Contributions from IASS Members

Sampling design data file

Seppo Laaksonen
University of Helsinki
E-mail: Seppo.Laaksonen@Helsinki.Fi

Keywords: Auxiliary variables, calibration, data quality, fieldwork, inclusion probability, non-response

Abstract: The paper first determines the term 'sampling design file' that is not commonly used in survey sampling literature. The methodology behind this term is, of course, used to some extent, but only 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, fieldwork and finally for estimation. The sampling design file consists of all the gross sample units and its variables include those that give opportunity to create sampling weights, to analyse the survey quality, and to estimate. 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 begin.

1. Introduction

The term 'sampling file' or more broadly 'sampling design data file' is rarely used in standard survey literature. One of this first users is the sampling expert panel of the European Social Survey (ESS) that was established in 2001 (see more information about this survey that initially started in 2002, europeansocialsurvey.org). The document of the panel says: "The Sampling design data file **(**SDDF) is routinely generated by an ESS country's National Coordinator after fieldwork has finished. It includes information on the implemented sample design such as inclusion probabilities and clustering. As such, it serves the sampling team with the data required for computation of design weights, design effects and as a general basis for benchmarking the quality of sampling. The ESS analyst may use it for several purposes such as incorporating cluster information in her/his analyses."

A SDDF is required for all types of surveys, thus for surveys from households, individuals, businesses and corporations. Here we concentrate on surveys of individuals who are members of households.

2. Basic targets of sampling file

The statistical units of the sampling file should ideally cover all the gross sample units of the survey. Such units are selected addresses in the case of address-based samples (but there are individuals behind these addresses or dwelling units), and selected individuals in the case of individual-based samples. In the end, the file of these statistical units thus covers the respondents, the non-respondents and the in-eligibles. It might be difficult to completely numerate in-eligibles for the file, since any contact for some individuals/addresses cannot be made and hence the file may be inaccurate, but all efforts to complete the file with appropriate information should be done. It follows that such a unit may thus be either an in-eligible or a non-respondent. Correspondingly, some bias in estimates necessarily follows.

The *first-order sampling file* is good to create while the gross sample has been drawn. In this case, the file includes:

- Non-confidential and confidential identifier
- Sampling frame variables and respective statistics
- Stratification variables, explicit strata in particular
- Implicit strata if they include useful information; implicit stratification specifies the order of the units selected by equidistance or other systematic selection but basically this design corresponds to a simple random selection
- Inclusion probabilities of each stage within explicit strata.

In the case of multi-stage sampling, all inclusion probabilities may not be available before the end of the fieldwork. This is typical in a three-stage sampling if the primary sampling units (PSU) are small-areas and the secondary sampling units (SSU), respectively, are addresses or households, but the third stage units are individuals. This missingness for the third stage units is due to the problem of contacting a dwelling unit or an address in order to know how many target population members there exists. Even in register-based countries such information is hard to get correctly, since the register is not up-to-date for a survey period.

It is possible and also useful to calculate the gross sample design weights immediately when the first-order file is available. This gives opportunity to check basic figures and the quality of the sampled file of this phase. For example, when summing up these design weights we should obtain the correct target population statistics that represent the final target population if no missingness occurs. In contrast, if the third-stage units, for instance, are missing, the target population of the households or addresses can only be computed.

The above variables derived from a sampling frame are minimal requirements but not sufficient. It is rational at the same occasion to download other useful information for the sampling file from the sample frame that we call here the *second-order sampling file*. For example, in register-based countries, the sampling frame has been created from the population register, that is reasonably up-to-date. The sampling design only requires aggregate population statistics by large region, age group and gender, for example. But the same information can be matched at micro level into gross sample units too. In addition, the same data source consists of many other variables that are beneficial to download to the second-order sampling file at the same time since it is basically free of charge. It is not common even in Finland to distend over the minimum although it is possible to expand the file with the following auxiliary variables, among others: marital status, year of marriage, multi-marriage, number of children, house size, type of house, citizenship, mother tongue, coordinates of the house and municipality at birth.

