Topics in Social Statistics
University of Helsinki
Part I

Seppo Laaksonen

Tuesday 13 September
Thursday 15 September

Content

What is survey?
Key concepts in surveys
From Survey Data Collection to Cleaned Survey Data

Through
- Designing the survey
- Designing the questionnaire
- Designing the sample(s)
- Data collection with alternative single and mixed modes
- Data entry
- Editing the raw data
- Imputing the data
- Weighting the data
- Adding other features into the data file

Part 3 extends some to Part 1, especially in analysing cleaned data. Part 2 is focused on measurement questions in data collection and analysis.

What is survey?

It is a demanding question to answer.

Wikipedia:

Statistical surveys are used to collect quantitative information about items in a population. Surveys of human populations and institutions are common in political polling and government, health, social science and marketing research. A survey may focus on opinions or factual information depending on its purpose, and many surveys involve administering questions to individuals. When the questions are administered by a researcher, the survey is called a structured interview or a researcher-administered survey. When the questions are administered by the respondent, the survey is referred to as a questionnaire or a self-administered survey.

What is survey?

It is a demanding question to answer.

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
- Families
- Enterprises
- Plants (Local units of enterprises)
- Local-kind-of-activity units of enterprises
- Villages
- Municipalities
- Other areas
- Societies

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

My recent definition

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, households' income and poverty, and their housing environments, or the characteristics and performance of businesses. Survey research is avoidably multidisciplinary 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. On the other hand, surveys exploit substantially many other sciences e.g. such as informatics, mathematics, cognitive psychology and theory of sub-matter sciences of each survey topic.

Comments?

Key concepts in surveys

Next pages concentrate on
- Populations in surveys
- Cross-sectional vs. longitudinal surveys
- Sampling design
- Missingnesses and other deficiencies
- Editing
- Imputation
- Sampling and other weights
- Meta data and para data

Populations in surveys 1

In statistics population is a key concept 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 to be determined including its period (a point of time or a time period).

The target population of the ESS e.g. "Persons 15 years or older who are resident 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. Give your own example.

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 Finnish 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 thus two frames, but it is possible that this number can be even four.

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.

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 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, …).

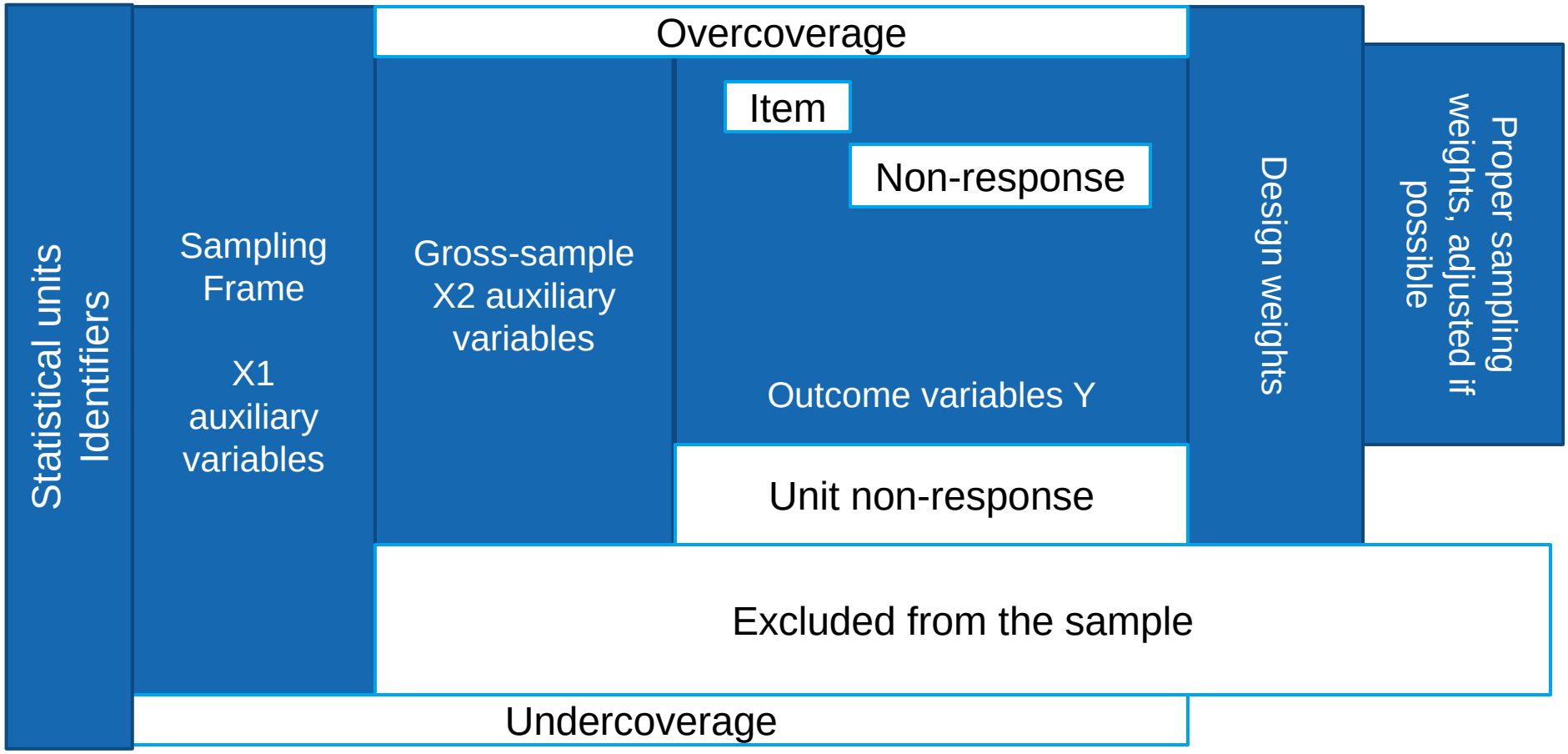
PISA and other student surveys use typically two units:
- 'PISA' Schools (or school classes)
and
- Students themselves

Give other examples.

