Survey Weighting
University of Helsinki
Part A

Inclusion probabilities and other basics

Spring 2016

Seppo Laaksonen

Linear solutions much used but they do not work always well.

Weighting in surveys

From Survey Data Collection to Cleaned Survey Data

- Designing the survey (should already be included in the project)
- Designing the sample (planned so that weighting in easily possible)
- Data collection (some information, auxiliary data collected during this)
- Data entry (some auxiliary data from registers etc if possible)
- Imputing the data (it may be competitor to weighting)
- Creation of the sampling data file; last version after fieldwork ended
- Weighting the data
- Documenting the weighting methods and tools

Survey Data Analysis: Using the best possible weights in each analysis.

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.
We can use both English and Finnish, let see.

The target is that almost everything will be possible to do during the official course hours when attending. Maybe some additional tasks outside these events. I hope that we will enjoy.

I will upload the basic course material on the course website but the data files and SAS codes will be delivered maybe by email.

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 = 5 if the training and its report is reasonably done.
- If more trainings (these options will be seen later), more credits possible after reporting.
- A small-scale exam during the last event. We will see how?

Questions?

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* (close to the term Target Group) 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 However, we do not need much to take care of this population since we will have always known
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).

Populations in surveys 2

We as well know the following populations if we need them but we only need to know the basic information about the third population. This basic information will be given but it is good to understand its meaning.

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:
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

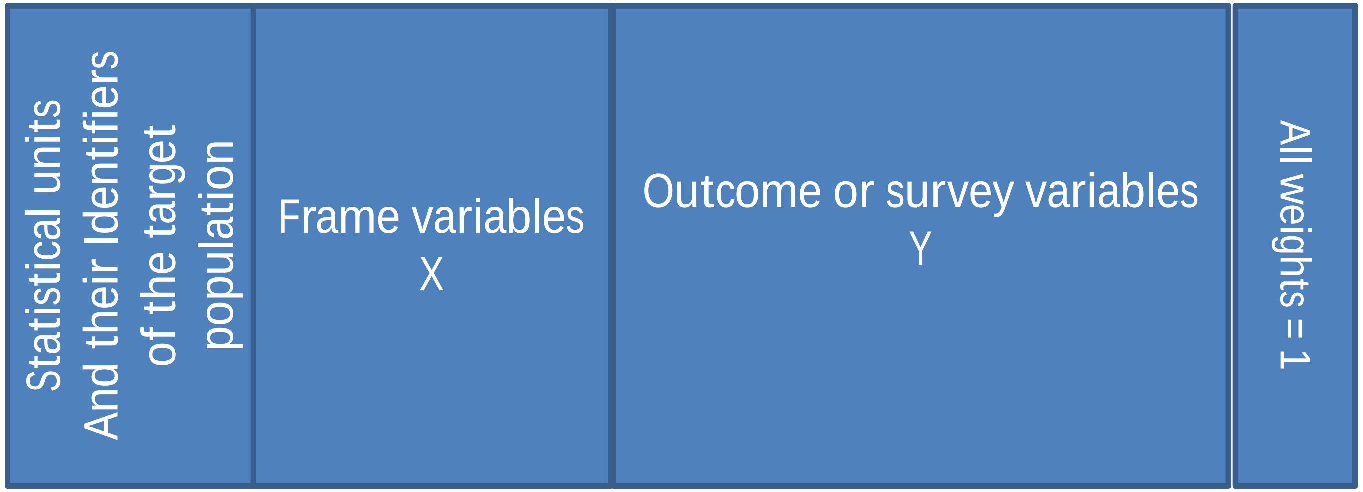
One key concept behind weighting thus is the target population that should be known when going to weight.

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. If all coverage and related problems are solved, the estimates can be generalized at the target population level. This means that the weights are needed in this estimation. Naturally, both point and interval estimates (standard errors, confidence intervals) are necessary to calculate but this course is focused on point estimates. Interval estimates are obtained using all survey instruments (strata, weights, clusters) and software like SAS, SPSS, STATA, R.

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.



Statistical units And their Identifiers of the target population

Frame variables X

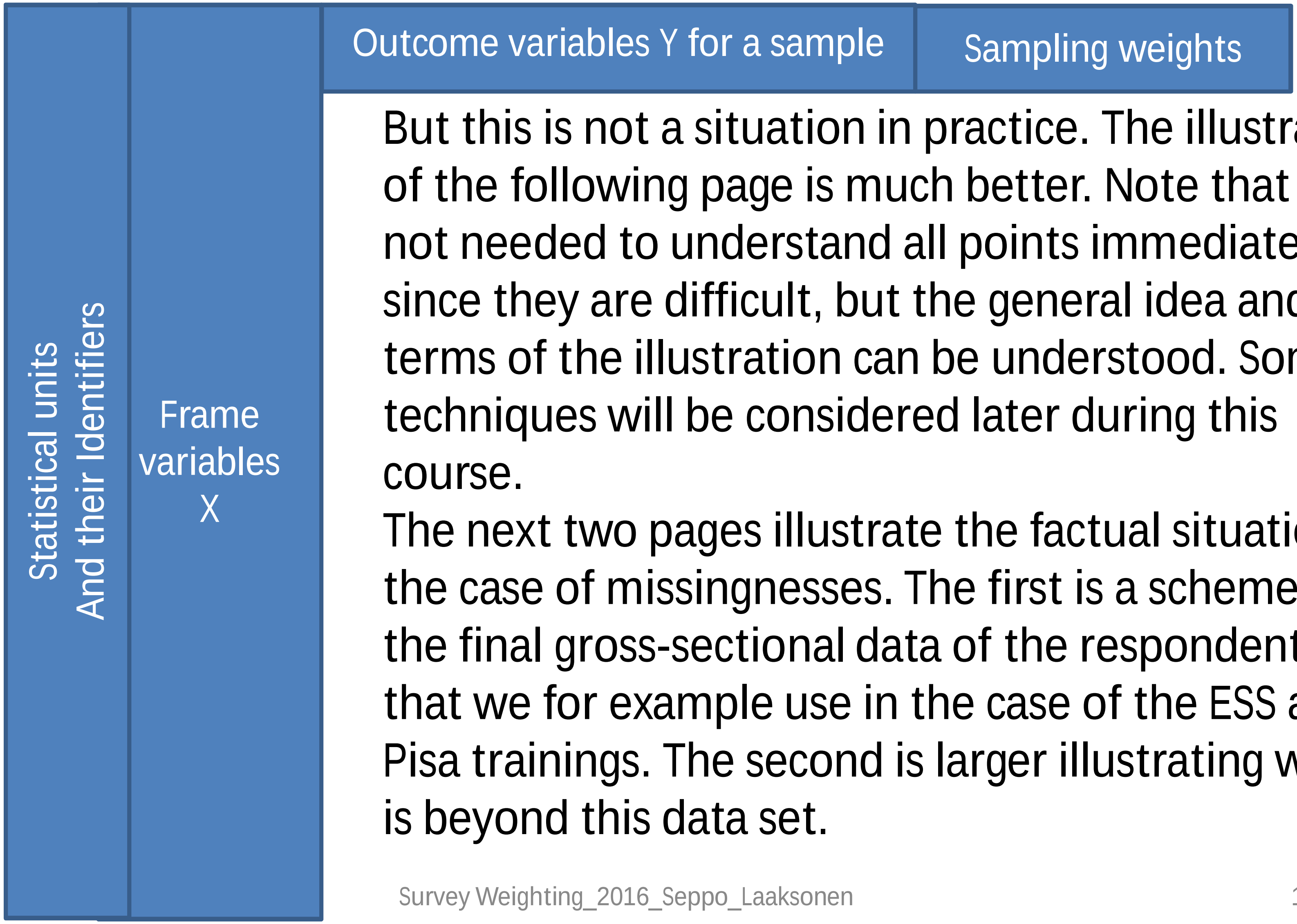
Outcome or survey variables Y

All weights = 1

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:



Statistical units
And their Identifiers

Frame variables X

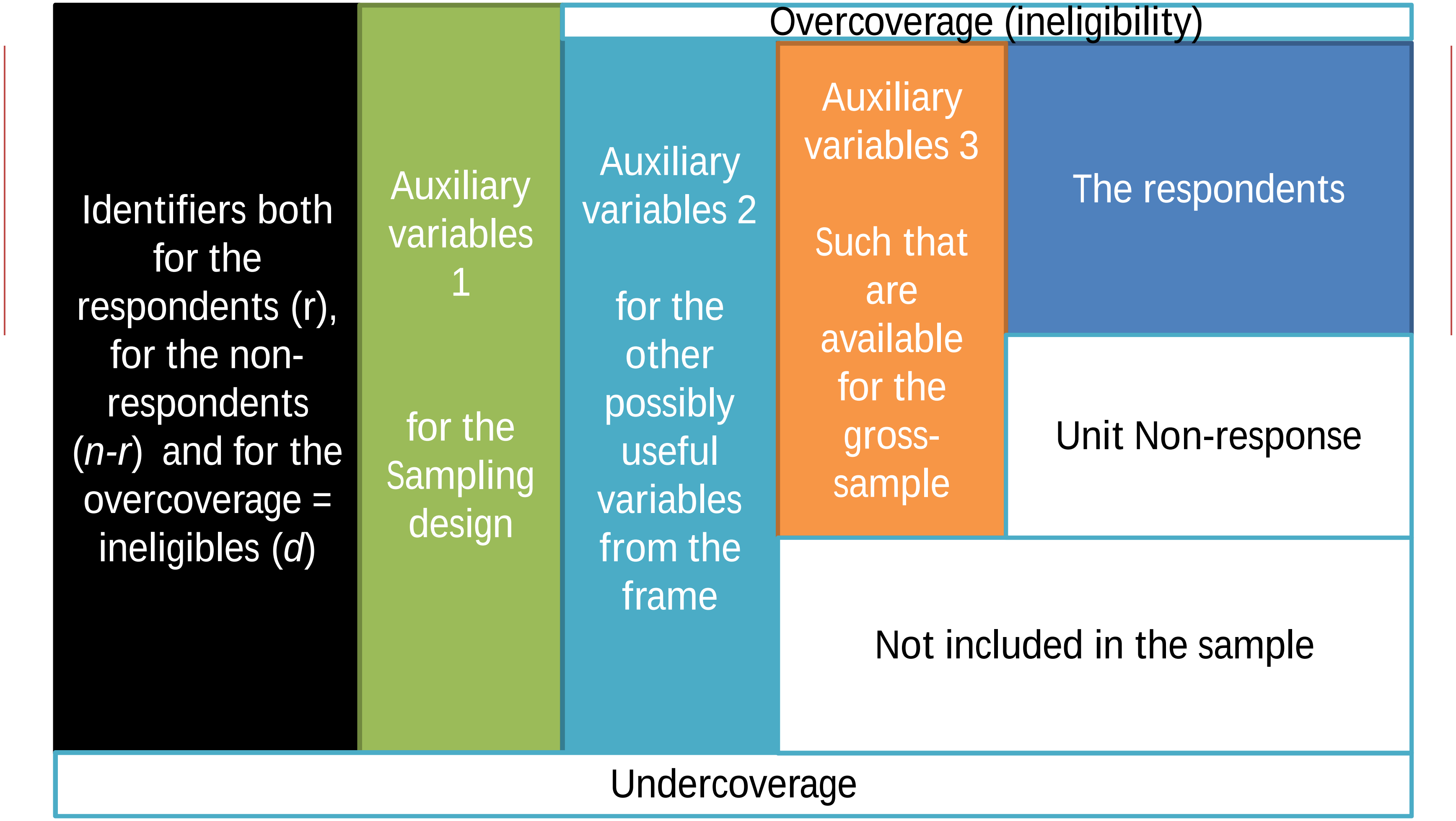
Outcome variables Y for a sample

Sampling weights

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



Overcoverage (ineligibility)

Identifiers both for the respondents (r), for the non-respondents (*n-r*) and for the overcoverage = ineligibles (*d*)

