.22	**-0.42**
Crave	**-0.62**	0.20	-0.15	-0.25	**0.68**
Reward	-0.25	**0.48**	**0.74**	-0.08	-0.01
Pleasure	0.13	0.39	**-0.49**	-0.03	0.08

Example: visualizing the results of multivariate methods

Correspondence analysis (factor dimensions, age and gender):

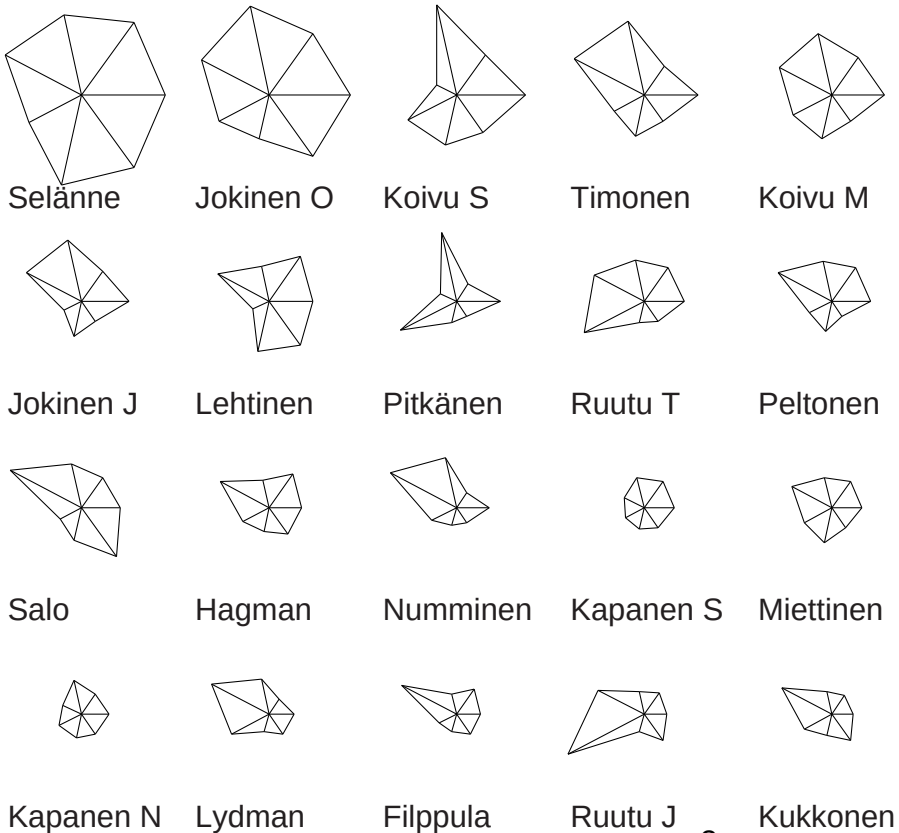




Example of visualizing player ("respondent") profiles

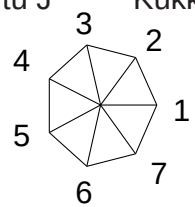
Suomalaistähtien pelaajaprofiilit

NHL:n runkosarja 2006-2007, yli 50 ottelua pelanneet



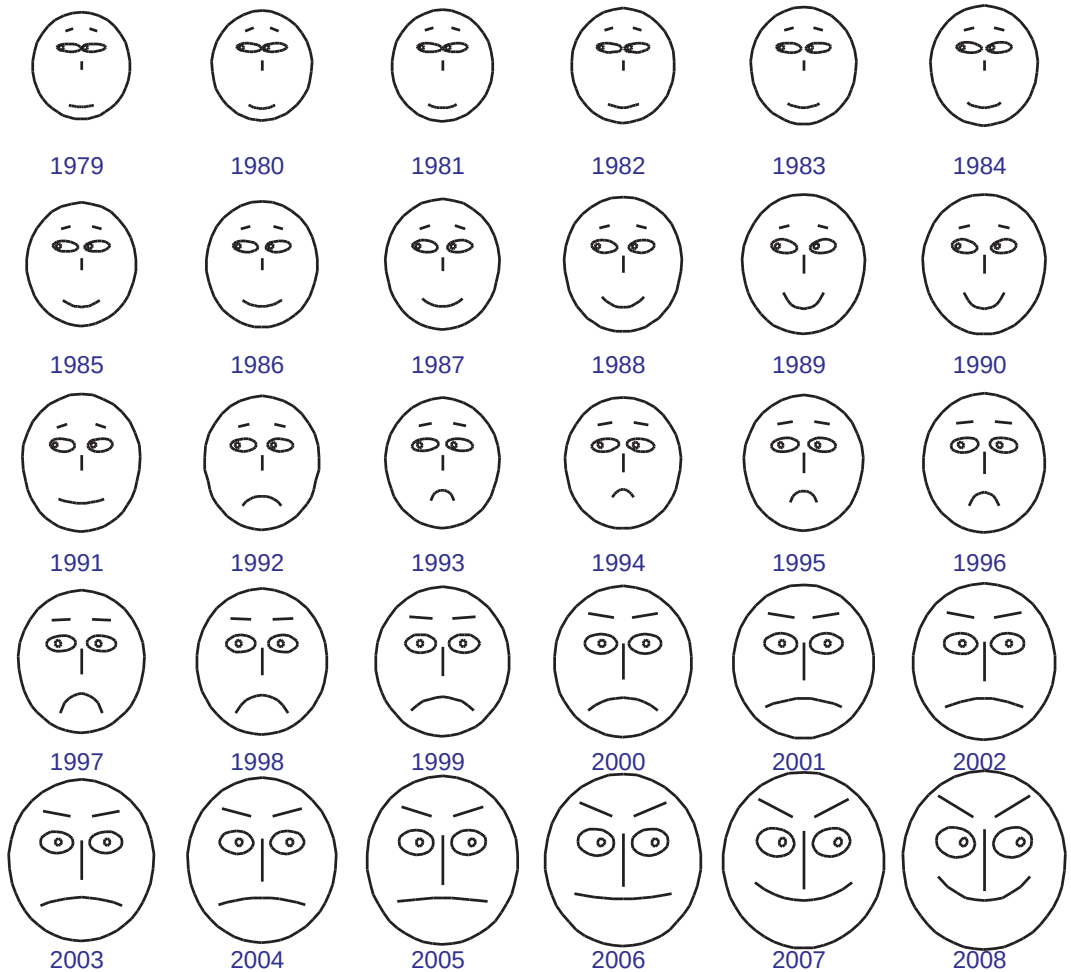
Muuttujat:

1: Pelatut pelit
2: Tehdyt maalit
3: Annetut maalisyötöt
4: +/- pisteet
5: Jäähyminuutit
6: Ylivoimamaalit
7: Voittomaalit



Example of visualizing time-series of aggregate data

Suomen vointi taloudellisin ilmein 1979 - 2008



Muuttujat: BKT | tuonti | vienti | työttömyysaste | kulutusmenot

Yhteydet kasvonpiirteisiin: *www.helsinki.fi/%7ekvehkala/naamat.html*

Aineiston lähde: Tilastokeskus *Chernoffin naamat:* **SURVO MM**

Some comments on the previous visualization

Type of the graph: **Chernoff's faces**, described in detail by `http://www.helsinki.fi/~kvehkala/naamat.html` *(translation in progress).*

Reference: Chernoff, Herman (1973). The use of faces to represent points in k-dimensional space graphically, *Journal of the American Statistical Association*, **68**, 361–368.

The five variables and their connections with the selected features of the faces:

- Gross Domestic Product
  - Shape of the head and width of the mouth
- Import of products and services
  - Length of the nose
- Export of products and services
  - Size of the eyes and direction of the look
  - Eyebrows (position, slant, size)
- Unemployment rate (men and women)
  - Curvature and vertical position of the mouth
  - Slant of the eyes
- Consumption expenditures (private and public)
  - Size of the head
  - Separation and eccentricity of the eyes

Source of data: Statistics Finland (`http://www.stat.fi/`), drawn by KV using Survo (`http://www.survo.fi`)