The next two pages illustrate missingnesses as well as some other crucial concepts in surveys.

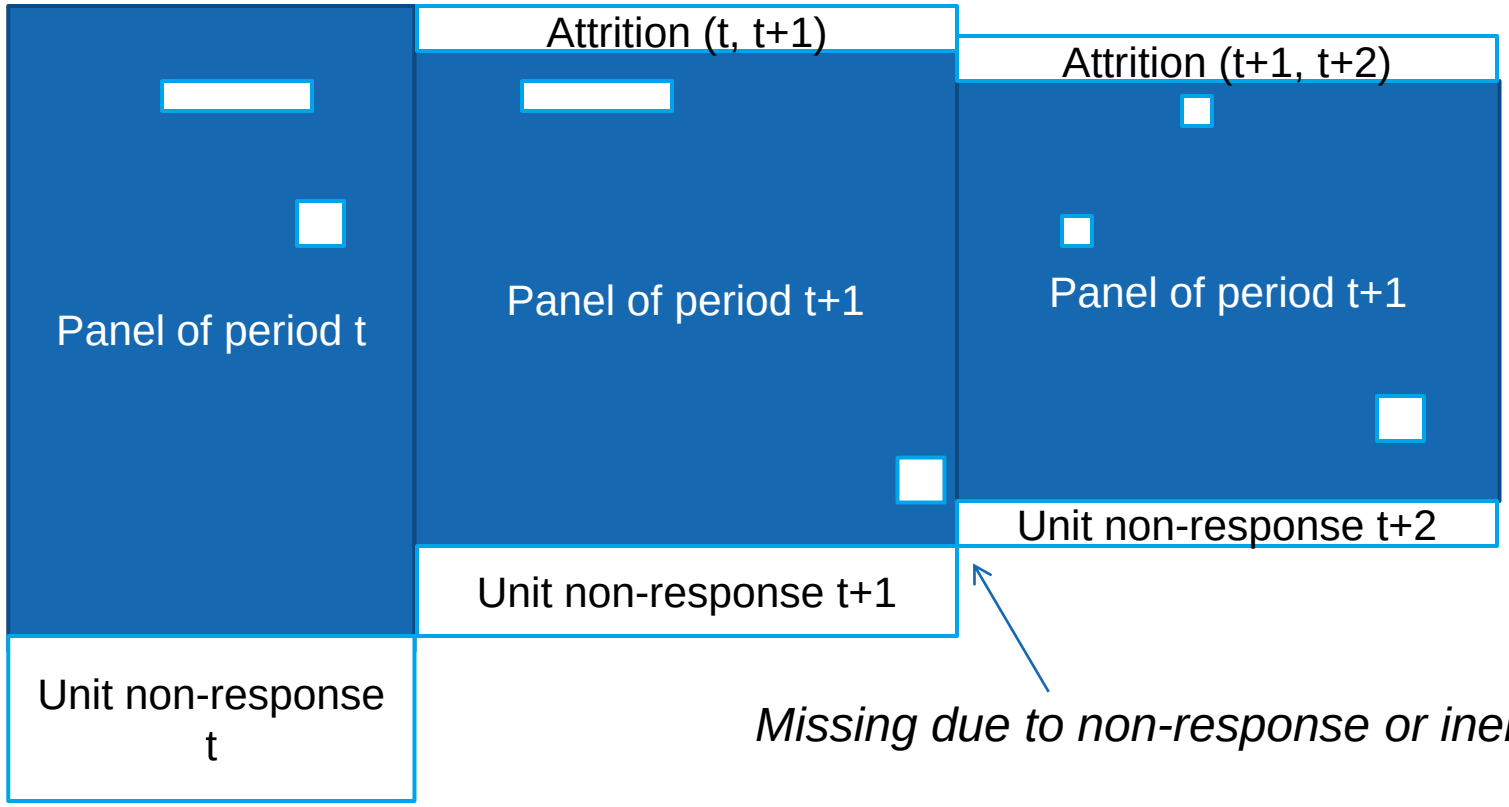
Micro data and Missingness

Now I focus on micro data where one can see various types of missingness. This is a cross-sectional case (white boxes are without data, missing for various reasons):



Micro data and Missingness

Cohort type of panel (longitudinal) example



Questionnaire and questioning

Basically four questions is required to be answered:

(i) How to contact/approach to a potential respondent?
Alternatives
- Mail
- Direct contact at home or another place
- Phoning
- Email
- General invitation in media, web, social media, poster
- Automatic invitation in the website to those who are present there
- Respondent has been contacted at street, shop, sport event, cultural event
- Respondent has been contacted from outside (driving, walking etc at certain area/point); she/he does not necessarily know that she/he has been picked up at survey data. This may lead to confidentiality problems.

