Survey Methodology
Part C

Fall 2015

Seppo Laaksonen

Sampling Principles
Missingness Mechanisms

Sampling design

You saw in the scheme of Part A that a sample survey data have some missingness, both intentional and inevitable. Intentional is mostly due to sampling, inevitable due to nonresponse, ineligibility, under-coverage and measurement errors.

Part C is focused on sampling that is used for several reasons but the most important one is that population is big but the resources are limited, hence a 100 percent sample is not possible. On the other hand, it is possible to get more precise estimates by sampling. The third key point is that sampling in its efficient form gives opportunity to get the desired estimates rapidly enough. I include some missingness questions in the sampling, since it thus is reality in almost all surveys,

I present a compact framework for sampling. This is called sampling design. Often a narrower framework has been given.

Sampling design 2

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 3

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: 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 4

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 5

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 6

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.

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.

Non-probability sampling, examples

Self-selection sampling is already discussed in Part B. This thus is much used in web surveys, but also in telephone surveys. The method may be acceptable in online television programmes when a journalist is invited the audience to comment certain actual topics using a simple binary question, 'Yes' or 'No.' Such answers cannot be generalized onto any concrete population. This is well understood by the major part of the audience itself, but it is possible that some people believe that the result represents the opinion of the residents of the country. This bias is more obvious if the journalist convinces the result being reliable, saying for example that a certain big number of answers were received, e,g, some thousands that look fine. Nevertheless, it is not possible to say anything about the quality of such a study. All estimates can be close to a true value with good luck.

Non-probability sampling, examples

Self-selection sampling may be in some cases the only way to get some light to a certain phenomenon. I would say that if any good sampling frame to approach to a target population does not exist, a preliminary estimate might be obtained using a well managed invitation to participate in a survey. The invitation could be published in a media that has been followed by those expected to participate and to respond to a survey. It is still danger that one respondent replies more times than once. This can be avoided to some extent in web surveys, at least, so that the answer can be not given two times from the same computer.

Non-probability sampling, examples

‘Hard-to-reach’ is a term used to describe those sub-groups of the population that may be difficult to reach or involve in research or public service programmes, among others. Application of a single term to call these sub-sections of populations implies a homogeneity within distinct groups, which does not necessarily exist. Different sampling techniques are attempted, the most common ones are briefly presented below.

Snowball sampling

Snowball sampling is a non-probability method used when the desired sample characteristic are rare or when the studied population is broader and more heterogeneous than that can be easily accessible through other more reliable sampling methods.

Non-probability sampling, examples

Respondent-driven Sampling (RDS)

The main criticism about chain-referral or snowball sampling is bias toward recruiting more cooperative subjects and masking which is protecting close friends or relatives by not referring them when specially there is a strong privacy concern associated with the subject of the study. It is also suggested that those with extended personal networks to be over-sampled and isolated people to be excluded in the study.

Indigenous field worker sampling (IFWS)

In this sampling method instead of using formal trained investigators, they are selected from local community. Then they undergo special training relevant with objectives of the study including interview skills and fieldwork protocol. The selected people should have privileged access to the study target population. It is believed that use of this technique can reduce masking, volunteer bias and under-reporting of socially undesirable behaviors.

Non-probability sampling, examples

Facility-based sampling (FBS)

Facility-based sampling refers to recruiting members of target population from a variety of facilities including correctional and drug treatment centers, sexually transmitted diseases clinics or general health centers and hospitals in certain sub-urban areas. Each of these facilities can be used to recruit individuals from hidden population, but similar biases may occur due to under-sampling of those who are reluctant to seek and obtain services especially when their behaviors are stigmatized.

Targeted sampling (TS)

The targeted or purposive sampling method has been developed to overcome the limitation of snowball sampling when we would like to include specific pre-defined subgroups of population in our sample. This sampling method generally includes an initial assessment aimed at identifying the various subgroups that might exist in the population of interest. The identified subgroups are then regarded as strata.

