Survey Methodology
University of Helsinki
Part A

Fall 2015

Seppo Laaksonen

Course Information
Introduction
Survey Concepts
Examples

Content of the Whole Course

What is survey?
Other key concepts in surveys

From Survey Data Collection to Cleaned Survey Data

- Designing the survey
- Designing the questionnaire
- Designing the sample(s)
- Data collection using best possible fieldwork practices, with alternative single and mixed modes
- Data entry as much as possible during the fieldwork, or automatically
- Editing the raw data
- Imputing the data
- Weighting the data
- Adding other features into the data file

Survey Data Analysis: Basics in order to make it correctly so that the features of survey data are taken into account. Demanding analysis in other courses.

Course practice

Language in delivered material is mainly in English, in some cases in Finnish, but my e-book is completely in Finnish with a dictionary from Finnish to English in the appendix. Finnish language is also used but I cannot say exactly when and how much. My e-book is much larger than the English material but the latter may include some new things as well since the survey world has been changed in two years. It is very dynamic time going on as almost always, including

- Problems in survey climate.
- New data collection tools.
- Web is growing, old tools are possibly disappearing.
- Media is abusing survey information.
- Social media is here.
- International surveys are most interesting while small scale surveys may still be needed like as pilots for real surveys, or of a specific topic.

Course practice

1. Lectures including discussion and debating
Wednesdays 16-19 (16:15 as long as about 3x45 minutes are spent)
2. Computer Class Training with real data sets (European Social Survey and the PISA, own data are possible for extra credits), Thursdays 16-18 (16:15-17:45) or 18-20 (18:00-19:30). I hope that some are coming to the latter event.
SPSS is the main package but SAS is OK as well. R is not easily possible because meta data tools are poor and since the meta data of our data sets is fine, you can loose too much when using R.
Excel is possible to use in summarizing.
The training tasks are sent by e-mail. The reporting can be made with WORD or POWER POINT or both. The template for reporting is explained in the first training event.

Course practice

I think that I will use emails in our conversation as well and you can submit your comments by email as well. Naturally, I am most happy if I will get your feedback any time face-to-face.

The credits from the course:

- Main option = 8 if the exam has been passed successfully and the training and its report is reasonably done.
- Minimum option = 5, if the course has been made 60% (exam + training)
- Intermediate options = 6 to 7 credits as agreed mutually
- More than 8 credits if additional work is done (max =12). Agreed with myself.

Questions?

What is survey?

Wikipedia 8/2015:

A field of applied statistics, survey methodology studies the sampling of individual units from a population and the associated survey data collection techniques, such as questionnaire construction and methods for improving the number and accuracy of responses to surveys.

Statistical surveys are undertaken with a view towards making statistical inferences about the population being studied, and this depends strongly on the survey questions used. Polls about public opinion, public health surveys, market research surveys, government surveys and censuses are all examples of quantitative research that use contemporary survey methodology to answer questions about a population. Although censuses do not include a "sample", they do include other aspects of survey methodology, like questionnaires, interviewers, and nonresponse follow-up techniques. Surveys provide important information for all kinds of public information and research fields, e.g., marketing research, psychology, health professionals and sociology.

What is survey?

For me:

Survey is a series of tasks that finally results a statistical file of statistical units and their characteristics (variables). These units may be:

- Individual people
- Households and dwelling units (‘register households’ in Finland)
- Families
- Schools et al public institutions
- Enterprises
- Plants (Local units of enterprises)
- Local-kind-of-activity units of enterprises
- Villages, municipalities and other administrations
- Other areas including grid squares
- Societies, associations, corporations

Such a data file may cover basically the whole population (including register) or it can be based on a sample (= survey sampling, survey statistics).

My published definition

Source: Encyclopedia of Behavioral Medicine, 2012. Springer
The encyclopedia cover can be viewed at
http://www.springer.com/medicine/book/978-1-4419-1004-2?changeHeader.

Survey is a methodology and a practical tool used to collect, handle, and analyze in a systematic way information from individuals. These individuals or micro units can be of various types, such as people, households, hospitals, schools, businesses, or other corporations. The units can be simultaneously available from two or more levels such from households and their members. Information in surveys may be concerned various topics such as people's personal characteristics, their behaviour, health, salary, attitudes and opinions, incomes, poverty and housing environments, or characteristics and performance of businesses. Survey research is unavoidably inter-disciplinary, although the role of statistics is most influential since the data for surveys is constructed in a quantitative form. Correspondingly, many survey methods are special statistical applications. However, surveys exploit substantially many other sciences such as informatics, mathematics, cognitive psychology, and theory of subject-matter sciences of each survey topic.

There are other concepts in the area, e.g. such that are mentioned on the website of a fairly new American Journal (J. of Survey Statistics and Methodology = JSSM). This journal have three sections.
http://www.oxfordjournals.org/our_journals/jssam/about.html

The Survey Statistics section will present papers on innovative sampling procedures, imputation, weighting, measures of uncertainty, small area inference, new methods of analysis, and other statistical issues related to surveys.
The Survey Methodology section will present papers that focus on methodological research, including methodological experiments, methods of data collection and use of paradata.
The Applications section will contain papers involving innovative applications of methods and providing practical contributions and guidance, and/or significant new findings.

JSSM (cont.)

Its objective is to publish cutting edge scholarly articles on statistical and methodological issues for sample surveys, censuses, administrative record systems, and other related data. It aims to be the flagship journal for research on survey statistics and methodology. Topics of interest include survey sample design, statistical inference, nonresponse, measurement error, the effects of modes of data collection, paradata and responsive survey design, combining data from multiple sources, record linkage, disclosure limitation, and other issues in survey statistics and methodology.

Statistics Canada publishes its Survey Methodology journal dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves.

Key concepts in surveys

Next pages concentrate on
- Populations in surveys
- Cross-sectional vs. longitudinal surveys
- Other basic concepts such as Meta data and para data

After that I will go to
- Sampling design with inclusion probabilities etc.
- Missingnesses and other deficiencies
- Statistical Editing
- Imputation
- Sampling and other weights

Populations in surveys 1

Population is a key concept of statistics, determined by Adolphe Quetelet in 1820's. This is not just one in surveys where I need even five populations:

1. *Population of interest* is the population that a user would like to get or estimate ideally but it is not possible always to completely reach and hence she/he determines
2. *Target population* which is such a population that is realistic. Naturally, this population should be exactly determined including its reference period (a point of time or a time period).

The target population of the ESS e.g. "Persons 15 years or older who are residents within private households in the country in the first of November."
Correspondingly to the EFSS (European Finnish Security Survey): 15-74 years old non-Swedish speaking residents in Finland 1st of October 2009.
The Programme for International Student Assessment (PISA) survey: fifteen-year-old school students (this is specified so that the full calendar year is covered). The grid-based study of Finland: People from 25 to 74 years of age living in southern Finland whose mother tongue is either Finnish or Swedish.

Discussion about two first populations

I think that it is good first to try to find an ideal population = population of interest. It is not possible in most cases, e.g. if the target is to get the voting population, it is not possible. Hence, the population who are eligible to vote is reasonable. Even though this population is not used e.g. in Finnish opinion polls where 80 years old or older are not included and people living outside Finland not either. The latter ones are hard to reach, but the olds are excluded because this has been considered to be easier for survey organizations (TNS Gallup, Taloustutkimus, etc.). Before the last elections this was criticized by well-known older citizens. Note that the ESS gives opportunity to analyze residents of all ages above 15 including their voting behavior. Naturally, the results are not fresh.

In general, as soon as you are designing the survey and its target population, use much time with your team to think and discuss what is your population of interest and how close to this you can go with a realistic target population and the frame population, next page.

Discussion about two first populations
There is another international survey SHARE (Survey of Health, Ageing and Retirement in Europe) without age limit in which people aged 50 and older are included. Finland has not been participated in this but 20 European countries from Sweden to Greece and Portugal to Estonia are. I think that this survey is becoming more and more important.
Naturally, there are many other areas that have the same role. May you think which survey topics could be important?

Populations in surveys 2

In order to get the target population you need

3. *Frame population and the frame* from which the statistical units for the survey can be found. Usually, the frame is not exactly from the same period as the target population (delay in Finnish population surveys is rather short i.e. 1-5 months, but for enterprise surveys much more, even some years).

The frame is not always at element level available as in the case of population register based surveys. Instead, the frame population can be as follows:

Stage 1: List of the electoral sections (e.g. in a certain country their number is 12,313 and they cover the whole territory of the country).
Stage 2: Lists of all households' addresses of the at the first stage selected units.
Stage 3: One or more members of the selected household/address

There are here thus three frames, but it is possible that this number can be even four such as municipalities, blocks or villages or census districts, addresses, people at certain ages, among others.

Populations in surveys 3

Due to the delay in the frame,

4. *Updated frame population* is useful for estimating the results better. Usually, the initial frame population has been used for estimation too. This may lead to biased estimates. Fortunately, this bias is not severe in most human surveys. At contrast, old frames can lead to dramatic biases in business surveys, if this is due to large businesses.

After the data collection or fieldwork we are able to determine

5. *Study population or survey population.*

It is ideal if this fifth population corresponds to our target population or even the population of interest. But if not, our estimates are somewhat biased.

Generalization of the survey results = Estimation

Before continuing with survey terms it is good to discuss about the question what is the purpose of these populations. Naturally, the first point is to approach to the targets of the survey as well as possible, and hence it is needed to know all steps and possible gaps passed or hopefully solved. But the final target is to estimate the desired estimates, such as averages, standard deviations, medians, distributions, ratios and statistical model parameters.

This can be made just calculating whatever ways but such figures cannot be generalized at any population level without using survey instruments that will be learned during this course. If all coverage and related problems are solved, the estimates can be generalized at the target population level that is just mainly considered in this course.

Generalization of the survey results = Estimation

If this population can not be well achieved, it is best to speak about the generalization at the study population level. It is not common to report the surveys in this way although the reality is that certain groups are not really represented among the respondents. For example: homeless, disabled, other marginalized, people who do not understand languages used in the survey.

It is possible to try the generalization in another way, for example using modelling, but this issue is special and cannot be considered in this course. This generalization is mainly concerned certain connections or explanations found from the data. It thus is possible to try to generalize such 'estimates' or other outcomes in some way.