Questionnaire and questioning 2

(ii) How the information uploaded into the file?
Alternatives:
- Interviewer asks, and saves the answers into the paper or another manual file
- Interviewer asks, and saves the answers into the electronic file.
- Respondent reads, looks and/or listens the questions and saves the answers into a file
- Respondent reads, looks and/or listens the questions and the interviewer saves the answers into a file
- IT system submits the questions to a respondent and she/he answers by email or by web questionnaire
- IT system collects the data automatically from the data base of the respondent (should be accepted by the respondent); this is typical for business surveys.

Questionnaire and questioning 3

(iii) What kind of formats the questionnaires use? Note that the format can be converted into a new format after initial data collection.
Alternatives:
- Paper that can be filled in manually or printed from an electronic file.
- Electronic local format such as a memory stick
- Text message, email, annex of the email,
- Specific driver on the web, open or closed

(iv) How to submit the data?
Alternatives:
- If the data are already uploaded into a electronic file, it is ready.
- The paper responses can be submitted by mail or after scanning by email etc.
- Electronic files can also be submitted forward by mail or email or saving on an appropriate location.

Questionnaire and questioning 4

Some terms

PAPI = Paper and Pencil Interview
CAPI = Computer Assisted Personal (phone or f2f or Skype) Interview
Face to Face Interview (*f2f*)
CASI = Computer Assisted Self Interview
CATI = Computer Assisted Telephone Interview (this can be conducted in a specific so-called CATI centre or used by an individual interviewer)
TSI = Telephone Self Interview
Postal Survey
Web/Internet Survey
Telephone self interview
CAI = Computer Assisted Interview

Mixed mode survey: e.g. first web, next phone, or first web, next f2f, or first mail, next web, next phone, next f2f

Questionnaire and questioning 5

The questionnaires themselves are very important and difficult to well prepare. I do not go here to details but we will discuss this issue.

For this purpose, download the ESS4 or ESS5 questionnaire on their website Eusropeansocialsurvey.org and look how different styles are used in different questions. Specifically, choose some interesting questions for our demonstrations on Thursday.

These questionnaires are for f2f surveys, that is, they are not maybe similarly available for phone or web. We can discuss also this issue.

Sampling design

You saw that some missingness is due to sampling. If the population is big, sampling is a natural tool to work on.
I present here a compact framework sampling. This is called sampling design. Often a narrower framework has been given.

My framework is for *probability sampling*, not for quota or other non-probability sampling. Voluntary sampling is nowadays becoming more common especially using web arsenals. These are also non-probability methods for sampling.

Also so-called access panels are created. These are attempted to make as probability based as possible but voluntariness means that they are completely such ones.

First some basic concepts:

Sampling design 2

Cluster = e.g.

- small area where residents, birds, students
- school where students
- household where its members
- address where residents, employees
- enterprise where employees

Inclusion probability: probability that a frame (target population) unit will be included in the (gross) sample. In probability sampling this probability must be >0 (=1 is accepted naturally). Otherwise some units cannot be drawn in the sample.

Primary sampling unit = *psu*: the unit that has been included in the sample in the first step in sampling, using probability principles. Next secondary sampling unit.

Stratum: group or sub-population that will be included definitely in the sample. Inclusion probability = 100%. The strata are independent of each other, like sub-populations.

Sampling design 3

A. Frame	Study units are explicitly in the frame or they are not there.
B. Sampling unit	The sampling unit is the study unit as well, or not.
C. Stage	Hierarchy to approach to the study units by using probability sampling. First going to the first-stage units (=*psu*'s), and then to the second stage units, …Terms: one-stage sampling, two-stage sampling, three-stage sampling

D. Phase	First a probability sampling applied for drawing a first-phase sample, and afterwards a new sample has been drawn at the second phase from the first sample.
E. Stratification	The population divided into several independent sub-populations.
F. Allocation of the sample	How a desired gross sample has been shared into each stratum.

Sampling design 4

G. Panel vs. cross-sectional study	If a panel is desired, it is needed to design also how to follow up the first sample units, and how to maintain the sample. Whereas a cross-sectional study is desired, it is good to design it so that a possible repeated survey can be conducted (thus getting a correct time series).
H. Selection method	How to select the study units - probability equal to all (srs, equi-distance, Bernoulli) or - probability varies unequally typically by size (pps =probability proportional to size)
I. Missingness anticipation	Trying to anticipate response rates and allocate a gross sample so that the net sample is as optimal as possible in order to get as accurate results as possible.

Thus: choose an optimal alternative from each A to I alternatives, and you will have a gross sample. Of course this task is not easy, since you have to anticipate many things. One such thing is intra-class correlation *rho* given that your *psu* is a cluster.

Sampling design 5 _ Cluster psu

The correlation $rho = \frac{Between\_\text{var}iation}{Total\_\text{var}iation}$ is an indicator for homogeneity of clusters. It varies a lot from a survey to the next.
For example in PISA 2006, where the clusters are school classes and the variables are scores of mathematical-statistical literacy, it is for Finland around 0.1 but for Germany around 0.6. What this means?
The ESS *rho*'s are much smaller, since *psu*'s are small areas that are not as homogenous. Typically *rho* is around 0.02-0.04. In Finland it is = 0 since clusters are not used.

So, when designing ESS samples we have to anticipate many things, also response rates. Unequal probabilities mean that the accuracy will worsen. Hence we also increase the gross sample size (analogously to cluster effect in which case a higher *rho* requires a larger gross sample. This is due to our target that all participating countries achieve an about same accuracy level. This has been measured at sampling design with *effective sample size* that should be 1500 at minimum.