Adaptive Sampling

Adaptive sampling is a sampling technique that is implemented while a survey is being fielded—that is, the sampling design is modified in real time as data collection continues—based on what has been learned from previous sampling that has been completed, Its purpose is to improve the selection of elements during the remainder of the sampling, thereby improving the representativeness of the data that the entire sample yields.

This technique has also used when trying to find new respondents in the case when nonresponse of certain sub-populations is too high after a certain time during the fieldwork. This thus means that more attempts to approach such units (persons or via persons) are made.

Non-probability sampling, examples

A survey without any proper sampling design when going to collect survey data, I have met this case when a subject-matter researcher creates a questionnaire that he/she is using first for his/her own clients in social services, for example. Since the results look interesting but any reference group does not exist for comparing, he/she has invited clients of other institutions to participate. Now the number of the respondents is growing and the results look more interesting. But when he/she wishes to publish them, it is not automatically possible without a generalization to a target population. What to do? The first point is to determine the target population and to get its statistics. The second point is to decide the sampling design afterwards, It is not possible well, of course, but if the selection of the respondents is random within each sub-group (stratum), the weights can be created. But if the selection is not random, it leads to a bias. However, the generalization is possible if the above assumptions (MARS) holds true.

Non-probability sampling, examples

Opinion polls of market research institutes are often based on CATI surveys. They have created strata or quota before calling, The quota are based in the Finnish case, for example, on the cross-classification of two genders, 5 age groups and 4 regions, altogether 2x5x4 = 40 quota. It is known from a recent population statistics, how many target population people belongs to each quota, let say N_h in which h is a quota. The client of a survey institute decides the overall size of the respondents (e,g, r=2000). The survey institute calculates the proportions of each quota $q_h = \frac{N_h}{N}$ and the basic option is to

allocate the number of the respondents relatively equally to each quota, that is r_h=q_h. It is not guaranteed that this number exactly will be reached during a fairly short fieldwork period. Hence the survey institute states a certain minimum for its CATI centre at least and a maximum respectively. This strategy also assumes that the mechanism MARS holds true.

Non-probability sampling, examples

Opinion polls of market research institutes

This method is partially probability based and the survey weights will be based on the assumption that the respondents are selected at random within each quota. This may hold fairly well but not completely at all due to the following reasons:

- If a person does not answer a telephone call, he/she will be automatically out of the survey.
- If person denies to participate, he/she will also be out.
- The first questions of the survey are concerned the quota's, gender, age group and region. If a quota to which a person belongs is already full (a maximum is reached), he/she does not need to answer at all.

All the above points mean that there is no nonresponse in some sense at all, since a non-respondent has been replaced by a respondent. This is the main problem of this method but works if non-respondents do not matter much as in case of voting behavior questions.

Sampling design 7

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 8

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, equi-distance, 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 9

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 MARS or 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

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

Respectively assuming that MARS holds true, the conversion to the respondents

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

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.

Equi-distance sampling (*eds*): The inclusion probability for each *k* is constant

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

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:

(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:

(ii) 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

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.

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

This method is rarely used alone, but thus at the first stage. The denominator is thus the sum of *x*'s in the frame (as the sum of school classes), 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 9

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.

pps can be used both with replacement and without replacement. The latter is used in most surveys as the ESS. This may lead to a inclusion probability higher than one. It thus should not be accepted, How to avoid this problem, it is discussed below.

Sampling and inclusion probabilities 10

Unequal inclusion probabilities:

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

If *pps* is used at one stage sampling, it is best to apply another selection method as usually but they are not necessary for ordinary surveys. The main option to avoid probabilities above 1 is not to use too big cluster sizes. The above formula gives the following condition for the maximum cluster size:

$$x_c \leq \frac{\sum_U x_c}{n}$$

Or respectively for the sample size

$$n \leq \frac{\sum_U x_c}{x_c}$$

We here see that either the sample size or the maximum cluster size or both should not be too big.

