Chapter 3

Sampling Designing, Missingness Mechanisms and Design Weighting

We present a compact framework for sampling. This is called **sampling design**. Often a narrower framework has been given. The 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, we 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 inquiries. In general, it is good to try to go as close as possible to probability based sampling even using non-probability approaches. This especially means that the selection of sample units is as randomized as possible.

This chapter covers also principles of design-based weighting. These weights are fairly straightforwardly calculated from the inclusion probabilities, as their inverses, thus not taking advantage of proper auxiliary variables which weights are considered in Chapter 7. There does not exist any unique terminology for these weights but as far as the design based weights of the gross sample weights are concerned, the simpler term '**design weight'** has been often used, the European Social Survey being one example. At contrast, several terms are used for the weights of the respondents, thus the weights that are needed in survey estimation. Some are still using the term 'design weights' but it may be confusing since these weights are not completely design based due to nonresponse and other missingness. We use the term '**basic weight**' in which case we assume that missingness is ignorable.

some basic concepts, both for probability and non-probability sampling

Sampling/sample unit = *unit that has been included in the gross sample.*

Primary sampling unit = *PSU*: the sample unit that has been included in the sample in the first step/stage. The PSU may be a final sample unit as in single stage sampling, or 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. The SSU is missing in single stage sampling.

Cluster is a group of 'individuals' who are close enough to each other.

E*xamples in surveys:*

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

Stratum: group or sub-population or quota that will be included definitely in the sample. 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.

Selection probability: The inclusion probability of one sample unit in each stage.

(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, leading to under-coverage automatically. If this is intentional, it should not be allowed.

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 (but a general problem is that all single inclusion probabilities are not known for all units. It is possible that these probabilities are not independent but we do not consider such cases in details (cf. Laaksonen et al 2015). *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.*

In this introduction to sampling, it is good to discuss missingness as well. As said earlier, some missingness occurs always, due to nonresponse and ineligibility, in particular. Under-coverage or measurement errors cannot be included well in the sampling process.

The term 'missingness mechanism' or 'response mechanism' is a practicable term to be used here. The below terms are mainly used in ordinary literature, but we have a bit extended this list.

MN (Missing No)
This thus is a survey with a 100 per cent sample, and without missingness.

MI (Missing Ignorable)
The sampling fraction is 100 per cent but some missingness occurs. Nevertheless, missingness has not been taken into account, and all

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

easy to handle the data. The assumption MCAR is much used even though it does not hold true.

*MARS (*Missing At Random Conditional to 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 quotas, but not between them.

*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 either 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

Non-probability sampling cases

Desired to generalize afterwards to a target population that might incorrectly determined, or it approximately such or hoped to be such or convinced to outsiders that it is such.

Non-probability sampling cases

Opinion polls of market research institutes
are often based on CATI surveys. They have created strata or quota before calling, The quota are based, for example, on the cross-classification of 2 genders, 5 age groups and 4 regions, altogether 2x5x4 = 40 quotas.

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 (proportionally) to each quota, that is $r_h = rq_h$.

If the respondents are a random selection of each quota target population, the MARS mechanism holds true and we can be happy with our data quality. It is hard to know how good the quality is, really.

This method thus is partially probability based and the survey weights will be calculated assuming that the respondents are selected at random within each quota. This may hold true 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 refuses to participate, he/she will also be out.

Self-selection

Most often very problematic but it can be tried. If a good survey has been conducted at the same, it is possible to reduce the bias and to get more respondents (and thus the precision will be improved). But this is not any main use of self-selection that is to get quickly some results that can be very biased.

Next page: The example in the leading Finnish Newspaper Helsingin Sanomat in which Self-selection is used. First, a medical doctor told on her unbusinesslike (bad) treatment as a patient and then the newspaper motivated readers to tell about their experiences. It was of course that negative experiences were expected to receive (63% mostly negative). This is not any research study even though more than 6000 people replied on the website. Note that a good research study about this topic is not easy either from the point of view of sampling, nor of the questionnaire designing nor of the data collection.

My rare experiences have been always positive.

Helsingin Sanomat 16 Sept 2016

Keskustelu sai alkunsa lääkäri **Riitta Korpisaaren** keskiviikkona 14.9. HS:ssa julkaistusta mielipidekirjoituksesta. Siinä Korpisaari kertoi muun muassa saaneensa potilaana epäasiallista kohtelua hoitajilta.

HS pyysi lukijoita kertomaan kokemuksistaan päivystysvastaanotoilta. HS.fissä kyselyyn vastasi torstai-iltapäivään mennessä yli 6 000 ihmistä.