Populations in surveys 4

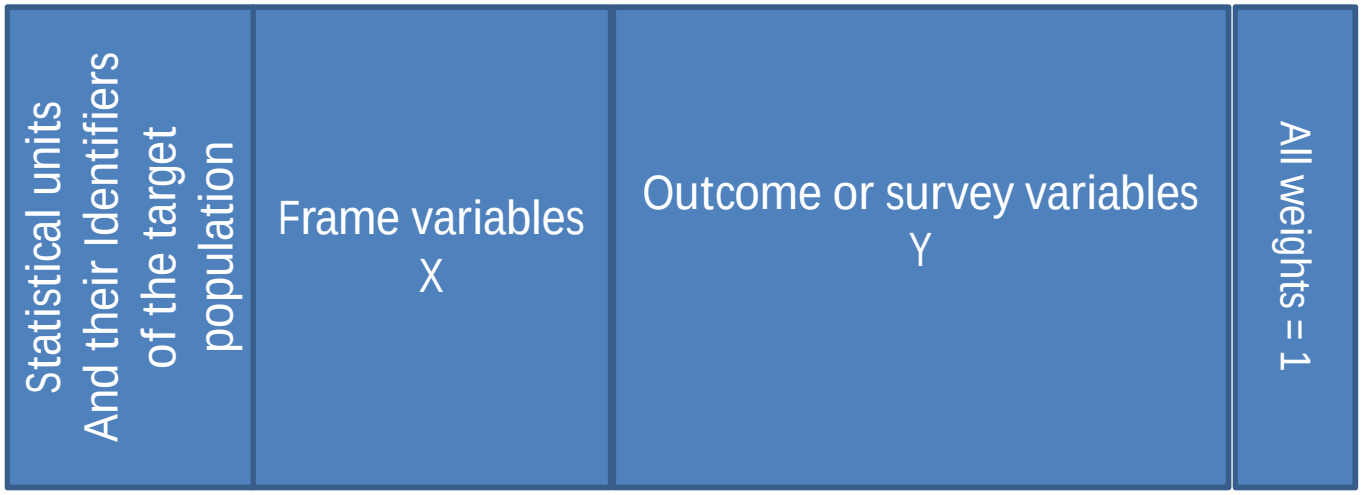
The units of the target population are equal to those of the study population but the units of the frame population(s) can be essentially different.
The ESS survey designs vary a lot from one country to the next. There are such countries where all the units are equal = individuals 15+ (Finland, Sweden, ...) but many countries have several units (small areas, addresses, households, 15+ years individuals, ...).

PISA and other student surveys use typically two units:
- 'PISA' Schools (or school classes)
and
- Students themselves

Give other examples.
The next two pages illustrate missingnesses as well as some other crucial concepts in surveys. The first is the case of a cross-sectional survey.

Micro data for the entire target population

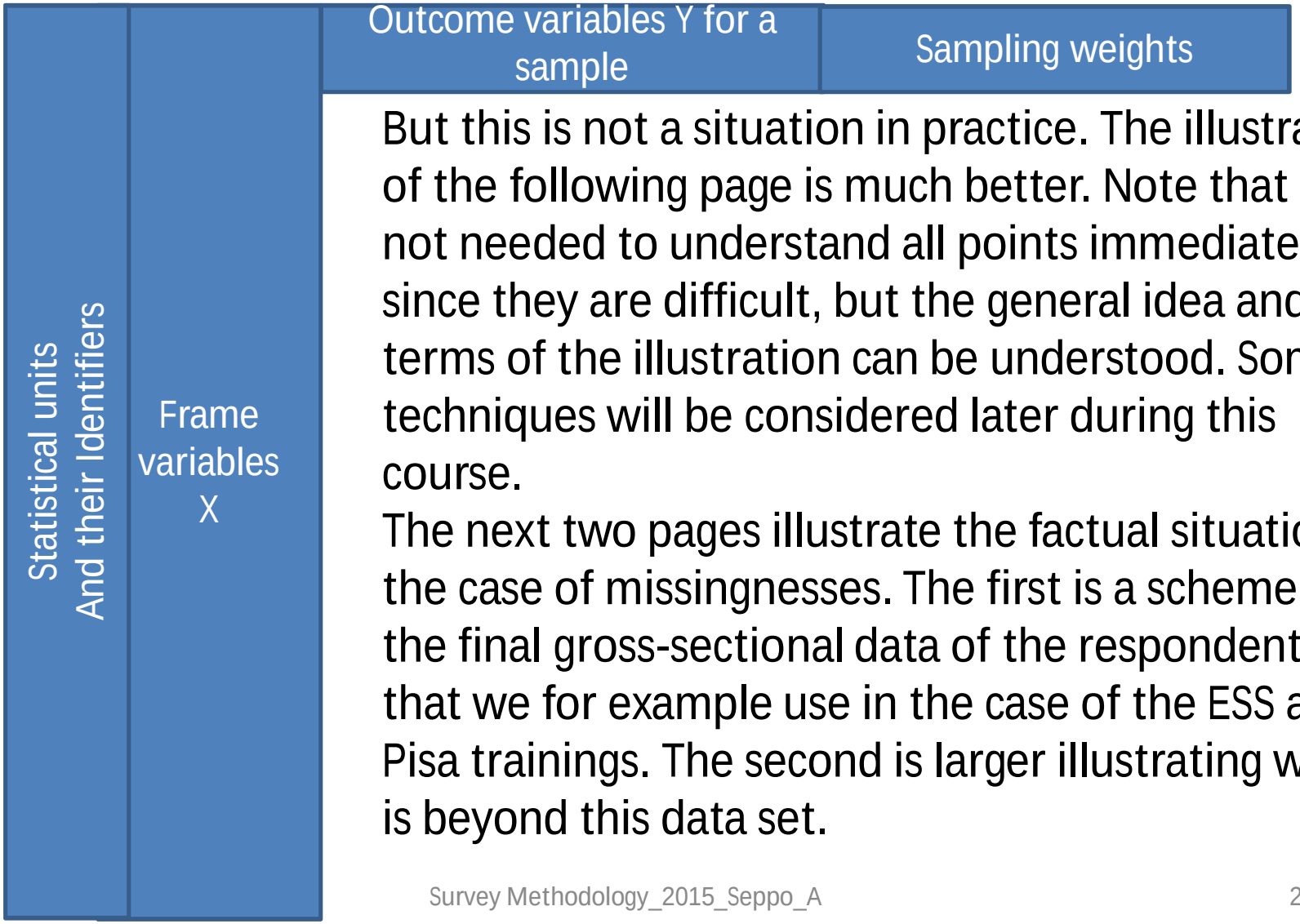
Now I focus on micro data. I start with a cross-sectional case. If the whole target population has been examined, and any missingnesses occur, it is simple as the following scheme illustrates.



We have here the term 'weight' that is a basic tool for generalizing the results. In this entire population data set, all weights are equal to one and hence the weights do not need to be used factually at all. The generalization is concerned estimates based on outcome variables. Frame variables are just getting values of these *Y* variables by a survey.

Micro data without missingness

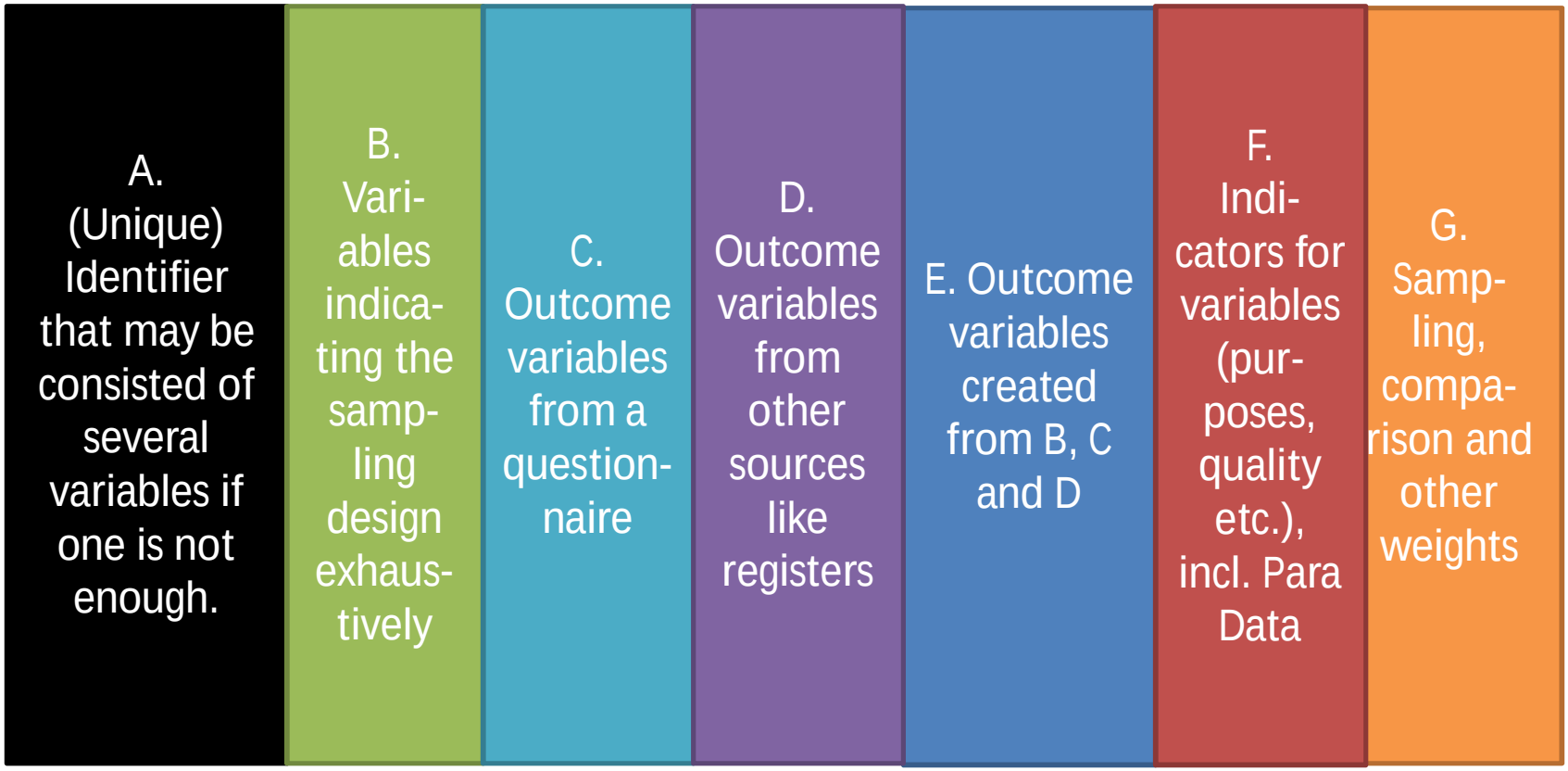
But most surveys are based on a sample. Below is a simple illustration for this case without non-response, that is, the gross sample units reply completely:



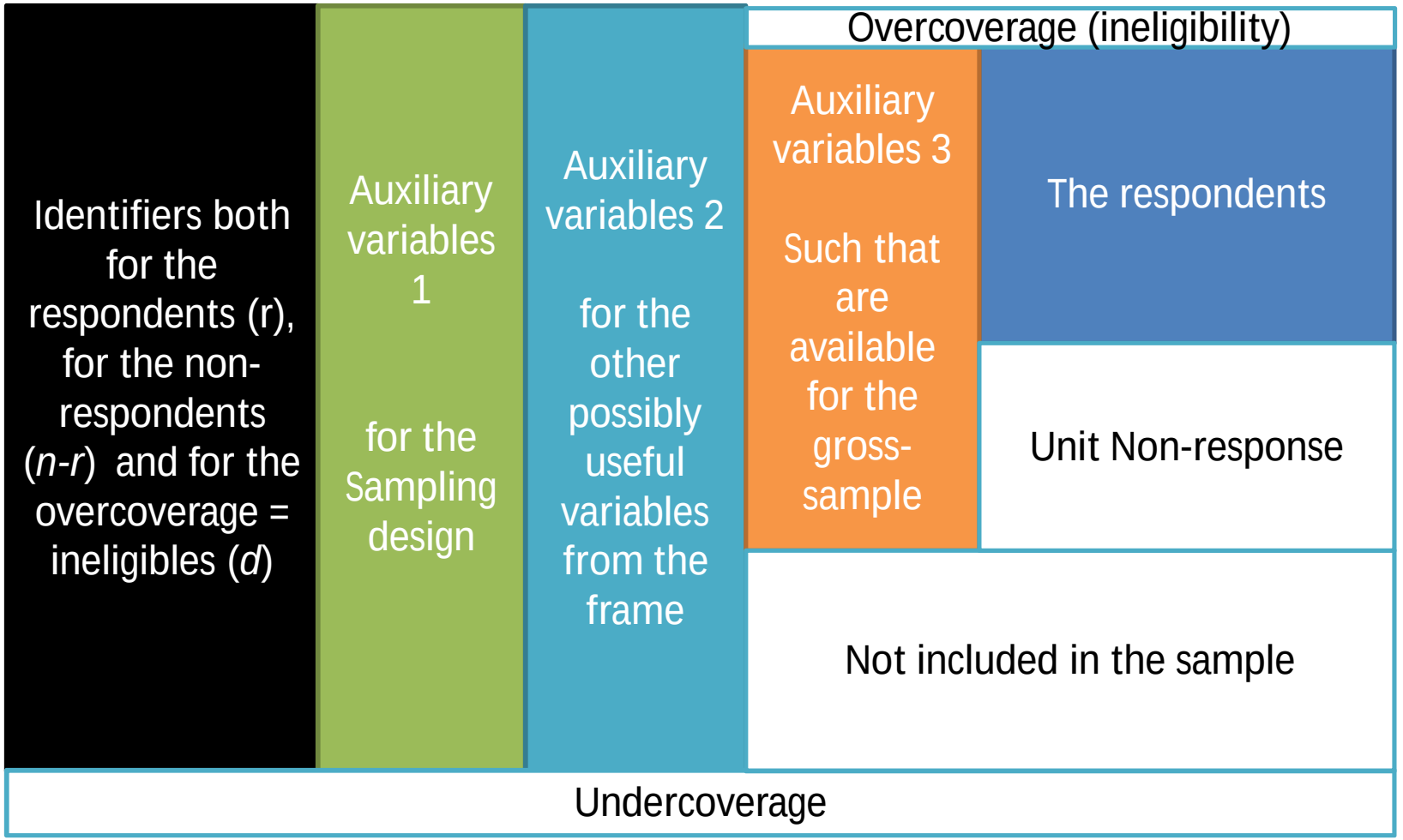
But this is not a situation in practice. The illustration of the following page is much better. Note that it is not needed to understand all points immediately, since they are difficult, but the general idea and the terms of the illustration can be understood. Some techniques will be considered later during this course.

The next two pages illustrate the factual situation in the case of missingnesses. The first is a scheme of the final gross-sectional data of the respondents that we for example use in the case of the ESS and Pisa trainings. The second is larger illustrating what is beyond this data set.

General structure of a micro level cross-sectional survey data file that consists of *r* respondents (rows of the matrix).
It is possible that there are outside this scheme other data, e.g. more para data, and content data. Good meta data should be available for all variables.



General scheme of the cross-sectional micro file in which the previous is one box (marked ———). Box sizes do not correspond to any real situation.



Micro data and Missingness

Examples of the terms in social surveys:

Overcoverage (in-eligibles): died, emigrants, errors in the frame

Some of them can be observed during the fieldwork, not all. This is worsening problem nowadays since if not contacted it is difficult to know whether a unit belog to this group or to unit non-response.

Undercoverage: new born, new immigrants, illegally living in a country, errors in the frame. Updated frame helps to find them.

Unit non-response: not contacted, disable to participate, refusals, ...

Sampling weights are of two types:

- Their average is for each target population = 1 and hence their sum = the number of the respondents
- Their sum = the number of the target population units (households or individuals, etc.) and each weight indicates how many units one unit represents in the target population; thus these weights are for generalizing the results?

A real survey file is not such as the scheme of the below page, except in some special cases like methodological experiments using simulations, for example. There are two real files:

- Sampling design (data) file that covers the gross sample units and auxiliary variables. I have a special (good) example in another document of the course website. From this file we usually create the sampling weights and other sampling design variables and merged these into 
- The file of the respondents that thus is used in analysis. In the end of this section, I give examples of such survey files and of course we will later use these in our analysis.

Comment on variables:

As noticed, there are *X* and *Y* variables and they have a special role. However, a *X* variable can be used as a *Y* variable as well in the analysis but in this case, their values are only for the respondents.

Moreover, *Y* variables can be of different kinds:

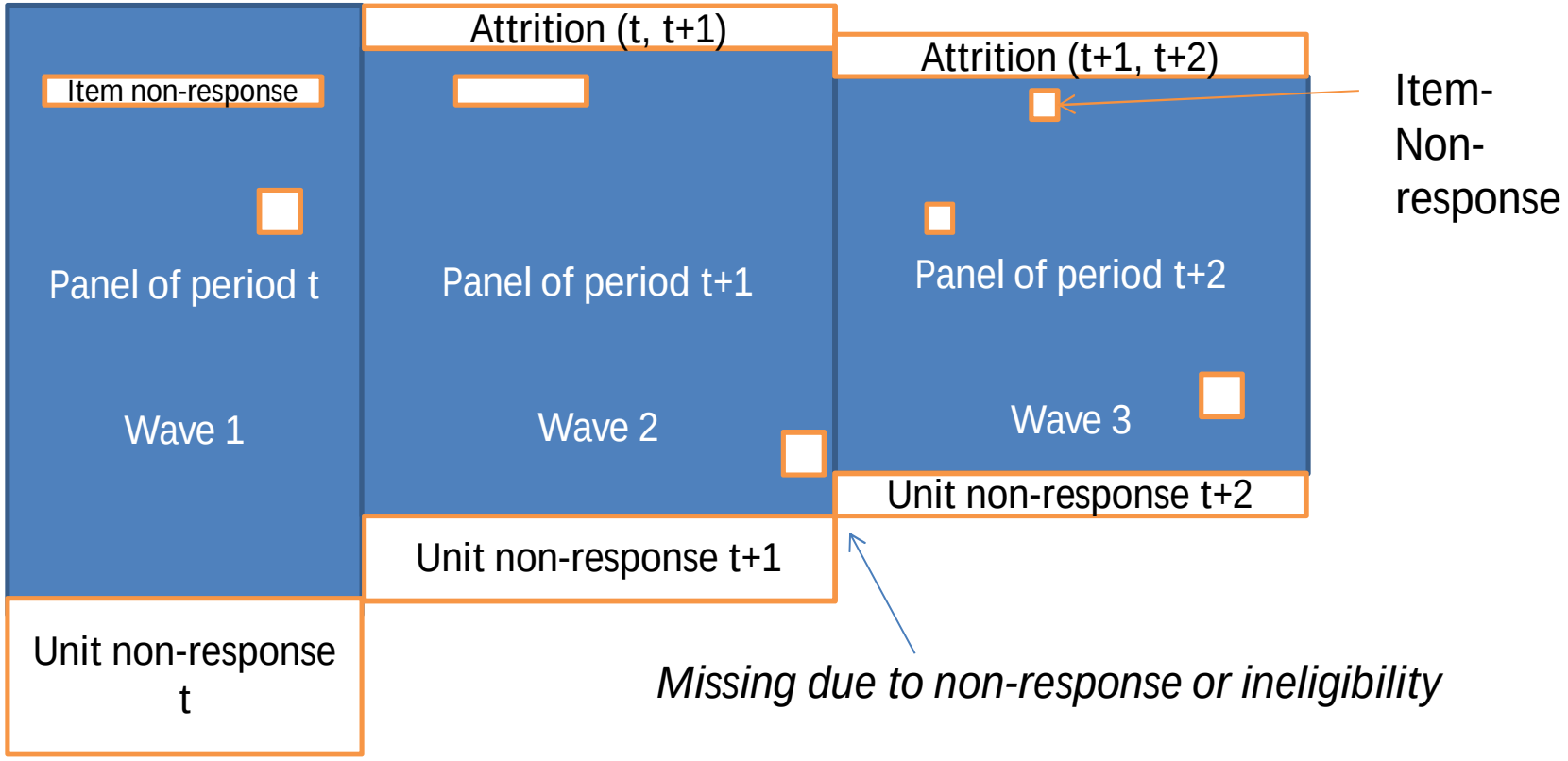
- Initial survey questionnaire variables and exactly in the same form as in the questionnaire.
- But these initial *Y* variables can be transformed into another form (a new scale) as well, in order to facilitate analysis.
- Summary variables from another source like student (PISA) exams, clinical examinations, ...
- Aggregated information e.g. from a living area characteristics (the same value for all living in this area).
- What else?