Sampling and inclusion probabilities 11

Unequal inclusion probabilities:

(ii) Other cases

The most common practical reason for an unequal inclusion probability is such in which the frame units are available from the other level than those of the study units. Let N = the size of the target population at a cluster level (e,g, household, address, dwelling), and the respective sample size is n. Furthermore, there are m_k individuals in each cluster and their target population size is of Nm_k. Note that n is different in these two options:

- If a sample with n individuals is drawn, we get as many sample clusters respectively, and thus the inclusion probability of cluster k is

$$\pi_k = \frac{n}{m_k N}$$

Sampling and inclusion probabilities 12

Unequal inclusion probabilities:

(ii) Other cases

- If the situation is opposite and the sample of *n* clusters is drawn, but the individuals are our study units, then the inclusion probability is

$$\pi_k = \frac{m_k n}{N}.$$

Here the numerator thus indicates the number of sample individuals.

Sampling and inclusion probabilities 13

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

The former of the above options is used in register countries since the register consists of the individuals of the desired ages and it is rather easy to draw those, and after that to form the clusters who are close enough to them. These can be formed at household level since the household composition can known when contacting sample units.
The latter option thus is opposite and used almost in all other countries when one stage is an address or a dwelling and one individual is interviewed at the next stage.

Sampling and inclusion probabilities 14

Unequal inclusion probabilities:
(ii) Other cases (Interpretation)
Both options have the disadvantage since the inclusion probability varies and not always in best way, It increases linearly by the cluster size in the first case. It is good thing if wished to get more sample units for big households, for instance. The number of single households will be smaller respectively that is not always a bad thing.
The individual inclusion probability is equal to the household inclusion probability in the second option.

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.

Example when household size varies and if it is used in sampling

To illustrate the impact of unequal inclusion probabilities, the below table is from the ESS. There are two variables, one of the household size and the second for the size who belong to the target population, 15 + years old, You see that the mean and the variation of the latter is smaller. The variation is one component of the accuracy or the survey estimates. This can be measured with the DEFFp= Design Effect due to varying inclusion probabilities that is approximately = 1 + cv^2. We come back to this and other DEFF's but its is good to notice this indicator of the accuracy (standard error is one measure for this). On other words, the same accuracy will be achieved when increasing the sample size as now with 21 percent (DEFFp=1,21). This indicator is invented by Leslie Kish in 1960's (His family had to leave Hungary after 1st World war).

Variable Label	Mean	Coefficient of Variation (cv)	DEFFp
Number of people living regularly as member of household	2,39	0,568	1,32
Number of 15+ years old people	2,00	0,455	1,21

Sampling and inclusion probabilities 15

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 I have seen, is a bit below 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 with strata is

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

Sampling and inclusion probabilities 16

Stratification in sampling

This method is also called stratified random sampling, It is obviously the most common method, in all kinds of surveys, including business surveys where stratification is necessary to include the large businesses of each industry class in the sample since their impact in most statistics in enormous. After the fieldwork, when the counts of respondents are known in each stratum, the inclusion probability can straightforwardly computed:

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

If r_h is zero or small, it 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.

Example of special inclusion probabilities

Stratified two-stage sampling for Pisa,

These formulas are given for each stratum but subscript *h* is not mentioned. You can add it if you wish. Strata in Pisa are used in all countries but it is not included in the micro data file in all, but in most countries. Strata are either based on regions or areas, or the combination of region and type of schools.

In the first stage school classes where are Pisa students are selected by *pps.* If x_s = the class size (Pisa students), and n_s = the respective sample size, then this probability is

$$\pi_s = \frac{n_s x_s}{\sum_U x_s}$$

In the second stage, a student is selected with simple random sampling, and thus the inclusion probability is simply

$\pi_p = \frac{n_p}{x_s}$ in which n_p = the sample size of students. However, if it is ≤ 35 then all are included.