Sampling design 6 _ DEFF

We use the concept *DEFF (design effect)* when planning the sample of the ESS and its gross sample size, in particular. This indicator is the ratio between the anticipated accuracy of this particular design and the corresponding design based on *srs* (although not used in most cases).

We have two *DEFF*'s:

-Due to unequal inclusion probabilities *DEFFp*
- Due to clustering *DEFFc = 1 + (b-1)rho (b=*average *net* cluster size*)*

The whole *DEFF* is the product of both *DEFF*'s

I give a theoretical example based on this whole strategy.

Sampling design 7 _ Summary

Operation	**Example calculation** (average-based, the figures may vary by stratum, cluster and another domain)
1. Target for the effective sample size (*neff)*	**2000**
2. Anticipated missingness due to unit nonresponse	30% i.e. 2000/.7 = 2857
3. Anticipated missingness due to in-eligibility	5% eli 2857/.95 = 3008
4. Anticipated Design Effect (DEFF) due to clustering including anticipated intra-class correlation (=0.025), average net cluster size (=5.3) and missingness (average gross cluster size = 8)	$DEFF_c$ = 1+(5.3-1)*.025 = 1.11 3008*1.11= 3338
5. Anticipated DEFF due to varying inclusion probabilities (calculated for anticipated respondents if possible)	$DEFF_p$ = 1.25 3338*1.25 = 4173
6. Risk factor, leading to increase the above Gross Sample Size Anticipated Net Sample Size	**4250** 2826

Crucial point:

All selections should be probability based but most important is to organize the last step (stage) really randomly. Usually, this stage is based on *srs* like drawing study units within *small-area psu's*. In practice, especially in f2f surveys, this basically simple operation is not simple, since random selection may be demanding for practical reasons. E.g. random route has been used, the listing of study units within psu's may be expensive and not really done correctly. But this step is extremely important.

Look at the sampling guidelines and other methodological documents of the European Social Survey **(ESS)**

http://www.europeansocialsurvey.org/index.php?option=com_content&view=article&id=80&Itemid=365

Missingness mechanisms 1

In order to handle missing data due to sampling and non-response auxiliary information is very important. Without such data can be something done but not well. On the other hand, it is important to assess the missingness (or response) mechanism.

There are four basic mechanisms good to think and make assumptions before starting the missing data handling (usually only three of these are presented in literature):

MCAR (Missing Completely At Random): If this could be reality, it is rather easy to decide which methods to apply. Most methods are workable and you do not need auxiliary variables either.

MARS (Missing At Random Under Sampling Design): Now missingness only depends on the sampling design variables. This is often used so that one assume that MCAR holds true within strata (pre-strata, or even post-strata). Here imputation or weighting is performed by strata.

Missingness mechanisms 2

MAR (Missing At Random Conditionally): Now missingness depends on both the sampling design variables and all possible other auxiliary variables. This assumption is much used when good auxliary variables are available.

MNAR (Missing Not At Random): Unfortunately this is the most common case in real-life to some extent. So, when all the auxiliary variables have been exploited much help has been received but still it is rather clear that our results are not ideal. So, it is good to interpret possible biases in results against general knowledge and missingness of good auxiliaries.

Data editing or statistical editing

This is often made together with imputation since editing can lead to imputation e.g. to replace missing or strange values with imputed ones.

Editing is basically a phase that can require a lot of time if entry data are dirty. This is the case more likely if

- Postal survey has been used i.e. a respondent has filled in his/her questionnaire.
- Pre-editing has not been tried in data entry (this can be the case in all modes but usually it is easy to give the limits for acceptable values in web, face-to-face and telephone surveys). Textual answers still need further editing usually.

However, a good survey provider always include a simple editing at minimum in data entry. Hence the initial 'raw' data file is slightly edited and it will easier to continue toward next editing steps.

Data editing or statistical editing 2

Nevertheless, editing is not only for correcting typing or other errors. It is also a development process:

- To learn about the whole survey process and

in particular

- Errors and not best practices of this current survey.

These learnings have been used in survey documentation since users/clients desire to know the survey quality from its all aspects. Secondly, the experience from each survey is beneficial to use in future surveys even though you have not in mind new surveys. It is important to save all editing (as software program).

Data editing or statistical editing 3

Some concepts from the UNECE data editing glossary (search more as yourself):

CHECKING RULE

A logical condition or a restriction to the value of a data item or a data group which must be met if the data is to be considered correct. In various connections other terms are used, e.g. edit rule.

CONSISTENCY CHECK

Detecting whether the value of two or more data items are not in contradiction.

ERROR LOCALIZATION

The (automatic) identification of the fields to impute in an edit-failing record. In most cases, an optimization algorithm is used to determine the minimal set of fields to impute so that the final (corrected) record will not fail edits.

GRAPHICAL EDITING

Using graphs to identify anomalies in data. While such graphical methods can employ paper, the more sophisticated use powerful interactive methods that interconnect groups of graphs automatically and retrieve detailed records for manual review and editing.

Data editing or statistical editing 4

I do not go to details in editing since it is specific for each survey, that is, it is required to check all values both

- One-dimensionally (like checking whether all values are acceptable)
- Two-dimensionally as well as possible (such as whether the values is acceptable also by background variables such as gender, age, region, education. E.g. A male cannot be a mother, a very young human cannot have a child or cannot be married).
- Distributionally
- Multi-dimensionally that can be made also using a multivariate model.
- In graphs that often facilitates to see the previous options more concretely.