Micro data and Missingness

Cohort type of panel (longitudinal) example

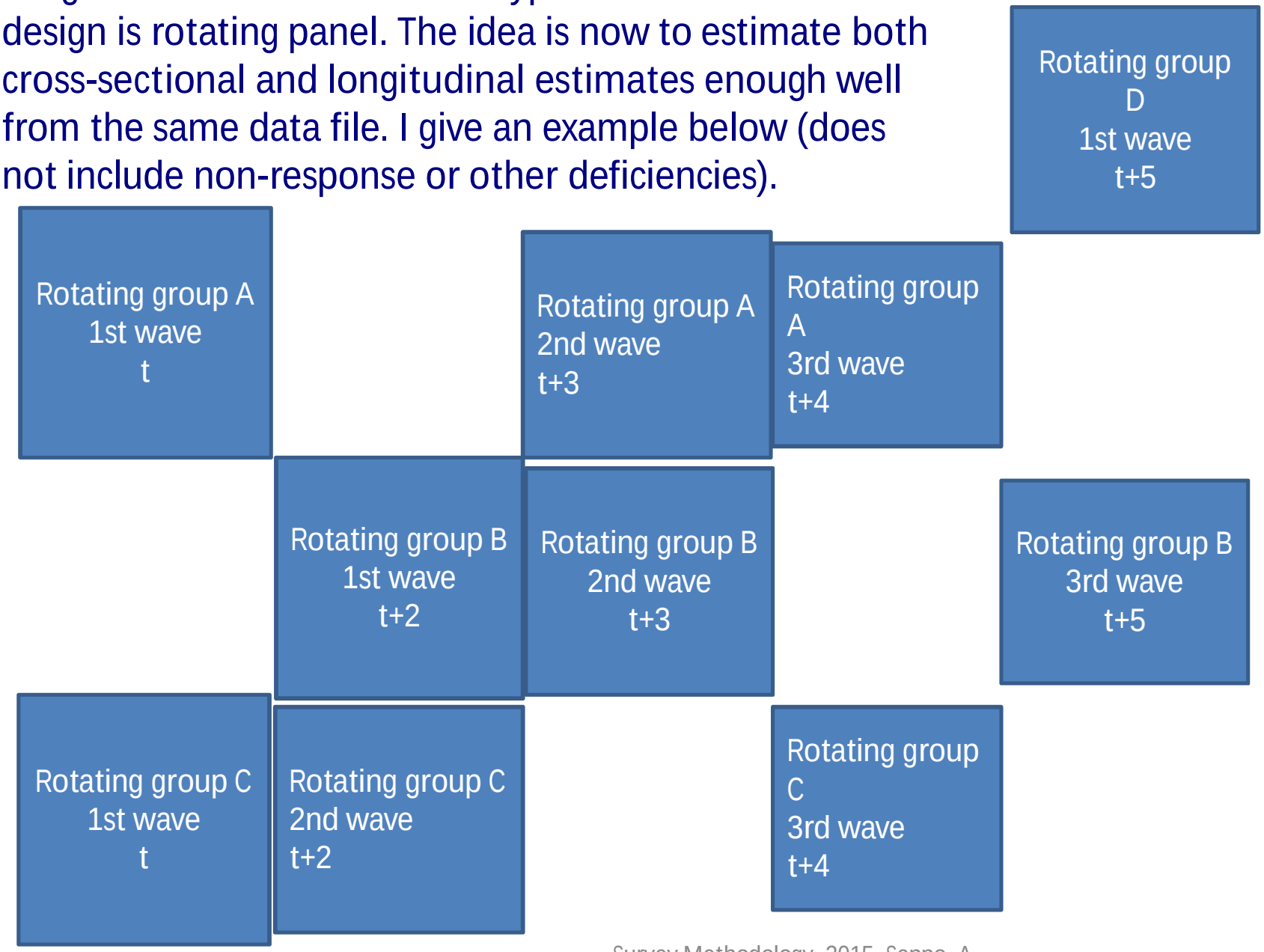
Here attrition does not include unit non-response as in some other studies

Obs. A big issue currently is how to update non-response units of older waves, since it is possible that they are no more non-respondents but ineligible (died, outside the target population).

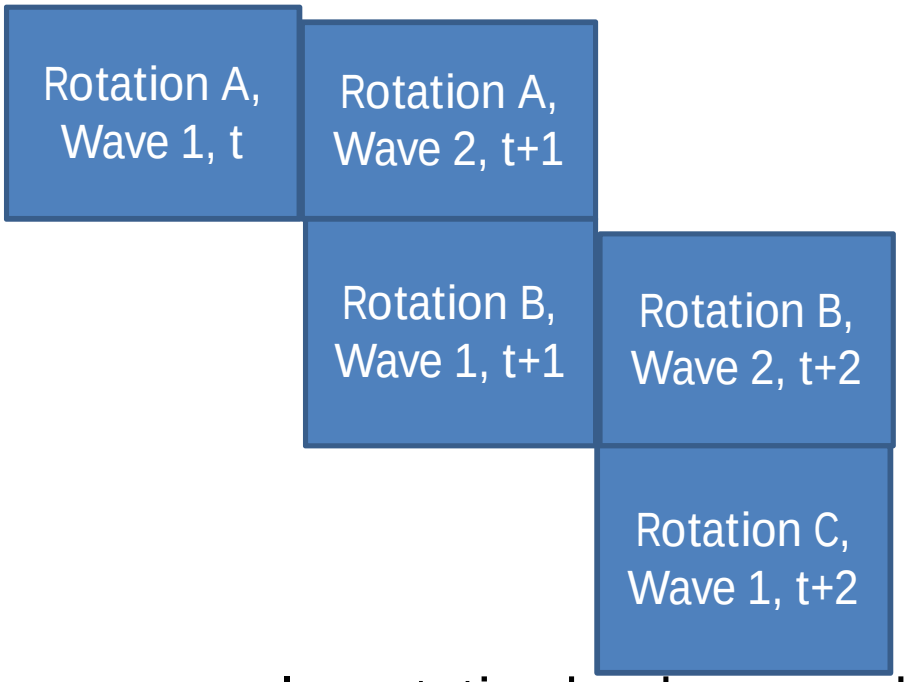


Rotating panel

Longitudinal can be of various types. One common design is rotating panel. The idea is now to estimate both cross-sectional and longitudinal estimates enough well from the same data file. I give an example below (does not include non-response or other deficiencies).



I have applied 20 years ago a rotating panel design for the Statistics Finland Income survey. This is maybe the simplest possible rotating design, since the panel covers only two years.



A more complex rotation has been used e.g. in Labour force surveys of all European countries. The reason is that they wish to e.g. estimate both unemployed rates and changes in these rates in their longitudinal meanings. Most income surveys have the same purpose, e.g. how permanent is poverty or riches?

X variables, auxiliary variables in more details

These variables can be found, collected and/or downloaded from the different sources, as follows:

- Population register (e.g. age, gender, members living in the same address, house type and size, kitchen type, area,...)
- Other registers such as tax register or job seekers' register, formal education register (e.g. tax income, unemployed, education, ...)
- Other administrative sources, often at aggregate level (e.g. % owner occupation, % social renting, % detached housing, % divorced, % under crowding, % 2 or more cars, 1 or more cars, % owner occupation, % unemployed, % long term unemployed, % social renting, % highly educated); the aggregate here may vary, being e.g. municipality, postal area code, grid square, block, village).

X variables, auxiliary variables in more details

- Using interviewer observations of the immediate vicinity of the houses of sample units about visible signs of neighbourhood disorder. These observations of neighbourhood disorder or decay can be linked to the ‘broken windows’ hypothesis. The neighbourhood has been classified into one or more variables by an interviewer using harmonised rules. This type of X variables is becoming more common but it is difficult get, regularly in particular. So far collected in pilot studies.

These X variables have several roles those will be more discussed later during this course but there are some below:

- Quality analysis of the survey data themselves
- Quality analysis of the data collection process
- Identify reasons for nonresponse and ineligibility
- Compute ineligibility rates = number of ineligibles/gross sample size
- Compute response rates = number of respondents/(gross sample size – ineligibles)
- Compute item nonresponse rates and other characteristics of them
- Use the data for weighting and reweighting
- Use the data for checking and other editing
- Use the data for imputations.

The meta data means information about the data. This information is available in different formats:

- The questionnaire is the most important concerning survey variables
- The survey methodology documentation is as important, including sampling, fieldwork, IT tools

The meta data of the survey micro file is currently well available e.g. in SPSS, SAS, etc. general software packages. These includes

- The variable label meta data that is usually a short description of the corresponding questionnaire text.
- In the case of a categorical variable, their labels respectively.
- Possibilities to include the information about missingness and the range of the values.

The following pages illustrate the role of the meta data for the European Social Survey.

Survey micro data that is hard to understand without meta data. If I say that that this is from the Pisa 2012, it helps but not enough. On the next two pages, there are first values with their labels and then the variable names as well.