Example of special inclusion probabilities

Stratified two-stage sampling for Pisa,

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

$$\pi_s = \frac{n_p n_s x_s}{x_s \sum_U x_s} = \frac{n_p n_s}{\sum_U x_s}$$

There does not the sample class size in the final formula but it thus is needed in order to do everything correctly. It however gives opportunity for bluff if bot inclusion probabilities are not publicly available. We present soon the tool that can and should be used for this purpose, the sampling design data file.

It is easy to take the inverse of the formula and to get the design sampling weight.

The basic sampling weight is straightforwardly obtained from the design weight replacing n_p's with r_p's. There are strict rules for responding, about 10 percent nonresponse only may be accepted. Class nonresponse is not accepted at all.

Example of special inclusion probabilities

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.

Example of special inclusion probabilities

Stratified three-stage sampling for some ESS countries

The first stage probability is here presented in a bit differently so that the cluster = primary sampling unit = psu, $\pi_{psu} = \frac{n_{psu} x_{psu}}{\sum_U x_{psu}}$

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 $\pi_a = \frac{m_{psu}}{x_{psu}}$ 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 $\pi_{3k} = \frac{1}{m_k}$ in which m_k is 15+ old years persons

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

Example of special inclusion probabilities

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 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 but their number will due to nonresponse, respectively. Thus k=1,...,r now, for the design weight k=1,...n.

Other sampling design issues

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.

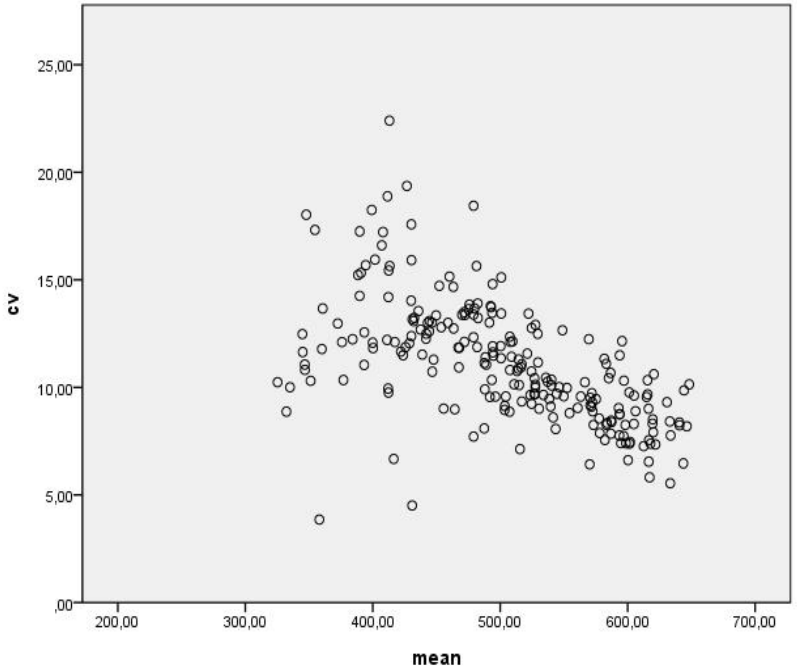
Here is an example of the fairly good sampling design data file
The size thus is the gross sample size