Auxiliary variables 1

for the Sampling design

Auxiliary variables 2

for the other possibly useful variables from the frame

Auxiliary variables 3

Such that are available for the gross-sample

The respondents

Unit Non-response

Not included in the sample

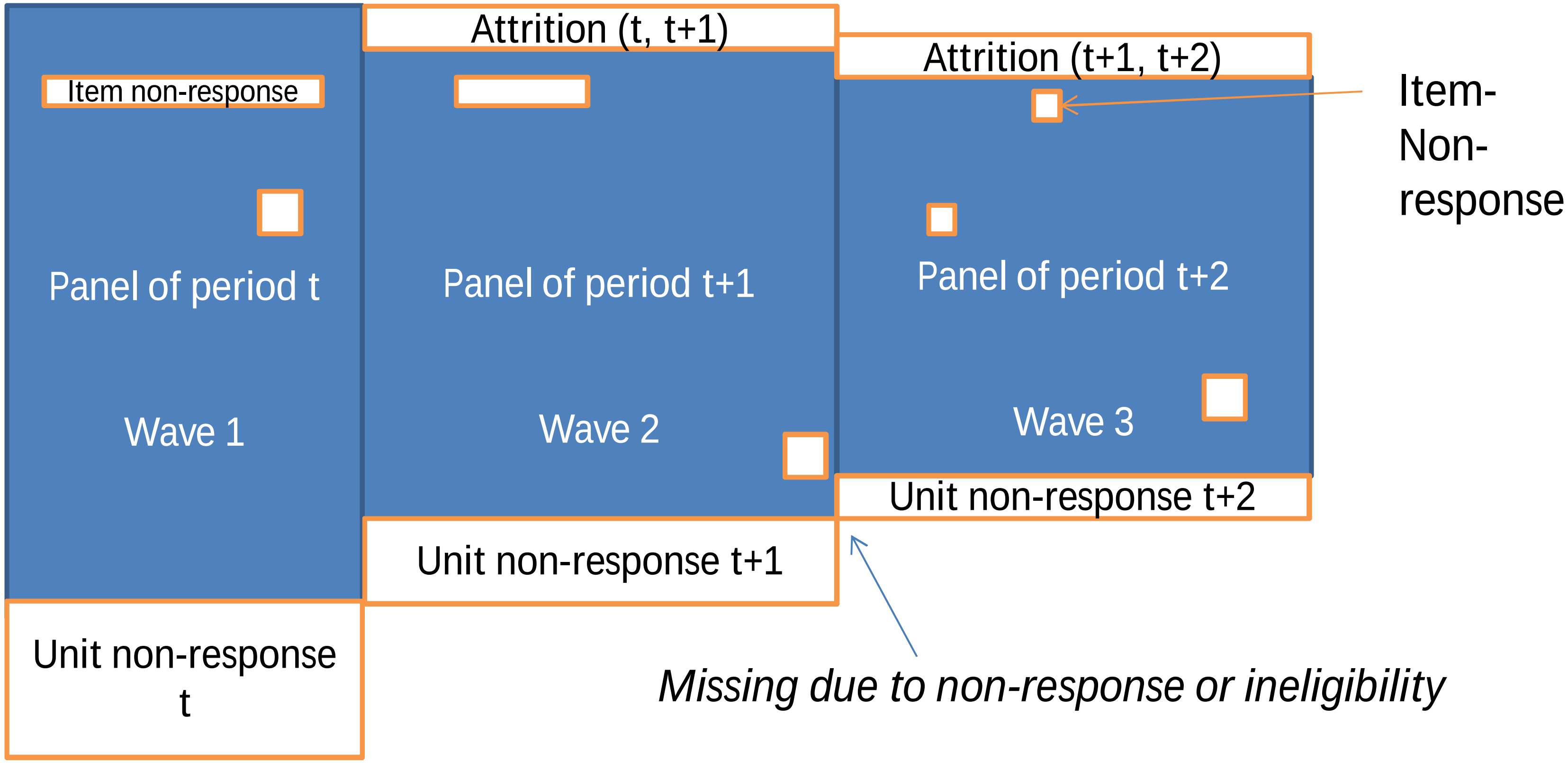
Undercoverage

Micro data and Missingness (we do not focus on this but include)

Cohort type of panel (longitudinal) example

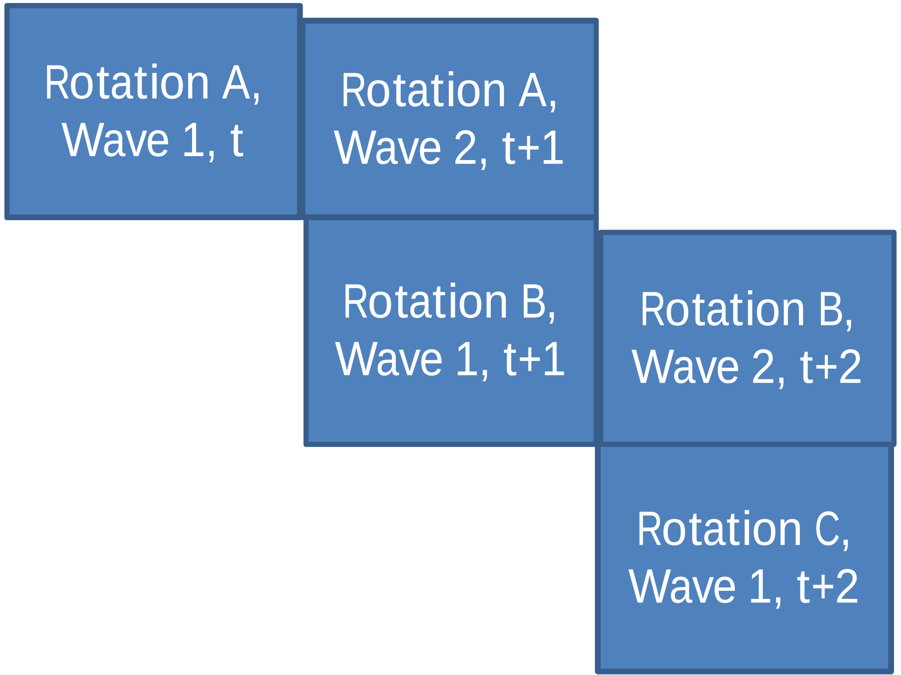
Here attrition does not include unit non-response as in some other studies

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).



Missing due to non-response or ineligibility

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.



Rotation A, Wave 1, t

Rotation A, Wave 2, t+1

Rotation B, Wave 1, t+1

Rotation B, Wave 2, t+2

Rotation C, Wave 1, t+2

A small example is possible in the end of the course.

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?

Micro data and Missingness

Examples of the terms in social surveys:

Overcoverage (ineligibles): 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?

Micro data and Missingness

- The average of sampling weights 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?

These two weights will be calculated in our training for all weights created.

The one-dimensional distributional indicators are to be the first result in our trainings. The second results are some estimates from our data sets. These will not be complex, that is, maybe the averages and other simple ones are enough, but there is an option to estimate more complex estimates, even models.

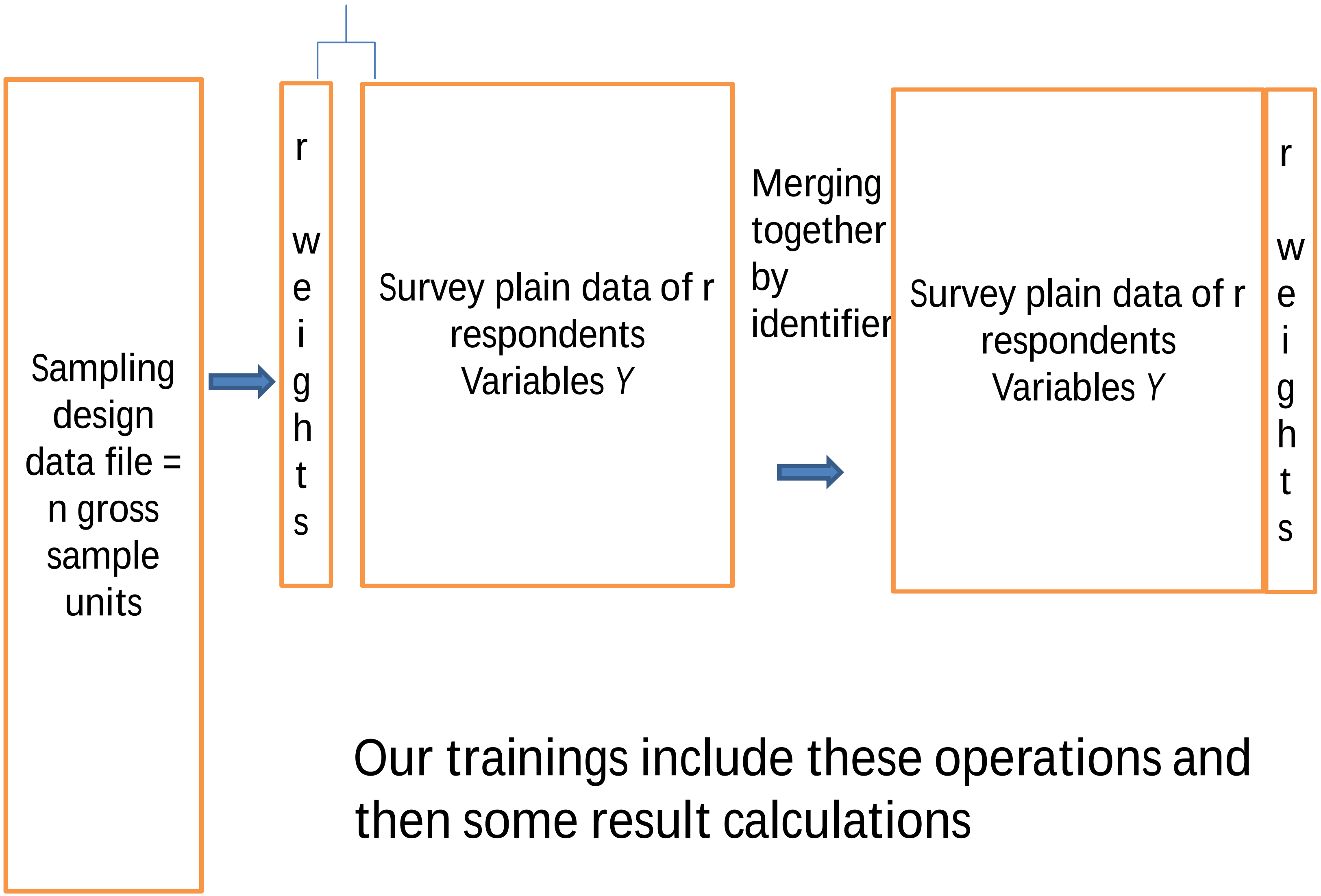
The training report thus includes these two types of results, and their interpretation; naturally it should be told how each weight was obtained.