	CNT	STRATUM	SCHOOLID	ST04Q01	ST11Q01	ATSCHL	IMMIG	PARED	SCMAT	W_FSTUWT	math_stat	science
1	ITA	ITA0401	0000424	1	1	.	1	13,00	.	109,31	427,57	372,75
2	JPN	JPN0203	0000031	1	1	.	1	16,00	.	253,63	471,42	502,46
3	SVK	SVK0005	0000045	1	1	,24	1	13,00	-,52	13,20	539,27	565,96
4	BEL	BEL0112	0000100	2	1	.	1	17,00	.	17,30	511,15	560,74
5	HKG	HKG0002	0000011	2	2	.	2	11,00	.	15,41	519,87	574,08
6	BEL	BEL0102	0000169	1	1	.	1	15,00	.	12,36	622,77	605,32
7	CAN	CAN0210	0000474	1	1	.	1	15,00	.	1,15	406,92	461,06
8	CAN	CAN0437	0000501	2	.	.	1	17,00	.	3,00	469,40	508,24
9	DEU	DEU9797	0000064	1	1	.	2	10,00	.	129,85	340,87	305,33
10	AUS	AUS0205	0000694	1	1	.	1	12,00	.	16,39	613,27	671,99
11	CAN	CAN0542	0000227	2	1	.	2	12,00	.	2,06	673,17	615,11
12	ARE	ARE0321	0000390	1	1	1,38	1	16,00	,88	4,23	471,81	506,01
13	FIN	FIN0016	0000232	2	1	-,64	1	12,00	1,38	2,89	613,27	566,71
14	CHL	CHL0316	0000100	2	1	.	1	16,00	.	12,79	646,14	631,15
15	BGR	BGR0007	0000131	2	1	-,64	1	17,50	,88	9,80	396,56	383,10
16	RUS	RUS9797	0000081	2	1	-,95	1	13,50	-,06	207,26	350,14	388,23
17	URY	URY0008	0000145	2	.	-,24	1	12,00	-,52	5,19	339,00	372,57
18	CAN	CAN0322	0000131	1	1	-1,46	1	17,00	-,76	5,39	376,70	446,23
19	FIN	FIN0001	0000288	2	1	.	1	12,00	.	5,05	509,74	517,38
20	ITA	ITA2101	0000515	1	1	.	1	12,00	.	7,51	528,13	569,41
21	JPN	JPN0202	0000164	1	1	,77	1	14,00	2,18	160,14	380,87	376,67

The previous file is SPSS, this is SAS as the next one too.

	CNT	STRATUM	SCHOOLID	ST04Q01	ST11Q01	ATSCHL	IMMIG	PARED	SCMAT	W_FSTUWT	math_stat	science	reading
1	Italy	ITA - REGION 04 stratum 01 : [Region 04] Emilia Romagna Licei	0000424	Female	Yes	M	Native	13	M	109.312	427.5666	372.7527	427.1
2	Japan	JPN - stratum 03 : 03: Private and Academic	0000031	Female	Yes	M	Native	16	M	253.631	471.4208	502.46176	502.40
3	Slovak Republic	SVK - stratum 05 : Region 2 / GYM	0000045	Female	Yes	0.24	Native	13	-0.52	13.2011	539.26624	565.96418	581.67
4	Belgium	BEL - stratum 12 : Flanders/Gen Ed and other type/network comm. Schls/Mixed ISCED	0000100	Male	Yes	M	Native	17	M	17.2979	511.14658	560.74224	472.64
5	Hong Kong-China	HKG - stratum 02 : Aided or Caput	0000011	Male	No	M	Second-Generation	11	M	15.4062	519.8707	574.07686	565.6
6	Belgium	BEL - stratum 02 : Flanders/Gen Ed only/private subsidized/ISCED3	0000169	Female	Yes	M	Native	15	M	12.3564	622.76834	605.31518	613.76
7	Canada	CAN - stratum 10 : prov 11 Fr. and Eng. 0 to ...	0000474	Female	Yes	M	Native	15	M	1.1471	406.92474	461.0593	452.84
8	Canada	CAN - stratum 37 : prov 13 Fr. 118 to ...	0000501	Male	M	M	Native	17	M	2.9962	469.39552	508.2432	408.89
9	Germany	Undisclosed STRATUM - Germany	0000064	Female	Yes	M	Second-Generation	10	M	129.849	340.87086	305.33382	382.46
10	Australia	AUS - stratum 05 : NSW_Government	0000694	Female	Yes	M	Native	12	M	16.3875	613.2653	671.98806	665.15
11	Canada	CAN - stratum 42 : prov 24 Eng. 35 to 305	0000227	Male	Yes	M	Second-Generation	12	M	2.0571	673.1656	615.10632	628.2
12	United Arab Emirates	RE - stratum 21 : SHARJAH,MOE,PUBLIC	0000390	Female	Yes	1.38	Native	16	0.88	4.2319	471.81024	506.00522	533.54
13	Finland	FIN - stratum 16 : Swedish/Not Aland-Urban-Low	0000232	Male	Yes	-0.64	Native	12	1.38	2.8857	613.2653	566.71018	564.47
14	Chile	CHL - stratum 16 : 16: Private / Primary and Secondary / HS	0000100	Male	Yes	M	Native	16	M	12.794	646.13646	631.14508	611.70
15	Bulgaria	BGR - stratum 07 : Region 07	0000131	Male	Yes	-0.64	Native	17.5	0.88	9.8047	396.56486	383.10332	325.41
16	Russian Federation	Undisclosed STRATUM - Russian Federation	0000081	Male	Yes	-0.95	Native	13.5	-0.06	207.257	350.1402	388.232	347.30
17	Uruguay	URY - stratum 08 : Public Technical Secondary Schools,Mixed	0000145	Male	M	-0.24	Native	12	-0.52	5.1942	339.00138	372.5662	330.7
18	Canada	CAN - stratum 22 : prov 12 Eng. 35 to ...	0000131	Female	Yes	-1.46	Native	17	-0.76	5.3938	376.70198	446.23276	471.50
19	Finland	FIN - stratum 01 : South-Urban-High	0000288	Male	Yes	M	Native	12	M	5.05	509.7445	517.3816	451.3
20	Italy	ITA - REGION 21 stratum 01 : [Region 21] Valle dAosta Licei	0000515	Female	Yes	M	Native	12	M	7.5087	528.12744	569.4144	568.17

Now our data are quite easy to understand due to meta data exploited. Naturally, in order to understand well, it is required to read other Pisa documents too.

	Country code 3-character	Stratum ID 7-character (cnt + region ID + original stratum ID)	School ID 7-digit (region ID + stratum ID + 3-digit school ID)	Gender	At Home - Mother	Attitude towards School: Learning Outcomes	Immigration status	Highest parental education in years	Mathematics Self-Concept	FINAL STUDENT WEIGHT	Mathematical-Statistical Literacy	science
1	Italy	ITA - REGION 04 stratum 01 : [Region 04] Emilia Romagna Licei	0000424	Female	Yes	M	Native	13	M	109.312	427.5666	372.7527
2	Japan	JPN - stratum 03 : 03: Private and Academic	0000031	Female	Yes	M	Native	16	M	253.631	471.4208	502.46176
3	Slovak Republic	SVK - stratum 05 : Region 2 / GYM	0000045	Female	Yes	0.24	Native	13	-0.52	13.2011	539.26624	565.96418
4	Belgium	BEL - stratum 12 : Flanders/Gen Ed and other type/network comm. Schls/Mixed ISCED	0000100	Male	Yes	M	Native	17	M	17.2979	511.14658	560.74224
5	Hong Kong-China	HKG - stratum 02 : Aided or Caput	0000011	Male	No	M	Second-Generation	11	M	15.4062	519.8707	574.07686
6	Belgium	BEL - stratum 02 : Flanders/Gen Ed only/private subsidized/ISCED3	0000169	Female	Yes	M	Native	15	M	12.3564	622.76834	605.31518
7	Canada	CAN - stratum 10 : prov 11 Fr. and Eng. 0 to ...	0000474	Female	Yes	M	Native	15	M	1.1471	406.92474	461.0593
8	Canada	CAN - stratum 37 : prov 13 Fr. 118 to ...	0000501	Male	M	M	Native	17	M	2.9962	469.39552	508.2432
9	Germany	Undisclosed STRATUM - Germany	0000064	Female	Yes	M	Second-Generation	10	M	129.849	340.87086	305.33382
10	Australia	AUS - stratum 05 : NSW_Government	0000694	Female	Yes	M	Native	12	M	16.3875	613.2653	671.98806
11	Canada	CAN - stratum 42 : prov 24 Eng. 35 to 305	0000227	Male	Yes	M	Second-Generation	12	M	2.0571	673.1656	615.10632
12	United Arab Emirates	ARE - stratum 21 : SHARJAH,MOE,PUBLIC	0000390	Female	Yes	1.38	Native	16	0.88	4.2319	471.81024	506.00522
13	Finland	FIN - stratum 16 : Swedish/Not Aland-Urban-Low	0000232	Male	Yes	-0.64	Native	12	1.38	2.8857	613.2653	566.71018
14	Chile	CHL - stratum 16 : 16: Private / Primary and Secondary / HS	0000100	Male	Yes	M	Native	16	M	12.794	646.13646	631.14508
15	Bulgaria	BGR - stratum 07 : Region 07	0000131	Male	Yes	-0.64	Native	17.5	0.88	9.8047	396.56486	383.10332
16	Russian Federation	Undisclosed STRATUM - Russian Federation	0000081	Male	Yes	-0.95	Native	13.5	-0.06	207.257	350.1402	388.232
17	Uruguay	URY - stratum 08 : Public Technical Secondary Schools,Mixed	0000145	Male	M	-0.24	Native	12	-0.52	5.1942	339.00138	372.5662
18	Canada	CAN - stratum 22 : prov 12 Eng. 35 to ...	0000131	Female	Yes	-1.46	Native	17	-0.76	5.3938	376.70198	446.23276
19	Finland	FIN - stratum 01 : South-Urban-High	0000288	Male	Yes	M	Native	12	M	5.05	509.7445	517.3816
20	Italy	ITA - REGION 21 stratum 01 : [Region 21] Valle dAosta Licei	0000515	Female	Yes	M	Native	12	M	7.5087	528.12744	569.4144

Para data

There have been mentioned the term para data already but now it is discussed more. This term is not at all meta data since para data is real data in some sense. On the other hand para data needs meta data as well in order to be understood. Para data gives information about survey process, its problems and success. Such data can be like ordinary data as on the next page example about the ESS, but it is often supplementary data, thus separately described. For example:

- Reasons for nonresponse and ineligibility
- Opinion of an interviewers about the quality of an answer of the respondent.
- Data of a survey for interviewers' after the fieldwork
- Number of attempts to contact an interviewee
- Incentives given to interviewee's
- Mode of the survey

Typology of the variables of a ESS file

Identifiers

Name	Type	Width	Decimals	Label	Values	Missing
name	String	36	0	Title of dataset	None	None
essround	Numeric	2	0	ESS round	None	None
edition	String	9	0	Edition	None	None
proddate	String	30	0	Production date	None	None
idno	Numeric	9	0	Respondent's identification number	None	None
cntry	String	6	0	Country	{AL, Albania}...	None

Weights

dweight	Numeric	4	2	Design weight	None	None
pspwght	Numeric	4	2	Post-stratification weight including design weight	None	None
pweight	Numeric	8	2	Population size weight (must be combined with dweig...	None	None