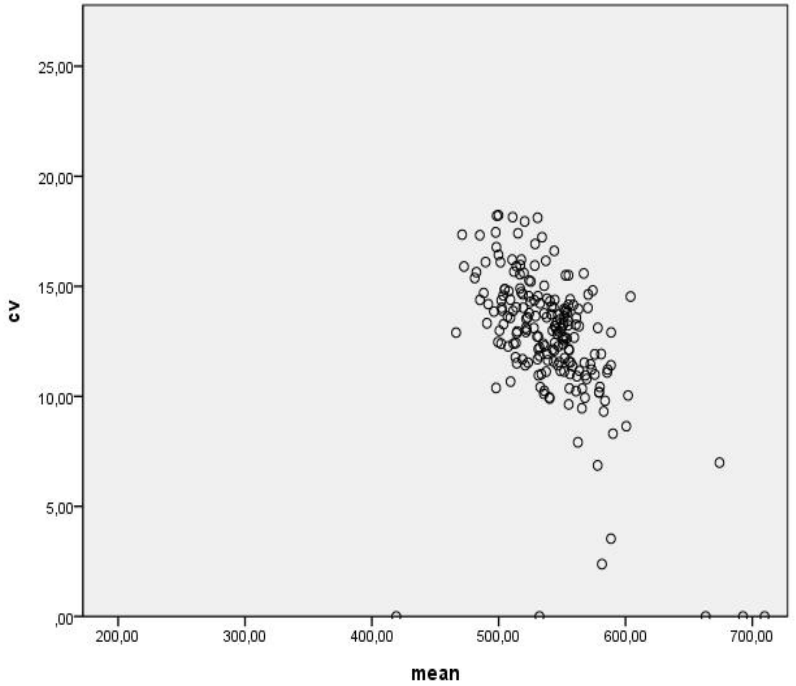
	Respondent's identification number	1st stage selection probability: town or rural area	2nd stage selection probability: village in rural areas	3rd stage selection probability: respondents in PSU	Primary sampling unit: village or town	Explicit stratum: Categories of PSU	Field outcome	Gender	Village vs. town size category	Number of inhabitants aged 15+ in PSU	Region	Year_married	Room_number	children	Kids	Housesize	education	Wife, husband, cohabiting partner
1	6	1.00	1.00	0.000	501	501	2	2	9	1495365	14	0	1	0	0	29	00	0
2	7	0.18	0.05	0.007	119	100	1	1	1	568	20	1993	5	2	0	119	6	0
3	15	0.12	0.10	0.006	201	100	1	1	1	686	26	1963	3	2	0	73	00	0
4	16	1.00	1.00	0.000	579	579	1	2	5	41979	18	2001	4	2	2	120	3	0
5	19	1.00	1.00	0.000	501	501	2	2	9	1495365	14	1980	5	2	0	130	5	0
6	26	1.00	1.00	0.000	520	520	1	2	7	199813	24	2002	4	3	0	110	3	0
7	33	1.00	1.00	0.000	586	586	2	1	8	559955	2	1965	2	2	0	43	3	0
8	34	1.00	1.00	0.000	576	576	2	1	7	357136	32	1972	5	2	0	107	3	1
9	45	0.16	1.00	0.000	175	300	1	2	3	8949	2	0	.	0	0	0	5	0
10	46	0.11	0.30	0.002	323	100	1	2	1	1954	18	1984	5	1	0	117	3	0
11	48	1.00	1.00	0.000	530	530	1	1	5	49251	24	1982	2	2	0	44	3	0
12	57	0.18	0.02	0.019	402	100	3	1	1	214	2	1979	.	3	1	0	3	0
13	60	0.16	0.10	0.004	324	100	2	2	1	961	18	0	3	1	1	113	5	0
14	63	0.10	0.02	0.031	168	100	1	1	1	127	30	0	1	0	0	32	00	0
15	69	0.24	1.00	0.000	313	400	1	2	4	17019	14	0	1	1	0	37	3	0