All errors or inconsistencies cannot be correctly definitely but it is not always needed, fortunately, unless their effect is fatal in estimates.

Just to do your best, and improve the data even during data analysis.

What is imputation?

It is to insert a value into the data in a more or less fabricated way. **Why**?

- Since there is no value in this cell, that is, it is completely missing.
- Since the existing value is partially missing (like given as an interval) but it is desired replace this with a unique value.
- Since the existing value does not seem to be correct, and consequently, it is desired to get a more reliable value to replace this.
- Since the current value seems to be too confidential, that is, and this individual unit should be disclosed. Motivation: the fabricated value can be considered as less confidential.

Imputation can be performed both for the macro and micro data but during this course I only consider the imputation methods of **micro** data. However, basically the same methods can be applied to macro data but usually this imputation is more limited.

Purpose of imputation

The purpose of imputation is twofold

-**Either** to replace a missing or partially missing or incorrect value with a such value that the estimate behind this variable will be more valuable than without imputation. Thus if imputation is advantageous from the estimation points of view, use it. Naturally, there are in surveys several estimation tasks and can be possible that a certain imputation is not advantageous in all respects. Hence, it is possible that some estimates are computed without imputation and some others with imputation. On the other hand, a big question is which imputation is best for each estimation. It is good to notice also that a bad imputation may worsen the estimation. Be careful!

- **Or** to make data more confidential. This leads to create certain incorrect values into the data that is not difficult but this should not be a purpose but to impute the confidential values so that their pattern gives opportunity to get as the reliable estimates as possible.

What can be imputed due to missingness

When looking for those schemes, we can find the following possible imputation affairs:

(i) Undercoverage that requires a new up-to-date frame. Very seldom possible.
(ii) Those units that are not selected into the sample. Done in threoretical (simulation) studies
(iii) Unit non-response, all or some variables. If done, called mass imputation. This is competitive to weighting methods.
(iv) Item non-response. This is the most common case.
(v) Deficient and sensitive values. Quite common.
(vi) Second, third etc wave missing values in cohort studies given that the previous value exists (or imputed correctly enough).

Most common tools for missing item handling without real imputation

(i) In the case of mass missingness, the weighting or the reweighting is is mostly exploited. This is possible only for the respondents. The respective imputed data thus covers the non-respondents too (or those non-respondents desired to include). Note that one imputation strategy is a kind of weighting method but its weights are more flexible than the standard reweighted sampling weights.

Most common tools for missing item handling without real imputation 2

(ii) Item-nonresponse is marked with a good and well-covered code such as:

- -1 = respondent candidate not contacted
- -2 = respondent refused to answer
- -3 = respondent was not able to give a correct answer
- -4 = missing for other reasons
- -6 = question was not asked from the respondent
- -9 = question does not concern the respondent

These codes are not much used but such as 7, 8, 9, 66, 77, 88, 99 instead. The negative values are easy to observe. Do not use a zero (0)!

Most common tools for missing item handling without real imputation 3

(ii) cont.

The good and illustrative codes are useful also when deciding the imputation methods itself. Thus a different method may be chosen for the different type of missingness.

Moreover, it is good to notice that the coded variable is full, without missing values. This kind of a categorical variable can be used as an explanatory variable in standard linear and linearised models, among others. But if desired to use it as continuous, real imputation is required.

Most common tools for missing item handling without real imputation 4

(iii) The values with missing codes are excluded from each analysis so that the observation number varies. This strategy does not give consistent results with each other.

(iv) Close to case (iii) but now the units with missing values have been excluded from each analysis. In this latter case, there are always the same number of observations. The standard multi-dimensional analysis makes this automatically for those variable patterns that are used in the multidimensional analysis. This strategy gives consistent results with each other.

(v) Pairwise analysis for multivariate purposes in such cases where e.g. the correlations are the basis for further analysis. This operation first computes pairwaise correlations like in case (iii) and when continues from the correlation matrix towards multivariate analysis.

Targets for imputation should be specified clearly

It is rather clear (except when imputation aims at protecting data)

(i) That a user is happy if the imputed values are as close as possible to the correct values. Success at individual level. Another point is that how to know how close they are, except in some cases. This may be often the too demaning target and hence

(ii) A user is still fairly happy if the distribution of the imputed values is close to the distribution obtained from true values. Success at distributional level. Of course this is hard to check but however easier than case (i).

(iii) The target to succeed at aggregate level is also statisfactory and specificly in NSI's where such estimates as average, total, ratio, median, decile and standard deviation are typical.

(iv) Some users hope to get the order of imputed values as correct as possible.

(v) Finally, success to preserve associations (like correlations) is also important in many studies.

Imputation process

Imputation is part of the data cleaning process. It can be considered to cover the following 6 **actions**:

(i) Basic data editing in which part the values desired to impute are also determined.

(ii) Auxiliary data acquisition and service incl. preliminary ideas to exploite these.

(iii) Imputation model(s): specification, estimation, outputs

(iv) Imputation task(s): use outputs of the model for imputation, possible re-editing if the imputed data are not clean and consistent.

(v) Estimation: point-estimates, variance estimation = sampling variance plus imputation variance.