Background variables, domains

Name	Type	Width	Decimals	Label	Values	Missing
cntry	String	6	0	Country	{AL, Albania}...	None
brncntr	Numeric	1	0	Born in country	{1, Yes}...	7, 8, 9
hhmmb	Numeric	2	0	Number of people living regularly as member of household	{77, Refusal}...	77, 88, 99
gndr	Numeric	1	0	Gender	{1, Male}...	9
agea	Numeric	3	0	Age of respondent, calculated	{999, Not available}...	999
marsts	Numeric	2	0	Legal marital status	{1, Legally married}...	66 - 99

Survey outcome variables

Name	Type	Width	Decimals	Label	Values	Missing
stfgov	Numeric	2	0	How satisfied with the national government	{0, Extremely dissatisfied}...	77, 88, 99
happy	Numeric	2	0	How happy are you	{0, Extremely unhappy}...	77, 88, 99
plinsoc	Numeric	2	0	Your place in society	{0, Bottom of our society}...	77, 88, 99
hincfel	Numeric	1	0	Feeling about household's income nowadays	{1, Living comfortably on present income}...	7, 8, 9

Para data variables

icpart3	Numeric	1	0	Interviewer code, lives with husband/wife/partner	{1, Respondent lives with husband/wife/partner}...	9
inwtm	Numeric	4	0	Interview length in minutes, main questionnaire	None	None

Data Summary: We thus have learned many data or variable names:

- Survey plain data
- Outcome, survey or study variable, and their initial or further developed forms
- Three types of auxiliary variables, and their different sources including the fieldwork
- Meta data
- Para data
- Cleaned survey data

I here want to add:

- Survey climate and the variables tried to use in this measurement
- Contextual data that describe the environment in which individuals reside and behave

Summary of the concept section with key symbols

U = target population

D = over-coverage or ineligibles

N = size of the target population (under-coverage may be a problem)

d = number of the ineligibles in a gross sample

r = number of the (unit) respondents = net sample size

n = number of units of the target population in gross sample

n+d = gross sample size

r(y) = number of responses to the variable y

k = statistical unit, e.g. for the respondents *k*=1, …, r

$$response\_rate = \frac{r}{n} \qquad nonresponse\_rate = \frac{n-r}{n}$$

$$ineligibility\_rate = \frac{d}{n+d} \qquad item-nonresponse\_rate = \frac{r-r(y)}{r}$$

Cleaned survey data

Next we go forward to operations that starts for designing data collection strategies and then getting raw data. The raw data need to be cleaned, as well.

Some more is needed:
-You have to document everything somewhere, and most important things into the electronic file, that is,
- You have to label your variables
- You have to label the classifications of your variables
- You have to add para data into the file; this is mainly derived from the fieldwork including e.g.
    - Interviewing time and place, length, interviewer code. mode
    - Reasons for missingness
    - Comments on data quality
- You have to save your file in a good format

And if you are releasing your data set outsiders
- You have to make the data confidential.

Examples about going forward from an initial questionnaire variables

As found, the initial survey data should be cleaned; this will be further considered during the whole course. The ESS public file is an example about such a cleaned data but it does not mean that all data analysis is ready to start even though using the survey weights available (their role will be discussed as well later).

As in my e-book illustrates, new variables especially are needed such those just give answers to your research questions.
I have found the two types of new variables only:

(i) New single variables by transformations about the initial ones
(ii) Combination a new variable from several initial variables. Transformations and other tools can be used in this 'summing.'

Examples about going forward from an initial questionnaire variables

(i) New single variables by transformations about the initial ones: there can be used linear transformation in this case; nothing else happens except that the transformation may facilitate the interpretation, in particular. My second example below is concerned this case but it includes the combining as well since.

(ii) The linearly transformed variables are first scaled into the same interval, and then the average of all are taken and a useful new summarized variable obtained.

Before to this example in detail, some other options are given on next page (all these and more are in my e-book, in its appendix ‘Asteikko- ja muunnosliite.’ Read it. Many things are useful in our training,

Examples about going forward from an initial questionnaire variables

(i) Common single variable transformations in addition to linear:

This transformation is often useful in order to linearize the relationship between the dependent and independent variable

- Categorization with equal or unequal intervals (possible both for categorical and continuous variables)
- Square, Cube, etc and Polynomial using several such variables (ratio-scaled variables or if handled as such)
- Square root (ratio-scaled variables or if handled as such)
- Logarithmic vs. exponential (ratio-scaled variables or if handled as such)
- Logit, probit, log-log and other transformations for binary variables
- Respectively logit, probit etc. for multinomial variables (variables with several ordered categories)

Examples about going forward from an initial questionnaire variables

(ii) New variables from several well scaled initial variables
Yhdistemuuttuja in Finnish

May be two main strategies:

- Using e.g. factor analysis to get an optimal number of dimensions of the initial variables. This example from the ESS is on following pages. Compare this with an example of my e-book “Asteikko- ja muunnosliite” that is from an earlier round.
- Using the average of the linear transformations. This example is found after the first factor analysis example.

Exploratory factor analysis example of the ESS rounds 4 to 6.

Shalom Schwartz has created a theory of basic human values that has been implemented with 21 questions. Their meta data (a shorter version of the question itself) can be seen from the results on the following pages. The scale of the questions is 1 to 6, thus the exact middle point is missing. Note that I excluded the unit with a missing code. Fortunately, their number was not high.

Exploratory factor analysis thus is based on principle components, and the rotation has been used to interpret the factor loadings. I have used four names for these dimensions, Equality, Enjoy, Tradition and Success, but as you see, the variables should be interpreted more broadly. The interpretation usually is mainly based on those variables whose loadings are higher. I have marked these in red.

The first part of the factor pattern. Note that factor1 is most important, factor4 least.

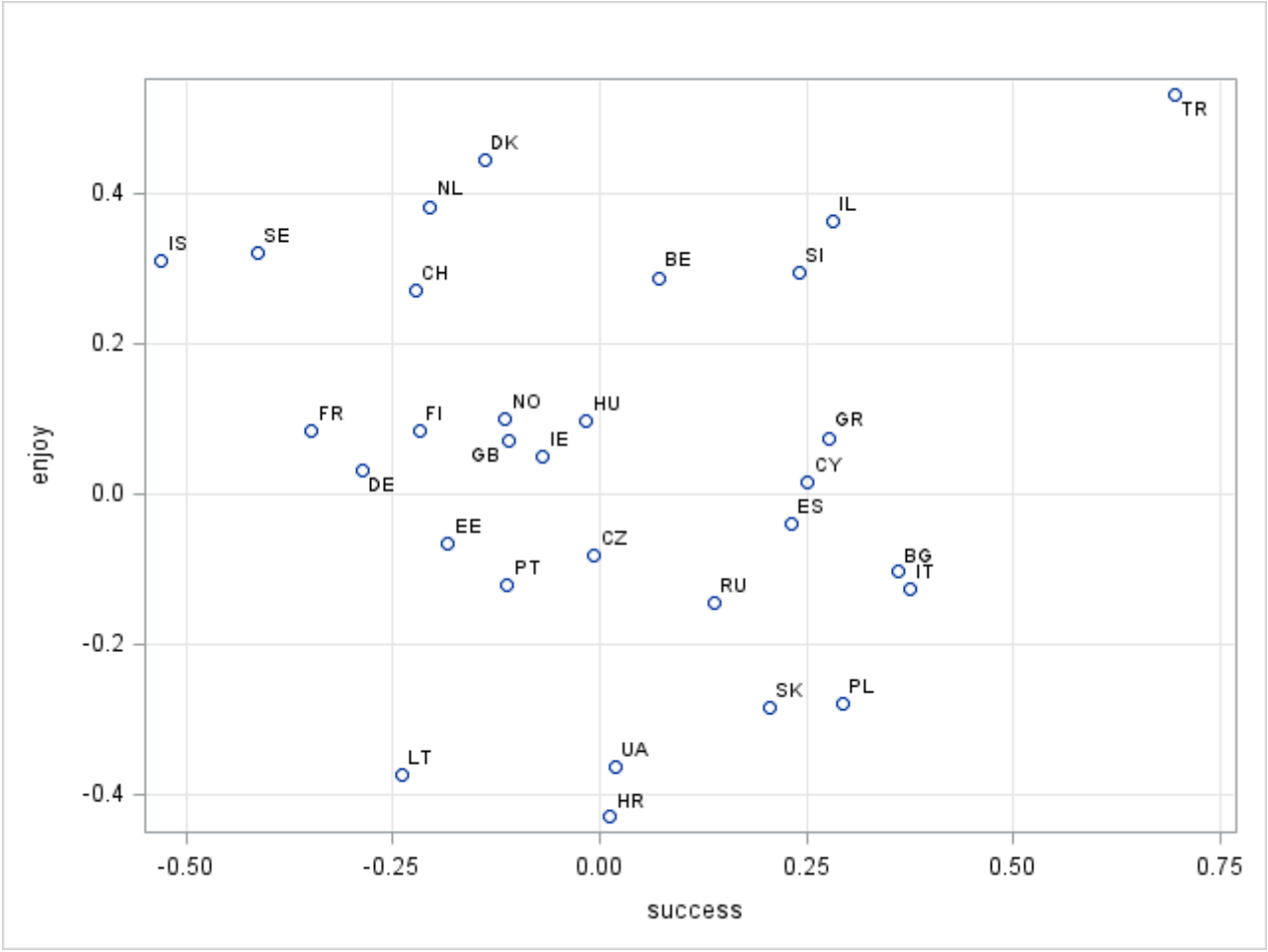
Rotated Factor Pattern					
		Factor1	Factor2	Factor3	Factor4
ipcrtiv	Important to think new ideas and being creative	0.42064	0.40146	-0.17938	0.24787
imprich	Important to be rich, have money and expensive things	-0.18329	0.32636	0.01188	0.64625
ipeqopt	Important that people are treated equally and have equal opportunities	0.66843	-0.01899	0.06328	0.08104
ipshabt	Important to show abilities and be admired	0.15371	0.29410	0.01395	0.68569
impsafe	Important to live in secure and safe surroundings	0.36263	-0.22912	0.37383	0.46182
impdiff	Important to try new and different things in life	0.25064	0.66772	-0.00579	0.16745
ipfrule	Important to do what is told and follow rules	-0.00878	0.04263	0.71049	0.14152
ipudrst	Important to understand different people	0.57664	0.23823	0.28624	-0.13787
ipmodst	Important to be humble and modest, not draw attention	0.27151	-0.02544	0.60747	-0.17834
ipgdtim	Important to have a good time	0.14236	0.67450	-0.02821	0.18612
impfree	Important to make own decisions and be free	0.51102	0.29504	-0.12943	0.25330