	Respondent's identification number	1st stage selection probability: town or rural area	2nd stage selection probability: village in rural areas	3rd stage selection probability: respondents in PSU	Primary sampling unit: village or town	Explicit stratum: Categories of PSU	Field outcome	Gender	Village vs. town size category	Number of inhabitants aged 15+ in PSU	Region	Year_married	Room_number	children	Kids	Housesize	education	Wife, husband, cohabiting partner
2318	9122	0.11	0.01	0.049	306	100	1	2	1	81	30	1978	1	2	0	52	3	0
2319	9125	1.00	1.00	0.000	576	576	3	2	7	357136	32	1972	4	2	0	96	3	0
2320	9126	0.53	1.00	0.000	336	400	1	1	4	37035	10	0	5	0	0	140	3	0
2321	9129	0.09	0.21	0.004	274	100	1	1	1	1098	16	0	3	0	0	70	3	0
2322	9132	0.08	1.00	0.001	117	200	1	2	2	4492	20	2003	3	2	2	115	00	0
2323	9136	0.42	1.00	0.000	359	400	2	2	4	28865	32	1979	2	2	0	59	3	0
2324	9137	1.00	1.00	0.000	566	566	1	1	8	498126	30	1991	2	2	2	55	5	0
2325	9142	0.11	0.01	0.045	409	100	1	1	1	88	6	0	1	0	0	34	3	0
2326	9143	1.00	1.00	0.000	501	501	1	1	9	1495365	14	0	3	0	0	120	00	0
2327	9152	0.21	1.00	0.000	247	300	3	1	3	11871	12	1982	4	2	1	120	3	0
2328	9159	1.00	1.00	0.000	565	565	1	2	6	108390	14	0	2	0	0	59	00	1
2329	9160	0.21	1.00	0.000	255	300	1	2	3	12157	28	0	3	1	0	83	00	0
2330	9164	0.20	0.14	0.002	312	100	1	2	1	1602	18	0	2	0	0	49	3	0
2331	9166	0.60	1.00	0.000	182	400	1	2	4	41659	24	0	4	0	0	83	3	0
2332	9168	1.00	1.00	0.000	553	553	2	2	8	681482	10	1965	3	5	0	75	7	0
2333	9171	0.09	0.20	0.004	127	100	1	2	1	1079	12	1977	4	2	0	95	8	0
2334	9175	0.26	1.00	0.000	170	400	2	2	4	18259	2	0	3	0	0	72	00	0
2335	9177	1.00	1.00	0.000	557	557	2	1	6	149324	28	0	4	0	0	110	00	0

Other sampling design issues, Custering

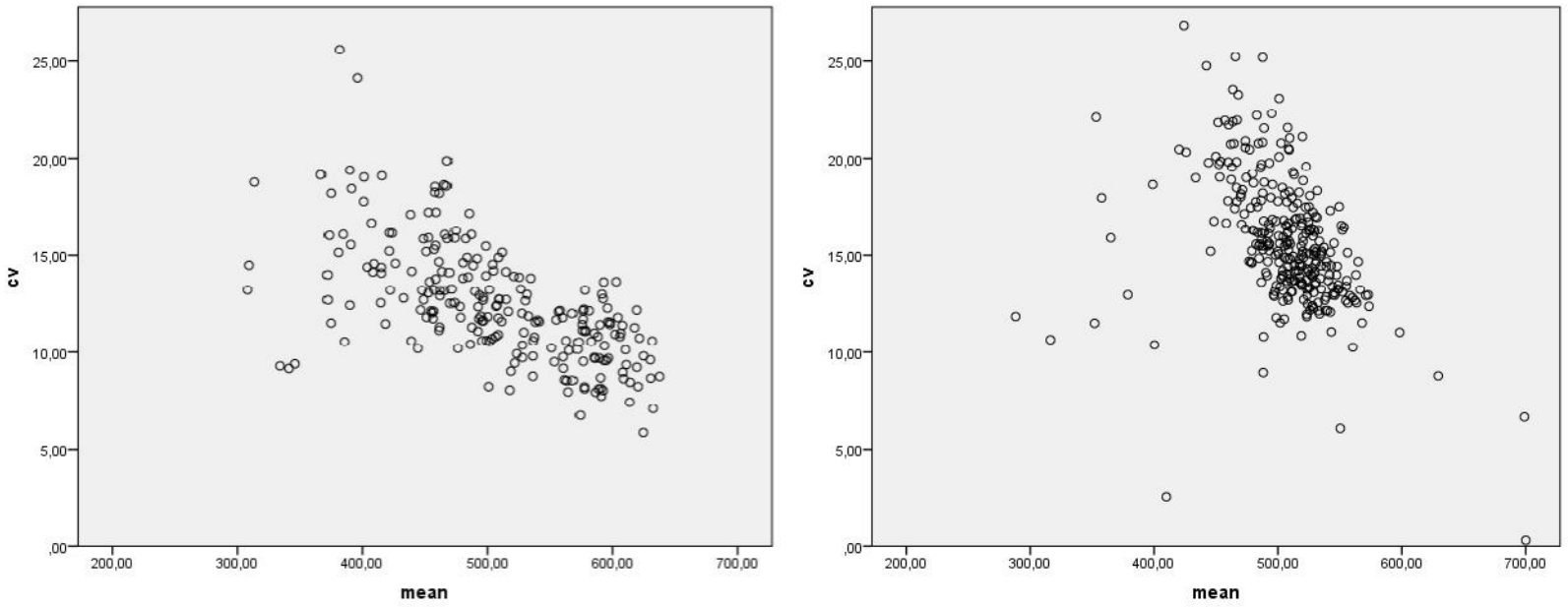
For the introduction: Here are two scatter plots from the 2009 PISA so that the x-axis is the mean of success rate for mathematical-statistical literacy, and the y-axis the coefficient of the variation, respectively. The plots are the PISA sample schools. May you guess, which graph is for Germany, and which for Finland?