HS:n kyselyn vastaajista 63 prosenttia kertoo kokemustensa päivystyksestä olevan pääsääntöisesti kielteisiä.

Helsingin Sanomat
16 Sept 2016

"Hoitajat vitsailivat potilaista"

HS:n kyselyn vastaajista 63 prosenttia kertoo kokemustensa päivystyksestä olevan pääsääntöisesti kielteisiä.

"Päivystyksessä ei hoidettu sisäistä verenvuotoa."

"Potilaita ei huomioida"

"Sain huonoa hoitoa"

"Henkilökuntaa liian vähän"

Snowball sampling and/or Respondent-driven Sampling (RDS)

In some cases maybe the only way to get some estimates, thus the frame information is missing as often in the case of special groups.

Adaptive Sampling

Not much different from the RDS.

A recent application: Responsive design: when the first part of the fieldwork has shown which groups do not seem to participate well, these groups have been tried to motivate more effectively in the remaining fieldwork period.

Online or Internet or Web panels

An Online panel is a group of selected participants who have agreed to reply survey questions by web during a reasonable time. What this time is, it depends on a survey organization but it is usually expected that it lasts more than one year. Some incentives have been usually given so that the incentive increases by the time of participation.

The Internet panel thus is not as a single web survey but can be such. The recruitments for the panel can be made using some conducted surveys asking in the end of the interviewing, for example, whether she/he could be willing in future to be the member of the panel. Naturally, such persons who responded in the survey, are more often asked than non-respondents. On the other hand, if she/he has access to Internet and enough ability to use it, could be good candidates to recruit.

It is of course possible to make the recruitment so that those gaps can be filled by giving the computer with Internet access to those who have not it, and train everyone to use in replying. At the same time, they can use the computer for their own purposes as a good incentive. This investment is naturally more expensive and not much used, one exception is the Dutch LISS panel (https://www.lissdata.nl/lissdata/).

The number of registered persons in Online panels can be big, several tens thousands. This gives opportunity to draw a special sample for each survey. Respectively, the workload for one person is suitable.

It has been discussed whether a good online panel could replace an ordinary survey since it may ensure a reasonable response rate, compared to that of the ordinary repeated survey.

What did you love or hate about your sampling or survey data analysis course?
I'm looking for two things basically:
1) Syllabi for sampling or survey data analysis courses you've taken or tought. Good and bad examples both wanted.
2) If you've take a course on this, we'd love your frank thoughts (again positive or negative) even if you can't share the syllabus. Direct message is fine if you don't feel comfortable commenting "out loud."
Colleague Stas Kolenikov and I are up to something.

Probability sampling framework

The below framework/taxonomy gives a comprehensive understanding about the factors needed in sampling and for implementing the sampling design. This 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(s)	If one frame only is required to get sampling/sample units, it is called 'element sampling.' But if several frames are required to get those units, it is more complex, see Question B.
B. Stage	Hierarchy to approach to the study/survey 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.
C. 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. The number of phases is rarely more than two.
D. Stratification	The entire population is divided into several sub-populations, and the sample is drawn from each of them separately and independently. The inclusion probability of each stratum thus is equal to one. If the sampling design method of two strata is different, it is called **two-domain design.** Respectively, if the method differs in several strata.
E. Sample allocation into strata	How a desired gross sample has been shared into each stratum? Alternatives: equal, proportional, minimum, Neyman-Tschuprow. Anticipated response rates can be taken into account as well (by strata), see H.
F. 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).
G. Selection method It leads to the inclusion probabilities when sample size is decided.	How to select the study units - probability equal to all (simple random selection (SRS) equidistance, Bernoulli) or - probability varies unequally typically by size = probability proportional to size (PPS)
H. 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. The anticipation is good to do by strata if possible, but for the whole sample at minimum.

Sampling question
A. Frame(s)
B. Stage
C. Phase
D. Stratification
E. Sample allocation into strata
F. Panel vs. cross-sectional study
G. Selection method It leads to the inclusion probabilities when sample size is decided.
H. Missingness anticipation or prediction

This is just one task of the sampling design but often it is the only recognized

G. Selection method It leads to the inclusion probabilities when sample size is decided.	How to select the study units - probability equal to all (simple random selection (SRS), equidistance, Bernoulli) or - probability varies unequally typically by size = probability proportional to size (PPS)

This last one is not metioned often at all even might be most influential

H. Missingness anticipation or prediction	Trying to anticipate response and in-eligibility rates and allocate a gross sample so that the net sample is as optimum as possible in order to get as accurate results as possible. The anticipation is good to do by strata if possible, but for the whole sample at minimum.