A real survey file is not such as the scheme of the above 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. From this file we usually create the sampling weights and take other sampling design variables and merged these into. The first file was already sent you.

- The file of the respondents that thus is used in analysis. This file will be given you later, maybe next week.

Next page includes an illustration. Before going to trainings, we have to learn something about sampling designs, first most simplest ones.



Sampling design data file = n gross sample units

r weights

Survey plain data of r respondents Variables *Y*

Merging together by identifier

Survey plain data of r respondents Variables *Y*

r weights

Our trainings include these operations and then some result calculations

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
- Aggregated information e.g. from a living area characteristics (the same value for all living in this area).

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).

These X variables have several roles including:

- 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.

Summary of the concept section with key symbols

U = target population (reference population of *N*)

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

D = over-coverage or ineligibles

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

Summary of the concept section with key symbols and SAS codes

$$response\_rate = \frac{r}{n} or = \frac{r}{n+d}$$

```
data resp; set z.srs;
if outcome=1 then resp=1; else if outcome=2 then resp=0;
if outcome=1 then resp2=1; else  resp2=0;
if outcome=3 then over=1; else  over=0;
```

$$ineligibility\_rate = \frac{d}{n+d}$$

```
proc means data=resp n mean;
var resp resp2 over; run;
```

Survey Weighting
University of Helsinki
Part B

Sampling design

Spring 2016

Seppo Laaksonen

Clusters are much used in sampling. These here are very homogenous compared to many others.

Sampling design 1

My framework is for probability sampling, not for quota or other non-probability sampling. Voluntary samplings are nowadays becoming too common especially when using web arsenals. These are often non-probability methods from a sampling point of view. Before going to probability sampling details, I present some views about non-probability sampling, focusing on such principles that are not working badly, or they may be the only alternatives for certain surveys. In general, it is good to try to go as close to probability based sampling even using non-probability approaches. This especially means that the selection of sample units is as randomized as possible,

Sampling design 2

First some basic concepts:

Primary sampling unit = *psu*: the unit that has been included in the sample in the first step (stage) in sampling. This may be a cluster as in multi-stage sampling.

Secondary sampling unit = ssu: the unit that has been selected at the second stage and within each psu.

Stratum or Explicit Stratum: group or sub-population or quota that will be included definitely in the sample, thus its inclusion probability = 100%. The strata are independent of each other. This means that a different sampling method can be used in each stratum. Even though the method is the same, rules may vary by strata.

Sampling fraction: the proportion of the gross sample units of the target population. This fraction can be equal in each stratum or it may vary depending on the precision targets of a survey.

Sampling design 3

Cluster, examples:

- small area where residents or birds,
- enumeration area, census district where people
- grid square (e.g. 250mx250m) where people
- school where students or teachers
- household where its members
- address where residents or employees
- enterprise where employees

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

Selection probability: The inclusion probability of one gross sample unit.

Sampling design 4

Final inclusion probability: an inclusion probability is first determined into each stage, stratum, phase or quota and then using these probabilities into the entire gross sample level. In the simplest case, the one level only is needed and this inclusion probability is the final respectively. But in the case of a more complex design, a new calculation is needed. If the sampling of each stage is independent. the final inclusion probability is the product of all single probabilities.

The same is basically concerned a phase but It may be more complex. The stratum and quota are like sub-target populations, and hence within them there are their own rules.

Design weight or basic sampling weight: the inverse of the final inclusion probability (design weight) or its conversion into the respondents (basic weight).

Sampling design 5

Second order Inclusion probability: a probability that two sample units belongs to the target population. This is not in details considered in this course since it may be demanding, It is however good to know that this probability is particularly needed for the variance estimation or standard errors or confidence intervals. Fortunately, their estimates in ordinary cases are available in good software's that can be used without knowing their detailed technics.

Obs. The literature is not clear as far as the selection probability is concerned. Some earlier ones do not use the term 'inclusion probability' at all, that is, the selection probability is the same as the inclusion probability. This is not ideal, since the inclusion probability requires the decision about the gross sample size even though the selection method is the same.

Sampling design 6

The below taxonomy is not always used to describe which questions should be taken into account when planning the sampling in practice. It is good to point out that even though this taxonomy looks large, it is not difficult, since there does not need to think many questions in each box.

Sampling question	Description
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 (ssu's), …Terms: one-stage sampling, two-stage sampling, three-stage sampling. The first stage method is usually different than at later stages.
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. The method may vary in each phase.

Sampling design 7

Sampling question	Description
F. Allocation of the sample	How a desired gross sample has been shared into each stratum.
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 It leads to the inclusion probabilities when sample size is decided.	How to select the study units - probability equal to all (srs, equidistance, Bernoulli) or - probability varies unequally typically by size (pps =probability proportional to size)
I. Missingness anticipation or prediction	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.

Sampling design 8

Thus: choose an optimal alternative and implement all from each A to I tasks, and you will have a gross sample. Next we go to more details of the most commonly used sampling methods. We here use symbols and formulas as well but trying to describe them so that the basic points can be understood by non-mathematicians as well. This is made in two steps, first everything for a single stage or a phase or a stratum, and secondly combining some of these together in the case of our survey examples,

The first part thus is concerned single inclusion probabilities in most common samplings. They thus can be applied similarly for strata, sampling phases or sampling stages. Hence we do not use a subscript in the first part but later when combining some methods together, You can add there an available subscript as stratum, phase or stage if it is not there,

Sampling and inclusion probabilities

We present in this section most commonly used sampling selection methods, both without missingness and then for the respondents assuming that the missingness mechanism is ignorable. The sample size *n* or its other forms are decided separately, trying to achieve a good quality, but we do not discuss these issues here.

Simple random sampling (*srs*): The inclusion probability for each *k* is constant

$$p_k = n\frac{1}{N} = \frac{n}{N}$$

Respectively assuming that missingness is ignorable, the conversion to the respondents

$$p_k = r\frac{1}{N} = \frac{r}{N}$$

Sampling and inclusion probabilities 2

Bernoulli sampling (*Bs*): The same as srs, but the achieved sample size is not necessarily a fixed *n* since it varies randomly. The variation is relatively small for a big population and for a big sample size.

Equidistance sampling (*eds*): The inclusion probability for each *k* is constant

$$p_k = \frac{1}{l} = \frac{1}{\frac{N}{n}} = \frac{n}{N}$$

Here *l*=the constant interval for the selection. The first *k* should be selected randomly. This interval is decided as soon as *n* is known as you see above. The interval cannot be changed for the respondents but now some sample units are missing. If this is not selective, it is possible to apply the same formula as for *srs*.

Sampling and inclusion probabilities 3

Equal inclusion probabilities: Each *kεU* have an equal inclusion probability via *srs, Bs* and *eds* to be selected in a sample. This is a necessary condition for probability sampling.
How this is done in practice?