Other sampling design issues, Clustering

Here are the same scatters from the 2012 Pisa. May you find any changes?



Other sampling design issues, Clustering

We saw fairly clear clusterings, especially in the first graph for Finland. Here the results between schools are fairly close to each if comparing with German results. On other words, the variation between German schools look large, and between Finnish schools small, both relative to individual variation that is here seen only within schools, but can be measured for all students. If this overall variation is called *Total_variation* and measured with the variance, and the respective variance between schools is symbolized by *Between_variation* then it possible compute their ratio. It is called intra-class correlation = rho

$$rho = \frac{Between\_variation}{Total\_variation}$$

This correlation get values within the interval [0, 1]. It is a usual concept in many meanings.

Other sampling design issues, Clustering

Intra-class correlation *is important in sampling* given that *psu's* are clusters. Thus if the psu is a single unit as a person or an enterprise, it is not needed to worry about intra-class correlation, since its value is equal to 0. Intra-class correlation may still be used for examining the interviewer effect for example, In this case, the interviewers form the 'class' and it is not good if this rho is high, since it means that the interviewers do not do their work objectively enough.

The intra-class correlation is an indicator for (in)homogeneity of clusters. It varies a lot from a survey to the next. For example in PISA 2009, where the clusters are school classes and the variables are scores of literacy, it is for Finland around 0.10 but for Germany around 0.5-0.6. This means that differences between schools are much higher in Germany than in Finland. The rho's are increased in 2012 (almost 0.15) for Finland but reduced for Germany (around 0.4).

Other sampling design issues, Clustering

We already mention the concept *DEFF (design effect) in one meaning, that is* Due to unequal inclusion probabilities *DEFFp.*

This is good to take into account when designing the sample size, It is also important to examine the design effect Due to clustering DEFFc since it is more or less above one, It is approximately

DEFFc = 1 + (b-1)rho (b=average *net* cluster size*)*

We thus have two *DEFF*'s:

The whole *DEFF* is the product of both *DEFF*'s = *DEFFp x DEFFc*,
It should be thus predicted or anticipated when designing the sample.
This is a special demanding job but the earlier data helps naturally.

Other sampling design issues, Clustering

The ESS Intra-class correlations are in general lower than in Pisa. This is due to clusters that thus are regions or areas. These thus are not as homogenous as school classes. Naturally, rho depends also on the variable. The table illustrates this. We see that rho is fairly high as expected since robbery rate e.g. is in some areas high, in some others not.