The second-order sampling file can further be completed from other sources at the same time as the first-order file has been created. This usually may require some additional administration and paper work but it is best to do as soon as possible since the data sources cannot be up-to-date for long, or even some data are destroyed. Typical other sources are: formal education, tax register information on income and wealth, jobseekers' register. Section 4 presents a Finnish example on this issue in more details.

The *third-order sampling file* can be created as soon as the fieldwork has been completed. In this case, the most important new variable is the outcome of the fieldwork that indicates who is a unit respondent, and who is a non-respondent and an in-eligible, respectively. As said above, the last two categories are often hard to definitely determine with accuracy. This seems to be a worsening problem in Europe due to more or less permanent absence of the official address (home). A reason for this is working outside the country over several months, or using a second home in another country, respectively.

A drawback, in many countries as said already above, is that all inclusion probabilities cannot be known after the fieldwork. In the ESS, the selection of one individual within the selected household or address is a good example. As a consequence, it is not possible to calculate a complete inclusion probability for the individuals of the gross sample, but only for the second stage address/household. The sampling weight for the respondent can be, nevertheless, calculated, assuming, for example, that the response mechanism for the third stage is ignorable within strata.

After the fieldwork, the sampling file can be further reinforced with other data on fieldwork. Opportunities for that are dependent also on the survey mode used. Face-to-face interviewers can collect information about the quality of the location where a potential respondent lives. For example, an interviewer can classify the quality of the living area or the type of house. This indicator is of course useful only if a valid measurement is available and the same information is available both for the respondents and for the non-respondents. Moreover, the interviewer information (e.g. their basic characteristics, attitudes toward this survey) can be added to the sampling file too.

3. What to do with sampling file?

The sampling file is necessary in order to calculate the sampling weights for the respondents. This requires that the inclusion probabilities are available in the file. Naturally, the identifiers of the respondents should be available in the sampling file in order to match the sampling weights and other sampling design variables into the survey data file of the respondents.

The narrowest correct sampling file is such that the sampling design is simple random sampling. In this case, the file consists only of an identifier and one constant inclusion probability, and the survey outcome variable that identifies the respondents, the non-respondents and the in-eligibles. These data allows the calculation of a single sampling weight for each respondent. No real non-response analysis can be done due to completely missing auxiliary data.

If a two- or three-stage design has been used, there are more variables, including PSU's as clusters, and SSU's, respectively. Even though there are no other strata or auxiliary variables, it is possible to review non-response by PSU and SSU, respectively. This gives the opportunity to adjust the weights to some extent since non-response may vary by SSU conditional to PSU. Hopefully, all PSU's are still in the file. Otherwise, the fieldwork has failed.

The sampling file is primarily needed to create the weights for the respondents although it is best to first create weights for the gross sample. Secondarily, the file is for analysing the success of the fieldwork. It is possible that a particular survey may use more than one survey mode, like in the case of a mixed-mode design. The sampling file naturally must include the mode used in data collection for all individuals. If two or more modes are used for one individual, this should be coded at variable level as well.

A good sampling file is naturally very useful to analyse survey quality. Auxiliary variables particularly are needed for this purpose. Also, we would be happy if some variables of the fieldwork file would be merged with the sampling file.

4. Auxiliary variables in the sampling file

We have above given examples of auxiliary variables of a good sampling file. Now, we concretise this issue. It is good to recognize that all such variables are given for individual gross sample units whatever they are. In the case of multi-stage sampling, such variables can be more problematic since they are first concerned with clusters of the target units. There can thus only be such variables that are related to clusters. If the clusters are small-areas, regional information is available. However, it is more difficult to know, for example, about the education of all cluster persons. This may not be necessary since it is more important to gather information about the education of the respondents and the non-respondents within this cluster.

Auxiliary variables can thus be either **macro** or **micro**. Both of these variables are useful and even for the same purpose; they can be derived from the same basis. For example, age of an individual can be used in non-response analysis in several forms, such as individual ages or as groups. However, the same variable is useful as target population statistics and thus a macro auxiliary variable would indicate how many target population members are in each age group. This is an example of the benchmarking information, and they can be used in calibration methods that require macro auxiliary data, that is, known population margins (e.g. Deville and Särndal 1992). There can be several population margins in calibration at the same time. And if such information is available in the sampling file, it is easy to compute the calibrated weights, respectively, using the French software Calmar 2, among others (see Le Guenne & Sautory 2005).