(vi) Creation of the completed data (or several data): includes good meta data such as **flagging** of imputed values, documenting of the whole imputation procedure and deciding what to give outsiders.

Next I focus on the actions (iii) and (iv).

Single and multiple imputation

Imputation can be performed for each desired value of the non-complete variable just once, or several times. The first is called *single imputation (SI)* and the second *multiple imputation (MI).* These are not the two different imputation methods as often said, since multiple imputation means that single imputation has been repeated several times. So, each single imputation should aim at succeeding as well as possible e.g. avoiding the bias. There are the strict rules how to repeat imputation properly. In this presentation these have not been much discussed. The rules are not always clear and hence often criticized.

Imputation model

Imputation model should be integrated strictly to the next step, that is, to imputation task. There are two options to determine the specification of the imputation model:

• To determine the model using smart information so that it predicts well the case required to impute. The model may a deterministic (or stochastic function) like *y = f(x)* (*+ e*) or a rule (like in editing) such as 'if so and so but not so then it is that.'

• To estimate the model using either the same data required to impute or other data that is similar (at least the structure) to the present data.

The previous models are often used in simple (conservative) imputations and in the same step as editing. Next I will focus on the latter models.

Imputation model 2

This second type of imputation model is always such in which it is purpose to predict something using auxiliary variables as independent variables.

The dependent variable of this imputation model can be of the two types only:

(i) either the variable being imputed itself

or

(ii) the missingness indicator of this variable.

Case (ii) can cover all possible forms, categorical including binary and continuous but in case (ii) the variable is binary.

Imputation model 3

These two models are estimated from the two different data sets:

(i) From the respondents (observed units)

(ii) Both from the respondents and the non-respondents.

But of course, the explanatory variables should be available from both the respondents and the non-respondents. Note my earlier comment that a categorical variable with the missingness codes may work reasonably in the imputation but many such variables maybe not unless these are concerned different units.

Imputation model 4

The model (ii) is concerned a binary variable (1 = responded, 0 = not) but the same model can used for the model (i) if the dependent variable is binary (e.g. 1 = employed, 0 = unemployed).

You know how to work with the binary model to predict. First you have to choose a link function, that can be:

-logit
-probit
-complementary log-log
-log-log .

There are no dramatic differences in explaining models between those link functions but of course some. Imputation thus requires to use this model for predicting the response propensities for all units (respondents and non-respondents). That is, the first outputs are those values between (0, 1).

Imputation model 5

In addition to ordinary models such as linear regression or probit regression, the imputation model can be nonlinear and nonparametric. An interesting example of the latter ones is *tree modelling*. If the dependent variable is categorical, we speak about *classification tree*s, whereas the model for continuous variable is *regression tree*. Moreover, neural nets often create analogous groups of the gross sample. This kind of a group is called in imputation terminology as *imputation class* or imputation cell.

Imputation cells can also be constructed manually or using smart statistical thinking. For example, strata can be rather good imputation cells. Given that the imputation cells are homogenous from the imputational points of view (especially if MCAR holds true within cells), these offer many advantages.

Imputation task

The two alternatives in general can be exploited after you have estimated the imputation model:

(a) Model-donor approach in which case the imputed values are computed deterministically (or stochastically) from the predicted values (adding noise) of the model.
(b) Real-donor approach in which case the predicted values (or adding noise) are used to find the nearest or a near neighbor of a unit with a missing value from whom an imputed value has been borrowed.

You see that the imputed values of case (b) are always observed values, observed at least once for respondents. The imputed values of case (a) are not necessarily observed except often for categorical variables.

Imputation task 2

To integrate model and task you see that we have the following options. So, the predicted values of the missingness indicator cannot be used for model-donor imputation directly.

	(a) Model-donor approach	(b) Real-donor approach
(i) either the variable being imputed itself	Yes	Yes
(ii) the missingness indicator of this variable	No	Yes

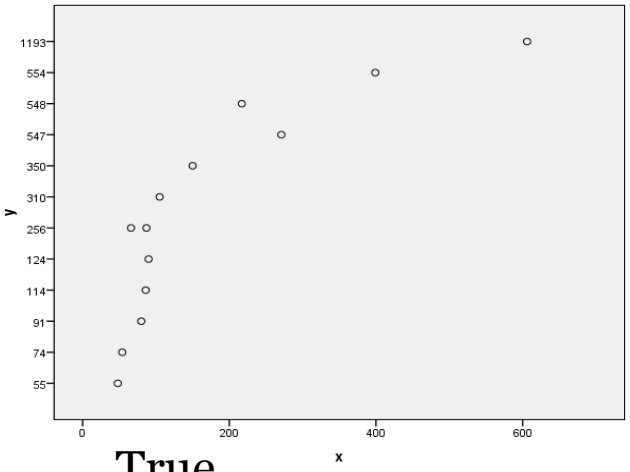
Example with very simple data

You have wondered that any commonly known imputation methods like mean imputation has not been mentioned in the text. This is due to my framework that covers of course those simple and usually inappropriate methods. The following example illustrates my framework in which the imputation model and imputation task is good to recognize even though the model is simple.