- (i) The frame is in an electronic form and an appropriate software package is available with a random number generator. A uniformly distributed random number within interval (0, 1) for each *k* is needed to create for a data file, e,g, variable *ran.* This number can be used for *eds* to select the first sample, and then using the interval drawing all others so that the entire frame has been passed. In case of *srs*, a technical option is to sort the units in the random order and then draw as many as needed from a whatever place forward and backward. Bs works such that if *ran*<the desired sampling fraction, it has been taken in a sample.

Sampling and inclusion probabilities 4

Equal inclusion probabilities (cont.):

(i) The frame is in an electronic form and an appropriate software package is available with a random number generator. A common practice is some ESS register countries is to use *eds* in the order of the population register. Since the members of dwelling units are there one after the other, several persons from the same dwelling are not drawn. This is often considered to be a good point. This method is also called *implicit stratification* but it has nothing doing with proper *stratification* that is also called *explicit stratification* if desired to avoid misunderstanding.

Sampling and inclusion probabilities 5

Equal inclusion probabilities:
(ii) The frame is not in an electronic form. The best solution is to upload it into an electronic form and continue as above. This is rarely possible if concerned a big population, However, this is often done at *psu* level such as villages and blocks that are not too big, e,g, below 300. This strategy is tried in several ESS countries when selecting households or dwelling but is not guaranteed how well it has been made. In Ethiopia, it was done so that the houses with households were marked in the first fieldwork day and then an equidistance selection was used to select the sample households. This could be an ideal method.

Sampling and inclusion probabilities 6

Equal inclusion probabilities:

(iii) The selection by an individual without random numbers. This is needed in the last stage of the multi-stage sampling, when an interviewer should select a 15+ years old person from those who are as old within a household or a dwelling that has been selected randomly by the survey organisation. The most common method of the ESS for this purpose is *last birthday method.* Its better version is such in which a survey interviewing day is randomised.

Sampling and inclusion probabilities 7

Unequal inclusion probabilities:
Just for clarification: the inclusion probabilities may vary by strata, quota or phase but this most common case is not considered here.

All methods demand one or more auxiliary variable to be used for the inclusion. This variable is in some sense ‘*size’* that is correlated with the inclusion probability. The ‘*size*’ variable is in most cases such improves the precision of the estimates. There are other reasons also that are mainly due to survey practice. We first present the case that has been used much in surveys where appropriate clusters are available.

Sampling and inclusion probabilities 8

Stratification in sampling

Stratification or more exactly ‘Explicit stratification’ is good to use in almost all samplings. Full simple random sampling is motivated to use in the case when any auxiliary variable for stratification for example does not exist. Of course, a good stratification maybe a challenging target, but should still be tried. In the simplest case, even using proportional allocation since it requires to get the certain statistics for stratification, the target population figures, in particular. This thus gives some light what is going to be met in final work. Let this statistics be N_h in which $h = 1,\ldots,H$ are these explicit strata. How big H could be, it is not clear but the minimum is one, usually however around 10, It is necessary that each stratum will have enough respondents finally. If the gross sample size is n_h then the inclusion probability when using simple random sampling within strata is

$$p_k = \frac{n_h}{N_h}$$

Sampling and inclusion probabilities 9

Stratification in sampling

If n_h is small, it is danger that the sampling design weight is not plausible .

Naturally, if $N_h = n_h = 1$, it is very OK.

After the fieldwork, when the counts of respondents are known in each stratum, the inclusion probability can straightforwardly be computed:

$$p_k = \frac{r_h}{N_h}$$

If r_h is zero or small, it is danger that the basic sampling weight is not plausible $w_k = \frac{N_h}{r_h}$. Naturally, if $N_h = r_h = 1$, it is very OK.

Our first training data set is for stratified simple random sampling.

That is, we come back to some other inclusion probabilities after this training.

Survey Weighting
University of Helsinki
Part C

Missingness and sampling data file

Spring 2016

Seppo Laaksonen

Missingness may be a difficult feature. If trees are missing somewhere as here, it is fairly normal but still good to know.

Missingness mechanisms 1

In this introduction to sampling, it is good to discuss missingness as well. As said earlier, missingness is due to nonresponse and ineligibility. in particular, in this context. Under-coverage or measurement errors can not be included ell in the sampling process. The term ‘missingness mechanism’ or ‘response mechanism’ is a practicable term to use here. The below terms are mainly used in ordinary literature, but I have a bit extended this list.

MN (Missing No)

This thus is a survey with a 100 per cent sample.

MI (Missing Ignorable)

The sampling fraction is 100 per cent but some missingness occurs. Nevertheless, missingness has not been taken into account, and all calculations done assuming a full response. It is not nice but used.

Missingness mechanisms 2

MCAR (Missing Completely At Random): If this were true, it would be rather easy to handle the data. The assumption MCAR is much used even though it does not hold true.

MARS *(*Missing At Random Under Sampling Design): Now missingness only depends on the sampling design. This is often used so that one assumes that MCAR holds true within strata or quota.

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 auxiliary variables are available. Thus without those, your assumption is MCAR or MARS.

MNAR (Missing Not At Random): Unfortunately this is the most common situation in real-life to some extent. So, when all the auxiliary variables have been exploited, the quality of the estimates have been improved but still it is rather clear that our results are not ideal.

Sampling design data file

The term ‘sampling design file’ that is not commonly used in survey sampling literature. The methodology behind this term is used, but implicitly. Its explicit determination facilitates many things in survey practice and also gives a clear target for one big part of a survey, that is, sampling and fieldwork. The sampling design file consists of all the gross sample units and its variables include those that give opportunity to create sampling weights and to analyse the survey quality. The file is possible to complete after the fieldwork, Its most important characteristics, including sampling design variables and weights, will be finally merged together with the real survey variables at respondent level, and then the survey analysis is ready to start.

Other sampling design issues

Sampling design data file, variables included with good meta data

(i) Inclusion probabilities of each stage

(i) Other variables directly relating to sampling design (psu that can be a cluster or an individual, explicit stratum, implicit stratum)

(ii) Outcome of the survey fieldwork (respondent, ineligible, non-respondent)

(iii) Macro auxiliary variables, statistics for the target population level (cluster psu's, explicit strata, calibration margins)

(iv) Micro auxiliary variables for individuals and their groups, e,g, gender, age, education level, regional or areal codes, language, ethnic or other background, household member data incl, children, civil status, employment status, register income, etc.

Handling unit missingness (nonresponse)

The methodology both for handling unit nonresponse and item nonresponse does not differ much from each other but what to do after that, it may differ more substantially. Thus the analysis itself in both cases is about as below:

1. To investigate the reasons for missing values.
2. To calculate the entire missingness rates, by reasons, and by domains (background variables).
3. To report and interpret the results, and publish the main points respectively.
4. To try to do everything better if possible in next surveys.

However, there are many differences in details, as described next.

Handling unit missingness (nonresponse)

2. To calculate the entire missingness rates, by reasons, and by domains.

This can be made in both cases creating a response indicator (and ineligibility indicator too) so that its value = 0 for missing cases and = 1 for non-missing cases.

Then

- The respective rates can be calculated. also by domains that are available in the data set (e.g. gender, age group, region, education, industry class, socio-economic status).

and finally.

These rates are to be calculated using each of our training data files.

Survey Weighting
University of Helsinki
Part B. CONTINUES

Sampling design

Spring 2016

Seppo Laaksonen

Clusters are much used in sampling. These are very similar compared to many others.