Sampling and inclusion probabilities

We present in this sub-section most commonly used sampling selection methods and their inclusion probabilities. These are here presented either without missingness and so that in the case of missingness its mechanism is assumed to be ignorable, or *MARS*. 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. Respectively, we have the number of the respondents with symbol *r.* The formulas of this sub-section can be called 'design-based.'

Simple random sampling (*SRS*)

The inclusion probability for each *k* is constant. The second term here and later is for the selection probability

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

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

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

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

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

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*. The conversion to the respondents gives the same formula as in *SRS*, but the certain units *k* are missing randomly.

Equal inclusion probabilities: Each $k \varepsilon U$ have an equal inclusion probability via

How this is done in practice?

- The frame is in an electronic form: An appropriate software package is available with a random number generator.

- The frame is not in an electronic form but still random numbers:
- Create a list of the frame
- More or less manually technically
- Last birthday method
- Coordinates, GPS, Google map

Unequal inclusion probabilities

Just for clarification: the inclusion probabilities may vary by strata, quota or phase. This most common case is not considered in this sub-section but later in this chapter.

All methods demand one or more auxiliary variables to be used for the inclusion, thus from the sampling frame. In his sub-section such a variable is in some sense '*size*' that thus is correlated with the inclusion probability. The '*size*' variable is in most cases such that 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 cluster PSU's are available.

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

The size x_c is inserted in the inclusion and selection probability as follows

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

The subscript *c* refers to a cluster PSU 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. The PSU size *n* is thus decided separately. It is not usually needed to convert to the respondents, since the missing sampled PSU's are rarely accepted. The non-response thus occurs within these PSU's, concerning individuals.

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 does not exist. Of course, a good stratification might be a challenging target, but should still be tried. In the simplest case, even using proportional allocation 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 are these explicit strata in which $h = 1,\ldots,H$. How big H could be, it is not clear but the minimum is around 10. On the other hand, the maximum depends also about the achieved amount of the respondents in each stratum. It thus is necessary that each stratum will have enough respondents. It is not possible give a simple answer to the question 'What is enough?' since it depends on many things.

If the gross sample size is n_h then the inclusion probability when using simple random sampling within strata is

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

This method is also called **stratified (simple) 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 is enormous. After the fieldwork, when the counts of respondents are known in each stratum, the inclusion probability can straightforwardly computed:

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

This is maybe the most commonly used sampling weight

If r_h is zero or small, it is danger that the basic sampling weight is not plausible. This weight thus is the inverse of the inclusion probability of the respondents (assuming *MARS*)

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

.

Naturally, if $N_h = r_h = 1$, the basic weight is not problematic but it may be such, for instance, if N_h would be 1000 and $n_h = 10$.

Several countries use an address or dwelling register, then selecting the sampled addresses or dwellings, and then the 15+ years old individuals. This leads to varying inclusion probabilities at the second stage. The table illustrates the practical situation from a dwelling register. We see e.g. that the most common dwelling is such that consists of two 15+ years old persons. The variability leads to variation in the inclusion probabilities of that level even though the first probabilities are equal to one. The coefficient of the variation of these inclusion probabilities is 45.2 per cent that is fairly common in countries using this design. This leads to the approximate **design effect due to varying inclusion probabilities** at this stage *DEFFp* = $1 + 0.452^2$ = 1.204. *DEFFp* will be considered better in Chapter 4 but it is now enough to understand if all inclusion probabilities are equal, *DEFFp* = 1. This thus is the design effect of simple random sampling without missingness; if the inclusion probabilities are not equal, it automatically leads to a larger *DEFFp*.

Number of 15+years old individuals	Inclusion probability of the individual	Frequency	Percent
1	1.000	842	26.51
2	0.500	1595	50.22
3	0.333	423	13.32
4	0.250	246	9.25
5	0.200	49	2.20
6	0.167	15	0.47
7	0.143	3	0-09
8	0.125	2	0.06
9	0.111	1	0.03

Label	N	Mean	Minimum	Maximum	Coeff of Variation	Sum
Ideal inclusion probability	3176	0.000447	0.00006	0.00125	66.3	1.42
Realized inclusion probability	1573	0.000457	0.00006	0.00125	65.6	0.72
Ideal design weight	3176	3214.1	802.5	16689	59.3	10207848
Realized design weight	1573	3119.2	802.5	16689	58.9	4906545

The term ‘sampling weight’ or ‘sample weight’ or ‘survey weight’ is a general label for the weights used in estimating the target population parameters. These parameters can be of two types:

(i) They are concerned totals of the target populations or its sub-populations (domains), thus sums, amounts or quantities.
(ii) They are concerned means, medians, percentages or other relative parameters.