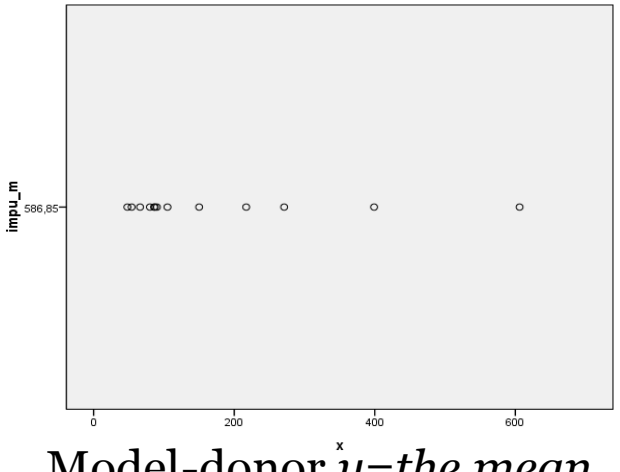
Example with very simple data 2

The data are artificial. Variable *y* is that required to impute to some extent. I have only one auxiliary variable *x*. These two variables are well correlated, *r*=0.92. The number of the units is 40, that of the non-respondents is 13. Missingess is not random, it is higher for small and large *y* values. Possibilities for successful imputation exist. My first imputation model is *y=the mean* but in the other four I tried the model *y=x,* also adding a normally distributed noise term. Results are below and on next page.

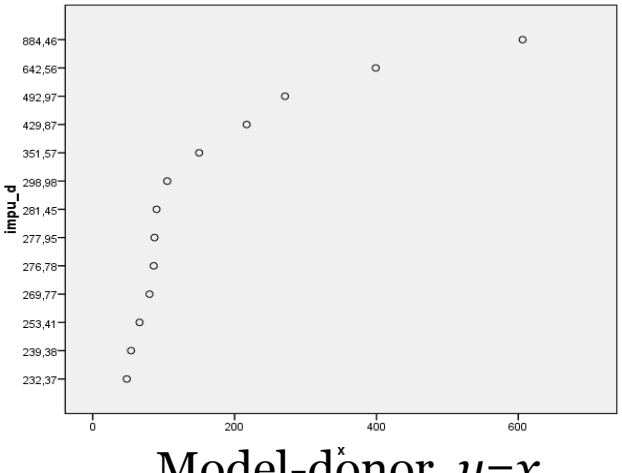
		Observations	Mean	Std deviation
True		40	507	317
Respondents		27	587	292
Model-donor				
Model	y= the mean	40	587	238
Model	y=x	40	519	279
Model	y=x+e	40	516	295
Real-donor				
Model	y=x	40	499	299
Model	y=x+e	40	534	299



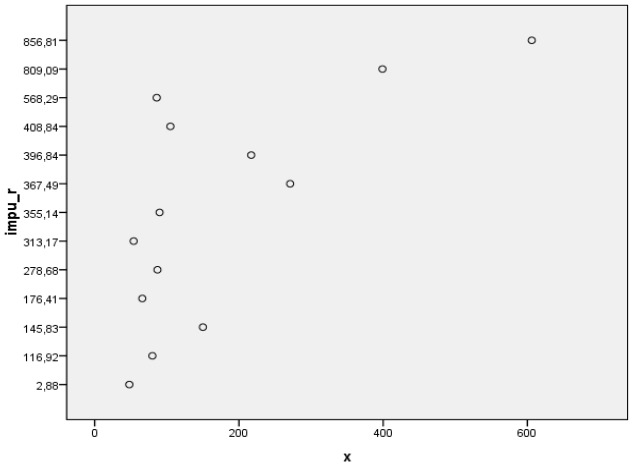
True



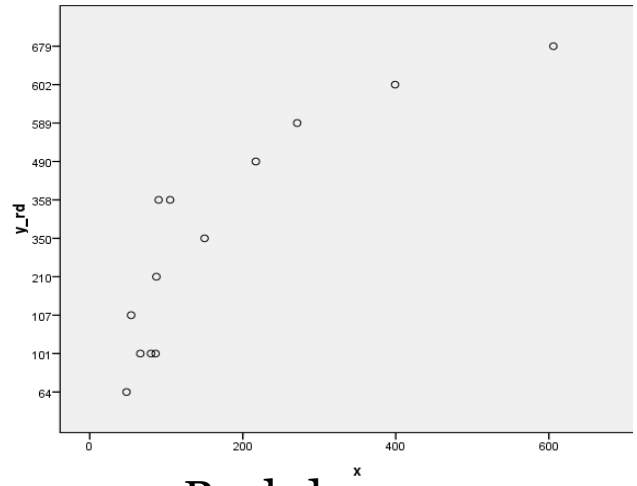
Model-donor *y=the mean*



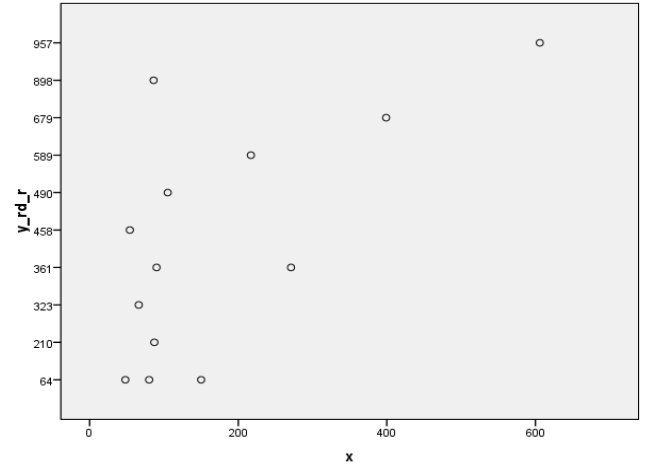
Model-donor *y=x*



Model-donor *y=x+e*



Real-donor *y=x*



Real-donor *y=x+e*

The scatters are for the imputed values. Thus compare with the true scatter.