Sampling and inclusion probabilities 10

Unequal inclusion probabilities:

(i) Probability proportional to size (*pps*)

The size x_c is inserted in the inclusion probability as follows,, The subscript *c* refers to a cluster that is used at the first stage of sampling. The ESS clusters are more or less small areas, whereas they are school classes in the Pisa.

$$p_k = n\frac{x_c}{\sum_U x_c}$$

This method is rarely used alone, but thus at the first stage. The denominator is the sum of *x*'s in the frame (as the sum of school classes, or all small area clusters), not any figure of the target population units of the survey (as the sum of the students in all schools). But if the second stage units are added, the sum = *N.*

Sampling and inclusion probabilities 11

Unequal inclusion probabilities:

(i) Probability proportional to size (*pps*), continued

The sum of the target population are obtained when the second stage has been added. This is presented in the case of a ESS sampling case below.

The design *pps* can be used both with replacement and without replacement. The latter is used in most surveys as the ESS since it is technically easier. This may lead to a inclusion probability higher than one. It thus should not be accepted. How to avoid this problem, it is not discussed here but the best strategy is to keep psu sizes as equal enough. We will not have this problem in our training data sets.

Sampling and inclusion probabilities 12

Unequal inclusion probabilities:
(ii) Other cases (Interpretation)

To get all inclusion probabilities it depends on the data sources of the country. If the central population is register is available, all alternatives can be used. This is not the case in many countries but in some there are local registers within small area psu's.

If any register of individuals does not exist, this leads to three or more stages so that the stages after the psu selection are

- an address or a dwelling and
- then one individual is selected and interviewed at the next stage.

Sampling and inclusion probabilities 13

Unequal inclusion probabilities:
(ii) Other cases (Interpretation)
If there are several stages, it is expected that the final inclusion probabilities vary more than in the single stage.
The first stage probability may be much varying and if there are address/dwelling/household stages then, their size naturally varies as well. But e.g. the number of single households will be smaller respectively that is not always a bad thing.

When the inclusion probability varies, it is expected that the respective sampling weight varies as well. The variation is one component of the accuracy of the survey estimates, and hence the variation is best to keep at suitable level.

Stratified two-stage sampling for some ESS countries

This design thus is used in the ESS so that the first stage units are areas, and the second stage units are individuals respectively drawing a proper random sample.

The fieldwork has been as successful that missing clusters are lacking or they are very rare. The nonresponse within clusters (for individuals) vary a lot, even going as high as 80 per cent in some countries. The formula still works but the estimates are obviously biased.

Stratified three-stage sampling for some ESS countries

The first stage probability is *pps* and here presented so that the cluster = primary sampling unit = psu, $p_{psu} = \frac{n_{psu} x_{psu}}{\sum_U x_{psu}}$

The second stage is most often an address or a dwelling, and the sampling is *srs* as well as possible, hence the inclusion probability is

$p_a = \frac{m_{psu}}{x_{psu}}$ Here m_{psu} is cluster sample size that varies from 4 to about 30 by country. A smaller size is advantageous from the precision point of view.

In the third stage one individual is selected from the sample address or the dwelling. This is *srs* selection as well and the inclusion probability $p_{3k} = \frac{1}{m_k}$ in which m_k is 15+ old years persons

(=1, …,12) within the address or the dwelling.

Stratified three-stage sampling for some ESS countries

The final inclusion probability is the product of these three inclusion probabilities

$$\pi_k = \frac{n_{psu} m_{psu}}{(\sum_U x_{psu}) m_k}$$

The inverse of this formula is the (gross sample) design weight. When continuing toward the basic weight, many numbers will be changed due to nonresponse but if the updated frame is available, even

$$\sum_U x_{psu}$$

is revised but this occurs rarely. The number of clusters n_{psu} should remain the same. The cluster size m_{psu} will be the net size due to nonresponse. The 15+ years old persons m_k is not usually changing. Thus k=1,…,r now, for the design weight k=1,…n.

This last probability is not possible to get in all countries as in our second data file, but our third file includes these completely.

Survey Weighting
University of Helsinki
Part B

Spring 2016

Seppo Laaksonen

Lagrange Multiplier is used in Calibration Weighting methods helping minimization

Content

Response Rates
Auxiliary variables, Covariates
Response propensity modeling
Design weights
Basic weights
Post-stratification calibration
Raking ratio calibration
Generalised regression estimation and calibration
Calmar and Calmar 2
PSW = Propensity score weighting (response propensity modeling weighting)
PSW plus calibration
General conclusion

Handling unit missingness (nonresponse)

We continue and repeat the first

2. To calculate the entire missingness rates, by reasons, and by domains.

This can be made in both cases creating a response indicator so that its value = 0 for missing cases and = 1 for non-missing cases.

Then

- The respective rates can be calculated. also by domains that are available in the data set (e.g. gender, age group, region, education, industry class, socio-economic status).

and finally

After that it is good to estimate a multivariate model, called

3. the response propensity model so that the response indicator is the dependent (response) variable, and all possible domains or auxiliary variables are attempted as independent or explanatory variables.

When we have learned enough about missingness and estimated the response propensity model with different link functions (logit, probit, complementary log, log-log), we thus understand our data quality to some extent, and thus we are ready to go to reweighting.

Weighting and Re-weighting process

It can be considered to cover the following 7 actions:

(i) Sampling design before the fieldwork
(ii) Weights for the gross-sample (*n*) using (i), 'design weights'
(iii) Sampling Design File before and after the fieldwork, this includes auxiliary variables from registers, other administrative sources, also from the fieldwork
(iv) 'Basic weights' for the net sample or for the respondents (*r*), assuming *MARS*
(v) Re-weighting strategies assuming MAR(C): specification, estimation, outputs
(vi) Estimation: point-estimates, variance estimation = sampling variance plus variance due to missingness.
(vii) Critical look at the results including benchmarking these against recent results (how plausible they are?)

Two types of Auxiliary variables (covariates)

(i) Macro or aggregate:

- Known (frame) population statistics by strata or post-strata or calibration margin (benchmarking): e.g. region, gender, age group, industry, number of employees and their share by gender, age group, occupation and education,

(ii) Micro:

- All variables (possibly useful but not known necessarily in advance) that are available both for the respondents and for the non-respondents: e.g. a code for region, area, psu, gender, age or age group, industry, education level, marital status, year of marriage, socio-economic group, dwelling unit size, number of children in a household, type of home, type of living area (grid), number of rooms, mother tongue, citizenship, employment status, living or not in a municipality born, R&D intensity, ownership, …

Towards weighting

We should always have a valid sampling design, that can be simple or more or less complex. Some examples soon.

Each sampling design is determined for a gross sample. But the data file after the fieldwork is available (for most variables) for a net sample only, that is, for the unit respondents.

The core variables in the sampling design data file include:

- Identity code (both confidential and non-confidential)
- All inclusion probabilities of the sampling design
- Other sampling design variables (stratum, psu, …)
- Auxiliary variables (above)
- Survey modes (single, mixed. multi)
- Fieldwork outcome
- Technical variables and good meta data

Inclusion probabilities

We should always have a valid sampling design, that can be simple or more or less complex.

Each sampling design is determined for a gross sample. But the data after the fieldwork is available (for most variables) for a net sample only, that is, for the unit respondents.
And we have calculated based on this design

- The design weights for the gross sample

But due to unit nonresponse, we also need the weights for the net sample, i.e., for the respondents.
I call these 'basic weights' or 'base weights' but some use 'design weights' for these as well, but note that this assumes that non-response is ignorable, e.g. within explicit strata but not necessarily within 'cluster *psu*'s'. However, the whole psu rarely is missing, instead units within *psu*'s.