The second part of the factor pattern

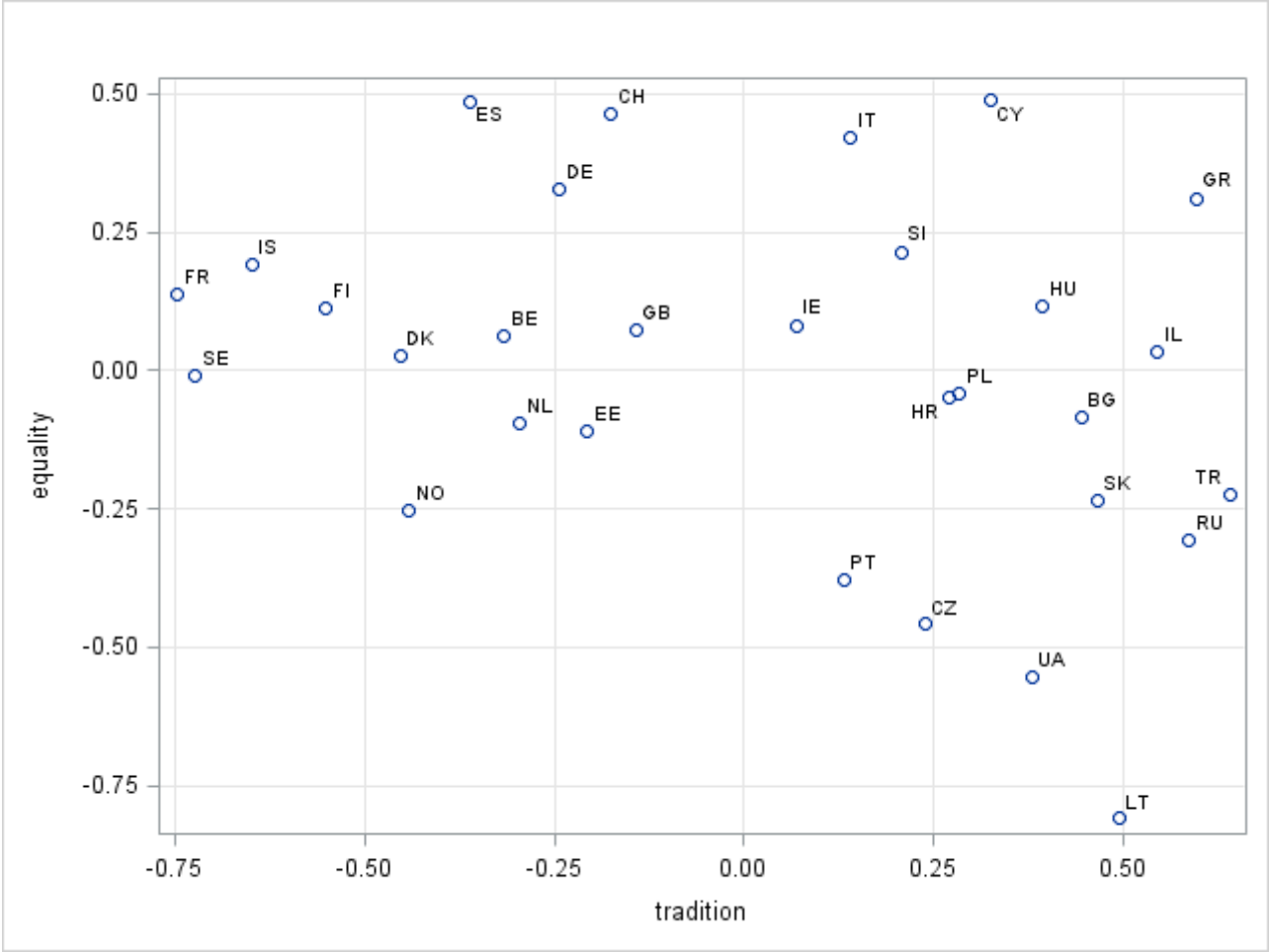
Rotated Factor Pattern (continued)					
		Factor1	Factor2	Factor3	Factor4
impfree	Important to make own decisions and be free	0.51102	0.29504	-0.12943	0.25330
iphlppl	Important to help people and care for others well-being	0.65237	0.15990	0.26566	0.00002
ipsuces	Important to be successful and that people recognize achievements	0.10657	0.38768	0.07134	0.66330
ipstrgv	Important that government is strong and ensures safety	0.39878	-0.15629	0.36066	0.42194
ipadvnt	Important to seek adventures and have an exciting life	-0.05477	0.76388	-0.07443	0.17957
ipbhprp	Important to behave properly	0.21864	-0.06116	0.69663	0.16950
iprspot	Important to get respect from others	0.04174	0.16523	0.33656	0.57308
iplylfr	Important to be loyal to friends and devote to people close	0.64541	0.15753	0.21644	0.03390
impenv	Important to care for nature and environment	0.61385	0.00501	0.24489	0.05327
imptrad	Important to follow traditions and customs	0.20008	-0.05525	0.58895	0.15283
impfun	Important to seek fun and things that give pleasure	0.10195	0.74655	0.04811	0.10559
	Factor name	Equality	Enjoy	Tradition	Success

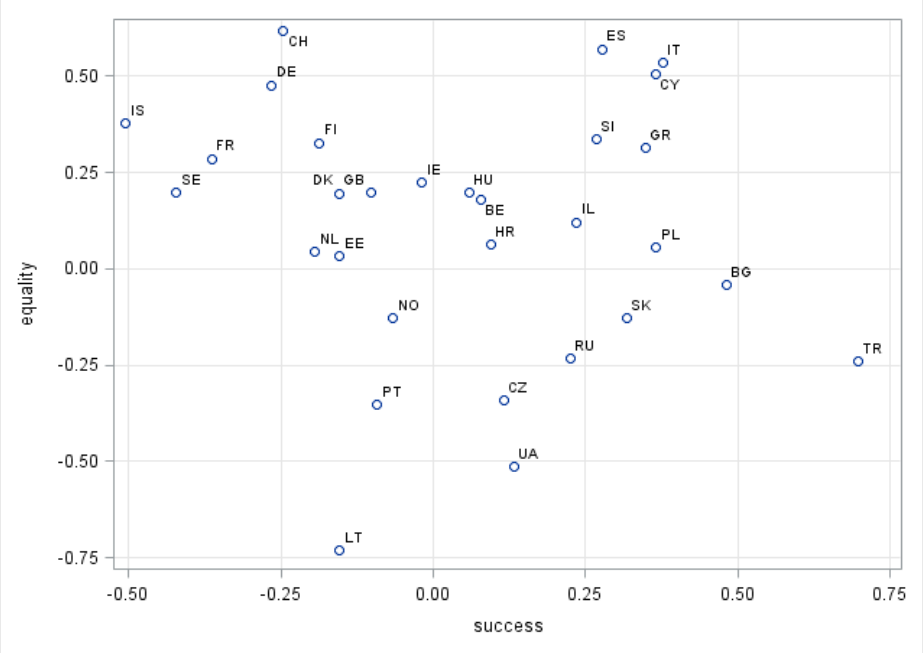
The above result of the exploratory factor analysis is already useful but it is good to go forward.
This is usually made so that the factor scores are computed. These are typically standardized so that their mean = 0 and the standard deviation = 1 respectively. These can be further transformed e.g. so as made in the PISA survey for its literacy exam variables (mean=500, standard deviation =100). The reason is to make the scores easier to understand by ordinary people.
Obs. The variable name should correspond to the values. Hence I have changed the sign of those original scores since the a higher value of the questionnaire is less positive there. Thus be careful when giving the name for each new variable.
All usual statistical methods can be used after scoring. On the next two pages, I have made a scatter plot by country. Your task is to interpret the results. Note that an outlier may be possible due to problems in data as sometimes mentioned.

Scatter plot by the factors ‘enjoy’ and ‘success’ in each ESS country participated in rounds 4 to 6

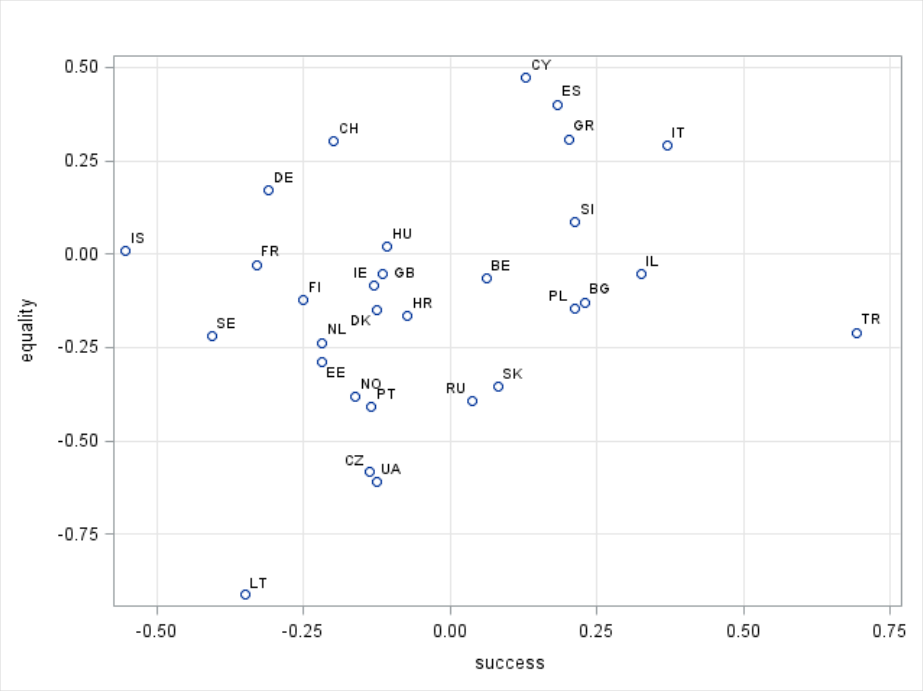


Scatter plot by the factors ‘equality’ and ‘tradition’ in each ESS country participated in rounds 4 to 6





Females



I made one type of scatter by gender too.