Variable	Rho	DEFF
Robbery rate	0.0294	1.43
Robbery fear	0.0198	1.29
Opinion: mothers should stay more at home, not to work	0.0287	1.42
Opinion: talented students should be more awarded	0.0028	1.04
Happiness	0.0005	1.01

Other sampling design issues, Clustering

The intra-class correlations thus influence on the *DEFFc*'s but the net cluster size as well, Hence the cluster size is good to keep small enough in face-to-face interviewing, but not too small, since then the interviewers should travel too much. On the other hand, the number of clusters should be enough big since it improves the quality. The Pisa requires that 150 schools at minimum should be included in the sample. This amount is achieved in most ESS countries, but many have much more.

Sampling design summary: gross sample size

The ESS *rho*'s are much smaller, since *psu*'s are small areas that are not as homogenous. Typically *rho* is around 0,02-0,05, in some countries higher (depending on the estimate). 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 summary: gross sample size

Effective sample size is an important concept when determining the gross sample size. It corresponds to the sample size in which case the micro data of respondents could give the same accuracy as the simple random (*srs*) gives. Thus is the net sample data can really be interpreted drawn from a target population with srs, we do not need even sampling weights for getting good accuracy estimates. Unfortunately, this is not the case in real life but its still good compare the achieved data relative to *srs* data.

Thus any estimate cannot be calculated as this *srs* that is the standard error of the mean, s^2 = ordinary sample variance and the data is of the respondents = *r*.

$$stderr = sqrt(\frac{\sum_r s^2}{r})$$

We will focus later on analysing the data correctly even not being *srs* based.

Sampling design summary for determining gross sample size if the effective sampling size is decided

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% 2857/.95 = 3008
4. Anticipated Design Effect (DEFF) due to clustering including anticipated intra-class correlation (=0,025), average net cluster size (=5.3) (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

Example: UK three-stage sampling design of the round 5 of the ESS
NO STRATA

	Sample stage	Sample size	Per cent
Inclusion probability at first stage of sampling (selection of PSUs)	1st stage	4640	100
Inclusion probability at second stage of sampling (selection of addresses)	2nd stage	4640	100
Inclusion of selection of dwelling unit at sampled address	2nd stage	4070	87.7
Inclusion of selection of household at selected dwelling unit	2nd stage	3835	82.7
Probability of selection of adult at selected household	3rd stage	3460	74.6
Unit response	Fieldwork	2366	51.0

Factually, there are five stages in the sampling data file. This is a usual case in countries that they can calculate missingness with a stage.

Example: Three-stage sampling for Russia with 10 regional explicit strata

	Sample stage	Sample size	Per cent
Inclusion probability at first stage of sampling (selection of PSUs)	1st stage	3982	100
Inclusion probability of household	2nd stage	3982	100
Inclusion probability of adult at selected household	3rd stage	2595	65.2
Unit response	Fieldwork	2595	65.2

You see that they did not try to get any detailed data on non-respondent households that is quite common outside register countries

As always, the response rates vary and this is good to take into account in sample allocation as was made in Russia. Today, too many countries do not take care of this problem but hopefully do in a future.

REGION OF SETTLEMENT = Explicit Stratum	Sampling fraction of indi-viduals, %	Gross sample size	Net sample size	Response rate, %
1	0.0024	451	233	51.7
2	0.0026	989	564	57.0
3	0.0027	190	154	81.1
4	0.0029	199	148	74.4
5	0.0030	436	310	71.1
6	0.0032	439	345	78.6
7	0.0032	491	356	72.5
8	0.0032	396	248	62.6
9	0.0034	222	130	58.6
10	0.0034	169	107	63.3

We find from the table that sampling fractions are not equal. There is a slight inverse relationship between the response rate and the fraction, especially in two first strata. This strategy is good since it gives opportunity to get more respondents at these strata. It would be good to go toward the same strategy in some other strata since some plots are spread. Naturally we do not know what has been their strategy but this is obvious.