Inclusion probabilities

The sampling file should include one inclusion probability variable at minimum. This is the case in one stage sampling. But the number of the probabilities is growing while more stages are used in the design.

Usually, the inclusion probabilities are independent of each other, that is, the final inclusion probability is the product of all stage probabilities.
There are designs in which case these probabilities are not independent but thus we do not consider these cases in details.
Note however that it is possible that all probabilities are not known for all units, i.e. there may be missingness for all or some nonrespondents.

Towards Re-weighting

As in the previous examples, we can have

- A specific sampling design, simple or more or less complex

And we have calculated based on this design

- The design weights for the gross sample

But due to unit nonresponse, we also need the weights for the net sample, i.e., for the respondents.
I call these 'basic weights' or 'base weights' but some use 'design weights' for these as well, but note that this assumes that non-response is ignorable, e.g. within explicit strata but not necessarily within 'cluster *psu*'s'. However, the whole psu rarely is missing, instead units within *psu*'s.

Inverse probability weighting is used in clinical studies for these weights.

Towards Re-weighting 2

Re-weighting thus starts from the valid basic weights that will be tried to improve so that the estimates will be less biased than the initial ones. Usually, there is not in mind to improve all estimates but some key estimates. The other estimates are often improved at the same time but not maybe all of them.

As already seen, good auxiliary data are necessary to make re-weighting successful. If you have little good auxiliary variables, you cannot do much. So, you have to work for the auxiliary data service hardly during the survey process.

Re-weighting methods

I do not try to explain all possible re-weighting methods since they are too many. Often it is however difficult to recognise what a certain method is about since so many various terms are used. I will not be an exception. My terms are somewhat new for you, I guess, but they are in my opinion quite clear, I hope.

I will concentrate on the two methodologies
Calibration and Propensity weighting (called also response propensity based weighting)
And their combination, or synergic application.
This could be called Joint Propensity and Calibration Weighting (JPCW).

Before that; I briefly describe Post-stratification that is possibly the most common reweighting method.

Post-stratification

Is a basic calibration method that is useful to apply if you have such population level data (macro auxiliary data) that are not yet exploited in the sampling design. This is often the case.

Post-stratification is not, unfortunately, simple still, since it is conditional to the initial sampling design. This means that there may be difficulties to compute appropriate post-stratified weights. A big problem is often that the data is too small in some post-strata. THIS is obviously the main reason why the other calibration methods are developed. We consider them later in this part. First however, we explain how to implement post-stratification, or how to create the post-stratified sampling weights?

Post-stratification 2

If the sampling design is simple random sampling, you can create the post-stratified weights:

- If your data file consists of a categorical variable for which the target population statistics are available.
- Naturally, the number of respondents should be big enough as in ordinary stratification.

The post-stratified weights have the same form as in stratification, that is, (in which *g* means a post-stratum *g=1,..,G.)*

$$w_k = \frac{N_g}{r_g}$$

This method is often used even though it is not known how close to *srs* the sampling is. For example, when obtained by CATI a number of respondents more or less randomly, the weights are calculated assuming that they are selected randomly within post-strata.

Post-stratification 3

I present another case that is maybe most common. Now the sample has been drawn by explicit stratification and with a certain allocation. The strata are symbolised by *h*. When the respondents are known, responding problems are found. For example, if the stratification is regional as it is often, the basic weights adjust for regional representativeness, not for anything else. However, it was found that females participated better than males, and educated people as well. This may lead to post-stratification given that the target population statistics are available at the same categories as in the survey data file. The tabulation of next page illustrates the situation.

Illustrating post-stratification

	Initial stratification = Pre-stratification						
	Region 1		Region 2		Region R		
Post-strata within pre-strata	Little educated	More educated	Males	Females	Little educated males	Little educated females	More educated males and females

It thus is possible to flexibly create the post-strata within each pre-stratum. Its purpose is either that the response rates vary by these post-strata, or the target is to reduce the sampling error that occurs if post-strata are more homogenous than initial strata. The weights are of the same form as all stratified weights

$$w_k = \frac{N_{hg}}{r_{hg}}$$

Response propensity models

The model is most commonly a binary regression model in which the link function is either logit, probit, log-log or complementary log-log.

A second alternative is to build a classification tree model with similar variables, but we do not here go to details of this model. Instead. we present those about the two most common models, logit and probit.

They can be implemented as all statistical models, trying to find best explanatory variables for the response indicator, the interactios are possible to try as well.

The model estimates can be interpreted as usually but it is also good to calculate the estimated response probabilities that are called often propensity scores.

The strategy for creating ‘propensity sampling reweights’ is as follows

(i) We have the gross sample design weights that are the inverses of the inclusion probabilities. Explicit stratification is used.

(ii) We assume that the response mechanism within each stratum is ignorable (MARS), and hence compute the initial (basic) weights analogously to the weights (i). These are available only for the respondents *k*, and symbolised by w_k.

(iii) Next we take those initial weights and divide these by the estimated response probabilities (called also response propensities) of each respondent obtained from the probit or logit model, and symbolised by p_{k}.

(iv) Before going forward, it is good to check that the probabilities p_{k} are realistic, that is, they are not too small, for instance. All probabilities are below 1, naturally.

The strategy for creating ‘propensity sampling reweights’, continues

(v) Since the sum of the weights (iii) does not match to the known population statistics by strata *h*, they should be calibrated so that the sums are equal to the sums of the initial weights in each stratum. This is made by multiplying the weights (iii) by the ratio

$$q_h = \frac{\text{å}_h w_k}{\text{å}_h w_k / p_k}$$

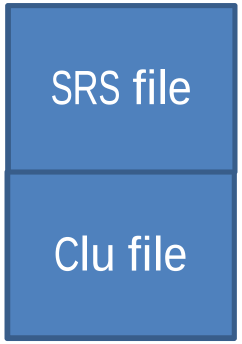
(vi) It is good also to check these weights against basic statistics. If the weights are not plausible, the model should be revised.

The above formula also here

$$q_h = \frac{\text{å}_h w_k}{\text{å}_h w_k / p_k}$$

In Training

Our target is to create response propensity weights for both data, i.e. for SRS = Stratified Simple Random Sample data and CLU = Stratified Three-Stage Cluster Sample data. First, it is required the basic weights that we have done for SRS. Note that the weights can be made in one occasion if those two data sets are pooled together.



Calibration

The basic idea of the calibration is thus to calibrate the re-weights so that the certain margins (macro auxiliary statistics) are correct. There are a number of strategies to succeed with this target. Usually the algorithm is such that the distance between the initial weight and the calibrated weight will be minimized. There are different technical and methodological tools to do it. The French INSEE software CALMAR 2 uses Lagrange multipliers that gives easier opportunity to apply several distance functions but in the case of linear function other tools are obviously easier.

Calibration

We here first apply the linear distance function that is most frequently used. It can be called linear calibration.

The macro auxiliary variables are needed in calibration. They thus are wished to be true values or true margins of the target population. These variables are next symbolized by x_p variables that number p is not big, let say about 3 to 7. Each such variable includes a number of categories with these true values, thus being vectors. The weight w_k is the variable that will be calibrated and the calibration weight respectively is c_k. Both these weights thus are for the respondents.