Macro auxiliary variables can thus be margins of known population figures giving opportunity to use these in calibration. They can also be relative frequencies of small areas like PSU's, concerning for instance register unemployment rates, rates of highly educated people, or crime and poverty rates. Such variables could be used for analysing reasons of nonresponse.

The richness of the auxiliary variables in the sampling file facilitates in analysing the success of the fieldwork. For example, unit non-response can be assessed against these variables and the multivariate response propensity model estimated as a result. This model may respectively be a good starting point for adjusting the sampling weights to take into account the variation in non-response (e.g. Laaksonen 2007, Laaksonen and Heiskanen 2013)).

The sampling file should be explicitly available, that is, for all gross sample units, and all inclusion probabilities should be in the file. Sometimes, these probabilities are only implicitly available. For simple random sampling it is most common since the inclusion probabilities are unique and only requires one population statistics figure and the gross sample size. Hence it is impossible to check based on this probability that everything has been done correctly. Thus, is SRS really good or not?

Another difficult situation is two-stage sampling when the equal absolute sample sizes are used in the second stage. This leads to final inclusion probabilities in which the PSU sizes of the first stage clusters will disappear. It means that this size is not necessarily needed in the formula of the inclusion probability. Unfortunately, there exists sampling files where there is only one 'final' probability of this kind. One example is in Burnham et al (2006) that Laaksonen (2008) criticised due to missing inclusion probabilities and in particular that all concrete information about first stage sampling is missing. So, it is possible that everything has not been done correctly since sampling design information is lacking. This is true for all designs, even in simple random sampling, since the sampling data file is so restricted that very little can be checked. Good auxiliary data (macro and micro) also lever confidence in the survey data and hence it should be recommended to collect.

5. A Finnish example

In the end, an example from the Finnish security survey (FSS) 2010 is presented (Aromaa 2010, Laaksonen and Heiskanen 2014). The characteristics of its sampling data file are given below.

The number of statistical units of the FSS is 7933. They are thus gross sample units. Table 1 illustrates the variables of the sampling file.

This list is rather long, and good in many meanings. It gave opportunity to analyse non-response by various auxiliary variables. Based on the data, we also created the so-called adjusted sampling weights. This first exploits the response propensity modelling and finally the stratification based on such calibration that the known population statistics match with our gross sample design weights by strata.

Naturally, we used the data also for survey quality including the analysis of problems in the fieldwork. This was possible for two reasons: (i) based on paradata, we were able to follow the interviewing time that was shortening during the fieldwork; the response time vary by mode as well so that web took least time and face-to-face the most time, (ii) we made a special survey for the interviewers and found that the point (i) was in telephone interviewing due to the busy call schedule at the end of the fieldwork. Naturally, the results were not ideal.

Our sampling file thus is rich but it is not common everywhere. The file content also depends on the survey practice. Our European Social Survey team has found various interesting contents that should be included in the sampling file. One is the so-called reserve sample that is initially created to guarantee that enough respondents will ultimately be found. It is clear that this reserve should be probability based, but if the reserve part is not included in the sampling file, it will be hard to follow the fieldwork well and even to calculate correct response rates. This reserve sample option is now in our template. It is interesting that a certain country found this option in our sampling file and incorrectly wanted to take a reserve sample even though this was not in their sampling file.

6. End notes

I sincerely hope that survey organisers will pay attention to create as good a sampling data file as possible and such that it would help in getting improving estimates from the survey. Unfortunately, this concept is not currently in standard literature. Hopefully it will be so in a future.

References

Burnham, G., Lafta, R, Doocy, S. and Roberts, L. (2006) Mortality After the 2003 Invasion of Iraq: a Cross-sectional Cluster Sample Survey. *The Lancet* **368,** 1421–1428.

Deville, J-C. & Särndal, C-E. (1992) Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 376-382.

Laaksonen, S. (2007) Weighting for Two-Phase Surveyed Data. *Survey Methodology,* December Vol. 33, No. 2, pp. 121-130, Statistics Canada.