The sampling weights presented above work well in both these cases, but it is good to be careful with each software in order to know whether it works correctly. This is one reason to understand the other type of sampling weight, called ‘**analysis weight**’ or ‘**analytical weight**.’ This is obtained from the proper sampling weight w_k by dividing with the average of all weights of this target population. This weight thus is relative whereas the proper sampling weight indicates ‘amounts’ or ‘totals’.

$$w_k\_analysis = w_k / (\sum_r w_k / r) = w_k / \bar{w}_k = w_k r / \sum_r w_k$$

The analysis weight is used more than amount sampling weights since it is well enough in most analyses, but it thus does not work when amounts are of interest. For example,

https://books.google.com/ngrams/graph?content=Statistics%2CComputer%2CMathematics&year_start=1700&year_end=2000&corpus=15&smoothing=3&share=&direct_url=t1%3B%2CStatistics%3B%2Cc0%3B.t1%3B%2CComputer%3B%2Cc0%3B.t1%3B%2CMathematics%3B%2Cc0

The final inclusion probability of the multi-stage design is the product of all single-stage inclusion probabilities.

This thus is in theory fairly simple but not in real life necessarily. The reason is that one or more probabilities cannot be known for the sample units with missingness. The missingness is not common in the second stage if register or other information is available but in following stages it is common since it is needed to contact such units to get the information for the probability calculation. This thus means that this final inclusion probability can be calculated for the respondents only. The below example illustrates this problem with our test data.

Example 3.4 The weights of the 2012 PISA survey

The public PISA data is not similar as the respective ESS in all respects even though all methodological information can be found. The PISA survey instruction variables are complete for most countries that is not the case for the ESS. This will be discussed further in the analysis chapter 13. However, the sampling weights can be used in both. The main sampling weight of the ESS is analytical, thus their average in each country is one but the PISA weights are ordinary ones. Their sum thus is the target population of the country.

Country code 3-character	Weight	Mean	Minimum	Maximum	Coeff of Variation	Sum
United Arab Emirates	PISA STUDENT WEIGHT	3,5	1,0	22,7	71,4	40612
	ANALYSIS WEIGHT	1,0	0,3	6,4	71,4	11500
Australia	PISA STUDENT WEIGHT	17,3	1,2	67,7	58,3	250711
	ANALYSIS WEIGHT	1,0	0,1	3,9	58,3	14481
Austria	PISA STUDENT WEIGHT	17,3	1,0	81,5	36,9	82225
	ANALYSIS WEIGHT	1,0	0,1	4,7	36,9	4755
Belgium	PISA STUDENT WEIGHT	13,7	1,0	44,8	38,9	117889
	ANALYSIS WEIGHT	1,0	0,1	3,3	38,9	8597
Bulgaria	PISA STUDENT WEIGHT	10,3	4,8	27,0	20,5	54255
	ANALYSIS WEIGHT	1,0	0,5	2,6	20,5	5282
Brazil	PISA STUDENT WEIGHT	435,3	1,1	2837,7	72,4	2397036
	ANALYSIS WEIGHT	1,0	0,0	6,5	72,4	5506

	ANALYSIS WEIGHT	1,0	0,5	2,1	22,8	5001
Denmark	PISA STUDENT WEIGHT	8,8	1,0	77,5	81,7	65642
	ANALYSIS WEIGHT	1,0	0,1	8,8	81,7	7481
Spain	PISA STUDENT WEIGHT	36,4	2,5	144,7	101,1	370862
	ANALYSIS WEIGHT	1,0	0,1	4,0	101,1	10175
Estonia	PISA STUDENT WEIGHT	2,4	1,3	8,0	44,0	11627
	ANALYSIS WEIGHT	1,0	0,5	3,3	44,0	4779
Finland	PISA STUDENT WEIGHT	6,8	1,0	54,1	79,3	60047
	ANALYSIS WEIGHT	1,0	0,1	7,9	79,3	8829
France	PISA STUDENT WEIGHT	151,7	103,4	293,9	15,1	699779
	ANALYSIS WEIGHT	1,0	0,7	1,9	15,1	4613
United Kingdom	PISA STUDENT WEIGHT	138,5	47,6	457,9	31,8	579422
	ANALYSIS WEIGHT	1,0	0,3	3,3	31,8	4185
Hong Kong-China	PISA STUDENT WEIGHT	15,1	8,3	55,3	25,2	70636
	ANALYSIS WEIGHT	1,0	0,5	3,7	25,2	4670