Calibration formulas

The first requirement is to minimize the distance function.

$$D(c_k, w_k) = \sum_U w_k G(\frac{c_k}{w_k})$$

In the case of linear calibration the distance function

$$G = \frac{1}{2}(x_p - 1)^2$$

Finally, the minimization is done so that *p* calibration equations holds true

$$\sum_r c_k x_p = \sum_U x_p$$

Some constraints can be added such as the ratio $\frac{c_k}{w_k}$ should be in a desired interval,
e.g. lowest=0.2 and its largest symmetric counterparty
= 1/0.2 =5.

Calibration practice

The first step to calibrate is to find the correct ‘true values’ of auxiliary variables. The following ones are often used in social surveys:

- Gender (two categories)
- Age group (5 to 10 categories)
- Large region (5 to 10 categories)
- Education level (4 to 7 categories).

These aggregates are then saved in a specific file and used in the calibration software so that the software tries to do its best to get such weights that satisfy these calibration margins. The algorithm usually succeeds but it is not guaranteed that the calibrated weights are ideal. It may be possible that they are negative or below one.

Calibration by Calmar 2

Calmar 2 is a new version of the initial Calmar that can be downloaded from INSEE website. Some maybe do not like that the manuals are in French, but it is good for everyone to learn some basics of this language.

Calmar 2 is a SAS macro as the initial Calmar as well. This means that you cannot do your own applications but insert necessary parameter values in the programme only. It was not easy to start to work with it but I found a person (Josiane Guennec) from INSEE who was willing to help us to use the software and the document: Sautory, Olivier ja Le Guennec, Josiane (2005). La macro Calmar 2: Redressement d'un échantillon par calage sur marges. Institut National de la Statistique et des Etudes Economiques Direction Generale.) .

Calibration by Calmar 2

The basic idea is thus to calibrate the re-weights so that the certain margins (macro auxiliary statistics) are correct (as true values as possible, or benchmarked values). There are a number of strategies to succeed with this target. Usually, the algorithm is such that the distance between the initial (basic) weight and the calibrated weight will be minimized. It is easy to notice that the distance function can be different. Calmar 2 gives opportunity to apply the five alternatives:

1. Linear
2. Raking ratio that is in fact exponential or logarithmic
3. Logit
4. Truncated linear
5. Sinus hyperbolicus

Calibration by Calmar 2

A user has to choose the starting weight that is ‘basic weight’ in our pure Calmar application. Secondly, he/she have to create a file that includes the margins. The number of margins and their categories are technically limited, but in practice, it is good to be realistic. There are in Calmar 2 also two margin levels possible, such as for individuals and for households, respectively. The third point needed is to choose one of these five methods. The methods give opportunity to put the certain constraints as follows: lower limit and the upper limit of the ratio of ‘calibrated weight/starting weight.’ This may be useful in order to avoid negative weights and other extreme weights. This option is both in method ‘Logit’ and ‘Truncated linear.’ Raking ratio and sinus hyperbolicus (both are exponential based) do not provide negative weights.

Calibration by Calmar 2

There are many nice things in CALMAR 2. For example, it shows what is the distribution of categories of auxiliary variables based on the initial weights and respectively, the true values that should be achieved by calibration.

CALMAR 2 thus gives such types of figures for all margins, and it is in principle possible to calculate an overall summary of these differences.

CALMAR 2 SAS MACRO

Calibration margins creating the SAS file ‘kkk’ if used three margins. The name can be whatever else.

Number of categories used, all cells should be there, if not available, insert the missing value code.

```
data kkk;
input VAR $ N MAR1 MAR2 MAR3 MAR4 MAR5 MAR6 MAR7 MAR8;
cards;
stratex 8 8565169 1642680 5207001 1450863 783288 647136 1147896 462240
Gndr         2 9924418 9981855 . . . . .  . .
Agegroup 5  3557207 4962788 5309909 3973562 2102807 . . .
;
Run;
```

OBS: the sum of margins should be exactly equal; otherwise the algorithm does not work. Here = 19906273

CALMAR 2 SAS MACRO

SAS MACRO

```
OPTIONS MSTORED SASMSTORE = cal;
   %CALMAR2 (DATAMEN=sample,MARMEN=kkk,
POIDS=w_resp,IDENT=idno,M= 1,DATAPOI=testi1,POIDSFIN=wcal1);
             run;
```

in which:

SASMSTORE = cal : the library where the Calmar macro is; I have used another folder than where the other files are.
DATAMEN = the file name in which the starting weights, e.g. 'W_BASIC' are available, but you can start from another weight as well
MARMEN = the SAS file of calibration margins
POIDS = the starting weight in file 'sample'
IDENT = identifier variable
M = Calibration method (1=Linear, 2=Raking ratio, etc.)
DATAPOI = the name of the output file
POIDSFIN = variable name of the calibrated weights.

New strategies

(i) Combine Calibration and Propensity Weighting

(ii) Use the design weights in estimating propensities. This often helps but does not influence much in other cases.

The procedure by CALMAR 2 is similar but it is only needed to choose the propensity weights as the starting weights. It is also possible to start from the post-stratified weights.

Our training data set gives opportunity to test different starting weights. It is not automatically clear which weights finally be best. However, it is expected that if more auxiliary variables are exploited (micro and macro), the estimates will be less biased. UNLESS: anything bad things have not been done such as low quality auxiliary data, non-good handling of outliers or small data.

Special weighting cases at general level

1. Sampling of individuals, estimates for clusters such as households

This occurs in the European Social Survey, among others.

Question: How to estimate household level estimates such as the household composition or average household size?

A: Since the individual weights are summed up to 15+ population, the estimates using such weights are concerned respective individuals. We however need to know the estimates of households. The number of households for each unit is obtained by dividing the individual weight with the number of 15+ persons, thus nothing needed if the 15+ size is =1 but in the case of larger households, the weight will be smaller respectively.

```
w_basic_HH=w_basic/members15Plus;
```

Special weighting cases at general level

1. Sampling of individuals, estimates for groups such as households

Results from our main SDDF using the basic weights
2.497
And with their better form
2.105

Thus: a big difference, should be found if done incorrectly that the individual weights lead to a too high household size.
Compare with other weights and also estimate the distribution of household size, and the number of households.

Special weighting cases at general level

2. Sampling and weights for households, estimates for individuals or other lower level

We thus have correct household level weights w_k that are used for household estimation, concerning *income*, for instance.
The file includes also the number of household members, with their ages and household position.
It is possible to estimate some figures so that their importance can be seen at individual level or other lower level. The weights e.g. for individuals are obtained by multiplying with the number of household members (or equivalent consumption units) . This thus means that the sum of the weights = the sum of individuals of the target population. It is another question which estimates can be now calculated if these variable values vary within households.

Special weighting cases at general level

3. Panel of two years

Let the weights of year *t* be $w_k(t)$, and those of year *t+1* respectively be $w_k(t+1)$, but the number of *k* is not necessarily the same, being smaller in the second year due to missingness. However the weights are concerned the same units (households, e.g.). Moreover, we wish to calculate the changes at unit level from *t* to *t+1*. How to do it in a best way?

A: If we cannot impute the missing values, we just exclude such units that are not correct in both years (<u>balanced panel</u>). Now our *k* is smaller than in year *t*, but maybe equal with that of year *t+1*. The three alternative weights can be applied:

- The weights of year *t* (corresponds to Paasche's index)
- The weights of *t+1* (corresponds to Laspeyres' index)
- The (geometric) average of both weights (Fisher's ideal index).

What do you choose?

Happy Re-Weighting
and Re-Winter,
not maybe doing
everything completely
linearly