Weighting and reweighting 1

You have seen that sample data need to be weighted, these are in generally called *sampling (sample) weights*. First the design weights have been calculated exactly following the rules of sampling, i.e. inclusion probabilities for gross sample.

Design weight = the inverse of inclusion probability of the study unit. This can be quite complex if the design is complex. The sum of the design weights = the size of the target population as such as it is known.

Due to unit missingness, new weights are needed. I call these here basic (base) weights. They are computed for the respondents only and their sum = the sum of the design weights. Basically, these weights are like the design weights since they are a simple reformulation of the design weights assuming that the missingness mechanism is MCAR in each stratum.

Weighting and reweighting 2

Base weights are not usually best ones. Hence it has been attempted to get better weights, called adjusted weights. These can be successfully created only if good auxiliary variables are available. These variables should be good predictors for missingness.

Several methods have been developed for adjusting the sampling weights:

- *Post-stratificatio*n (in which case new strata have been constructed within initial strata = pre-strata)
- *Homogeneous groups* (these are like post-strata but based on gross sample data whereas post-strata are created from population level).
- *Calibration methods* (these take benefit of known population level data). Post-stratification is the simplest (but often good) calibration method. More advanced calibration methods can use population margins of several auxiliary variables so that the estimates of these margins will be as correct from the sample data as those margins are.

Weighting and reweighting 3

My favourite is post-stratification unless the pre-strata are already well workable (they can be if you have made a good stratification and enough many strata) and if you have not a good pattern of auxiliary variables in *your sampling file*. As mentioned, it is good to work for getting a good sampling file with a number of variables that can predict missingness.

If you have a good sampling file, you have many other options for adjusting weights. The next 3 slides present my recommendation for constructing these weights.

It is good to note that there are two types of sampling weights whether they are adjusted or not:
-*Count weights* as above that should be summed up to the target population counts (including strata and post-strata counts).
- *Scaled weight = count weight divided by the average of the count weights.*
What is their sum?

Adjustments by response propensity modeling 1

This technique consists of the following steps:

(i) an initial weight for the respondents needs to be available. This weight may be whatever such as a basic weight, a post-stratified weight or weight based on homogenous groups. Naturally, if the initial weight is already good and there are no new auxiliary variables, the further adjustments do not make any difference or very little.

(ii) The creation of a binary response indicator, let say *resp*, so that *resp =1 for respondents and = 0 for non-respondents or deficiencies. The weight thus will be made for the respondents.*

(iii) Looking forward to good auxiliary variables to explain and to predict the variable resp, and estimate the respective model. The four different link functions can be used in this model: logit, probit, log-log or complementary log-log. So, choose one of these and estimate against your data set testing different forms of these explanatory variables (transformations, interactions).

Adjustments by response propensity modeling 2

(iv) Look at the estimates and try to find a such model specification that could predict response probabilities (and propensities) as well as possible. So, estimate these propensities too and include in the data set. The symbol for these probabilities is p_k . Check also their distribution, especially outliers and how plausible these are, that is, which types of units they concern. Design weights are useful to exploit in estimating the model.

(v) Compute the adjusted weights using the formula

$w_k(res) = (w_k/p_k)q_c$

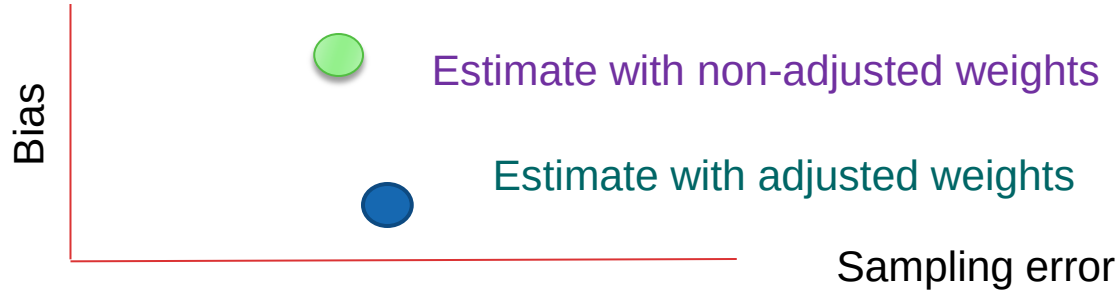
Here q_c = a scale or calibration factor that transforms the preliminary weights w_k/p_k so that the sum of the weights corresponds to the known population figures. This last requirement can be achieved with real calibration but in most cases the results are good if this scaling has been made at stratum level (either pre-stratum or post-stratum if the latter ones are available).

Adjustments by response propensity modeling 3

In the case of stratification, for each stratum *h*, it is rather easy to scale or calibrate the preliminary weights with the following ratio which form has been used earlier in computing basic weights

$$q_h = \frac{\sum_h w_k}{\sum_h w_k / p_k}$$

(vi) Check the weights and compare those with possible other weights, and start to estimate that is exactly similar as with other weights except if you wish to take correctly into account non-response effects in standard errors. This is not an easy task since there are obviously more bias in non-adjusted based estimates than in respective adjusted ones. I do not go details but look below.



Cleaned survey data

After all former operations starting from getting raw data you have cleaned data set, **almost**.

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.