Laaksonen, S. (2008) Retrospective Two-Stage Cluster Sampling for Mortality in Iraq. *International Journal of Market Research* 50, 3, 403-417.

Laaksonen, S. and Heiskanen, M. (2014). Comparison of Three Modes for a Crime Victimization Survey, *Journal of Survey Statistics and Methodology 2 (4): 459-483 doi:10.1093/jssam/smu018*

Le Guenne, J. & Sautory, O. (2005) CALMAR 2 : Une Nouvelle Version de la Macro Calma de Redressment D'Échantillion Par Calage. http://vserver-insee.nexen.net/jms2005/site/files/documents/2005/327_1-JMS2002_SESSION1_LE-GUENNEC-SAUTORY_CALMAR-2_ACTES.PDF

***Table 1.* The key variables of the sampling data file for the Finnish Security Survey.** Symbols in the column 'Source': SO = Created by survey organisation, M = Computed by methodologist, PR = Population Register, FER = Formal Education Register, ER = Employment Register, TR = Tax Register. The alternatives for use: R = Merging with respondent data, S = Sampling, U = Unit non-response, W = Weighting, E = estimation

Variable	**Source**	**Use for**
Identifier	SO	R
Survey mode (face-to-face, telephone or web)	SO	W E
32 Explicit strata, anonymous code	M	S U W E
PSU's, anonymous code	M	S U E
PSU size, Stratum size (incl. size of the target population)	M	S E
1st stage inclusion probabilities for PSU's and 2nd stage probabilities for households, 3rd stage probabilities for individuals	M	S W E
Age in years Age group respectively, both micro codes and macro statistics	PR	S U W
Gender, code and macro statistics	PR	S U W
Regional variables including municipality, postal code, co-ordinates of home, code and for some also macro	PR	S U W
Marital status with different options, year of marriage, number of marriages, code	PR	U W
Native language and citizenship	PR	U W
Occupational or socio-economic status (fairly rough only available)	PR	U
Household composition including number of children at different age groups	PR	U W
House variables such as size, number of rooms and type of kitchen	PR	U W
Level and field of education	FER	U W
Unemployed or not, number of months unemployed	ER	U W
Taxable income	TR	U W
Fieldwork outcome (respondent, non-respondent, in-eligible)	SO	M U W E
Neighbourhood variables in face-to-face surveys (option)	SO	U W
Reserve sample indicator if used, responsive design indicator respectively	SO	U W
Reason for non-response (well for face-to-face, badly for web)	SO	U
Para data, e.g. interviewing time, responding time	SO	E

Upcoming Conferences and Workshops

RIO
DE JANEIRO
iSi 2015
60th World Statistics Congress
26-31 July 2015
60th World Statistics Congress – ISI2015

<u>60th ISI World Statistics Congress</u>

Organized by:	International Statistical Institute
Where:	Rio de Janeiro, Brazil
When:	26.07.2015 to 31.07.2015
Homepage:	http://www.isi2015.org

We are delighted to invite you to the 60th ISI World Statistics Congress (WSC), which will take place in Rio de Janeiro, Brazil, during 26–31 July 2015.

The WSC is the flagship conference of the International Statistical Institute (ISI) and its seven associations. It is a biennial conference with a rich tradition, and IBGE is pleased to host and organize ISI2015 in Brazil.

The congress will bring together members of the statistical community to present, discuss, promote and disseminate research and best practice in every field of Statistics and its applications. The Scientific Programme of the 1512015 will include a wealth of activities that will cover stimulating topics and will offer delegates innovative and well-balanced presentations, as well as plenty of opportunities for discussion and exchange.

A rich and exciting Social Programme is also being developed, with plenty to see and enjoy for participants and their accompanying persons, hoping to make your trip to Rio and taking part in ISI2015 a truly unforgettable experience.

The venue - Riocentro - is located in Barra da Tijuca, a district surrounded by natural beauty but also many sophisticated bars, restaurants and several malls and close to a variety of historical and cultural programs that only the Wonderful City can offer.

We are confident that all the ingredients are in place to ensure that the 60th ISI World Statistics Congress will be a memorable statistical event!

For further information please email Francisco Samaniego fjsamaniego@ucdavis.edu