Males

This example uses 6 initial variables with two different scales below

Question B 31
How about people from the poorer countries outside Europe?
Instruction(s): Pre: STILL CARD 13
Post: Use the same card **Variable name and label:** IMPCNTR Allow many/few immigrants from poorer countries outside Europe
Values and categories
1 Allow many to come and live here 3
2 Allow some 2
3 Allow a few 1
4 Allow none 0
7 Refusal
8 Don't know
9 No answer

```
The same scale in variables
IMSMETN IMDFETN
```

Question B 32
Would you say it is generally bad or good for [country]'s economy that people come to live here from other countries?
Instruction(s): Pre: CARD 14
Post: Please use this card **Variable name and label:** IMBGECO Immigration bad or good for country's economy
Values and categories
00 Bad for the economy
01 1
02 2
03 3
04 4
05 5
06 6
07 7
08 8
09 9
10 Good for the economy
77 Refusal
88 Don't know
99 No answer

```
The same scale in
variables
IMUECLT IMWBCNT
```

As found there are some answers that cannot be scaled well without imputation. Since we do not know what imputation is and even though we could, it is not necessarily any easy operation. Hence it is often best to exclude these answers from the transformations and the combined variable. This has been made first. The second stage is to create a new variable, `Foreigner_positive`. The purpose here is to get the variable in which least positive to foreign based people gets the value = 0, and most positive = 100, respectively. Note that these are basically ratio-scaled even though formal researchers maybe do not like this. A big advantage is that the results are easy to interpret, since they are like percentages.

```
if IMSMETN>4 then IMSMETN=.;
if IMDFETN>4 then IMDFETN=.;
if IMPCNTR>4 then IMPCNTR=.;
if IMBGECO>10 then IMBGECO=.;
if IMUECLT>10 then IMUECLT=.;
if IMWBCNT>10 then IMWBCNT=.;

Foreigner_positive=mean((4-IMSMETN)*100/3,(4-
IMDFETN)*100/3,(4-IMPCNTR)*100/3, 10*IMBGECO, 10*IMUECLT,
10*IMWBCNT);
```

Note that the operator 'mean' takes each value that is not missing, not needed to answer each question.

Analysis Variable : Foreigner_positive								
Country	N Obs	N	1st Pctl	25th Pctl	Median	Mean	75th Pctl	99th Pctl
BE	1869	1868	3,3	40,6	55,0	52,0	65,0	90,0
BG	2260	2125	0,0	37,3	57,3	54,5	75,0	100,0
CH	1493	1491	16,1	47,8	60,0	58,8	68,3	95,0
CY	1116	1115	0,0	16,7	29,4	31,0	42,2	81,7
CZ	2009	1964	0,0	28,3	41,7	40,7	53,3	95,0
DE	2958	2958	10,7	50,0	62,2	61,9	75,0	100,0
DK	1650	1643	8,9	43,9	58,9	57,5	70,0	96,7
EE	2380	2371	3,3	37,8	50,6	50,7	63,9	96,7
ES	1889	1879	0,0	40,6	57,8	56,0	71,7	100,0
FI	2197	2195	12,2	45,0	56,1	56,5	68,3	96,7
FR	1968	1968	0,0	35,6	52,8	50,0	63,9	96,7
GB	2286	2279	0,0	33,3	48,9	47,8	63,3	94,4
GR	2715	2712	0,0	16,7	31,1	32,3	45,0	85,0
HU	2014	1977	0,0	28,9	42,7	42,4	55,0	95,0
IE	2628	2624	0,0	37,8	54,4	53,4	68,3	100,0
IL	2508	2439	0,0	28,3	45,3	44,5	60,6	100,0
IS	752	749	22,2	57,2	70,0	68,6	83,3	100,0
IT	960	956	0,0	36,7	56,1	52,4	68,3	98,3
LT	2109	2054	0,0	39,4	55,0	53,8	68,3	100,0
NL	1845	1845	8,3	45,0	58,3	56,1	67,8	93,3
NO	1624	1622	11,7	49,4	61,7	61,0	72,2	98,3
PL	1898	1884	6,7	48,3	63,3	61,7	76,1	100,0
PT	2151	2130	0,0	21,7	40,0	40,9	60,0	93,3
RU	2484	2456	0,0	24,7	40,0	40,0	55,0	94,4
SE	1847	1845	16,7	58,3	68,3	68,5	83,9	100,0
SI	1257	1249	0,0	38,3	55,0	52,2	67,2	96,7
SK	1847	1824	0,0	30,0	45,0	44,9	59,4	100,0
UA	2178	2125	0,0	35,6	52,8	51,7	68,3	100,0

Some results by country, the ESS round 6, GR from Round 5.
Make your interpretation.

It is easily possible to calculate other statistical results from this, e.g. by gender, age group, religion etc. This new variable can also be used in models.
Look at next page when sorted by median.

Analysis Variable : Foreigner_positive								
Country	N Obs	N	1st Pctl	25th Pctl	Median	Mean	75th Pctl	99th Pctl
CY	1116	1115	0,0	16,7	29,4	31,0	42,2	81,7
GR	2715	2712	0,0	16,7	31,1	32,3	45,0	85,0
PT	2151	2130	0,0	21,7	40,0	40,9	60,0	93,3
RU	2484	2456	0,0	24,7	40,0	40,0	55,0	94,4
CZ	2009	1964	0,0	28,3	41,7	40,7	53,3	95,0
HU	2014	1977	0,0	28,9	42,7	42,4	55,0	95,0
SK	1847	1824	0,0	30,0	45,0	44,9	59,4	100,0
IL	2508	2439	0,0	28,3	45,3	44,5	60,6	100,0
GB	2286	2279	0,0	33,3	48,9	47,8	63,3	94,4
EE	2380	2371	3,3	37,8	50,6	50,7	63,9	96,7
FR	1968	1968	0,0	35,6	52,8	50,0	63,9	96,7
UA	2178	2125	0,0	35,6	52,8	51,7	68,3	100,0
IE	2628	2624	0,0	37,8	54,4	53,4	68,3	100,0
BE	1869	1868	3,3	40,6	55,0	52,0	65,0	90,0
LT	2109	2054	0,0	39,4	55,0	53,8	68,3	100,0
SI	1257	1249	0,0	38,3	55,0	52,2	67,2	96,7
FI	2197	2195	12,2	45,0	56,1	56,5	68,3	96,7
IT	960	956	0,0	36,7	56,1	52,4	68,3	98,3
BG	2260	2125	0,0	37,3	57,3	54,5	75,0	100,0
ES	1889	1879	0,0	40,6	57,8	56,0	71,7	100,0
NL	1845	1845	8,3	45,0	58,3	56,1	67,8	93,3
DK	1650	1643	8,9	43,9	58,9	57,5	70,0	96,7
CH	1493	1491	16,1	47,8	60,0	58,8	68,3	95,0
NO	1624	1622	11,7	49,4	61,7	61,0	72,2	98,3
DE	2958	2958	10,7	50,0	62,2	61,9	75,0	100,0
PL	1898	1884	6,7	48,3	63,3	61,7	76,1	100,0
SE	1847	1845	16,7	58,3	68,3	68,5	83,9	100,0
IS	752	749	22,2	57,2	70,0	68,6	83,3	100,0

The same result but the countries sorted by the median. In general, if you have results of several domains (like country), this type of sorting is illustrative. I have seen too often that everything has been published in the initial order although it makes difficult to interpret the results. It would be possible here to add the standard errors but it is not done. It is fairly easy for the mean but not for all other parameters. Hence I have computed these by SAS; the same can be made by SPSS or another good software package. Next page.

Domain Analysis: Country						
Country	Variable	N	Mean	Std Error of Mean	95% CL for Mean	
CY	Foreigner_positive	1115	31,0	0,6	29,8	32,3
GR	Foreigner_positive	2712	32,3	0,4	31,5	33,2
RU	Foreigner_positive	2456	40,0	0,5	39,0	41,0
CZ	Foreigner_positive	1964	40,7	0,5	39,6	41,8
PT	Foreigner_positive	2130	40,9	0,6	39,7	42,0
HU	Foreigner_positive	1977	42,4	0,5	41,5	43,3
IL	Foreigner_positive	2439	44,5	0,5	43,6	45,5
SK	Foreigner_positive	1824	44,9	0,7	43,6	46,3
GB	Foreigner_positive	2279	47,8	0,5	46,7	48,8
FR	Foreigner_positive	1968	50,0	0,6	48,9	51,1
EE	Foreigner_positive	2371	50,7	0,4	49,9	51,5
UA	Foreigner_positive	2125	51,7	0,6	50,5	52,8
BE	Foreigner_positive	1868	52,0	0,5	51,1	52,9
SI	Foreigner_positive	1249	52,2	0,6	50,9	53,4
IT	Foreigner_positive	956	52,4	0,9	50,6	54,1
IE	Foreigner_positive	2624	53,4	0,5	52,4	54,4
LT	Foreigner_positive	2054	53,8	0,6	52,6	54,9
BG	Foreigner_positive	2125	54,5	0,6	53,2	55,8
ES	Foreigner_positive	1879	56,0	0,6	54,9	57,2
NL	Foreigner_positive	1845	56,1	0,5	55,2	57,0
FI	Foreigner_positive	2195	56,5	0,4	55,7	57,3
DK	Foreigner_positive	1643	57,5	0,5	56,4	58,5
CH	Foreigner_positive	1491	58,8	0,5	57,9	59,7
NO	Foreigner_positive	1622	61,0	0,5	60,1	62,0
PL	Foreigner_positive	1884	61,7	0,5	60,7	62,6
DE	Foreigner_positive	2958	61,9	0,4	61,1	62,7
SE	Foreigner_positive	1845	68,5	0,5	67,6	69,4
IS	Foreigner_positive	749	68,6	0,7	67,3	69,9

Here the countries are sorted by the mean. You can compare these results with the previous one.
I still continue to a graphical presentation that may be even more illustrative. Next page.

Now you can easily see whether a difference between nearest countries is significant at 95 % Confidence Interval (CI) level. If you wish to do the same at another level, you have to make your own calculations using the standard errors. E.g. Iceland (IS) and Sweden (SE) do not differ from each other but from all others.
Obs. The scaling and the graphics format here can be changed. What else you could apply?