Seppo Laaksonen

Survey Methodology and Missing Data

Tools and Techniques for Practitioners

Springer

Survey Methodology and Missing Data

Seppo Laaksonen

Survey Methodology and Missing Data

Tools and Techniques for Practitioners

Springer

Seppo Laaksonen
Social Research, Statistics
University of Helsinki
Helsinki, Finland

ISBN 978-3-319-79010-7 ISBN 978-3-319-79011-4 (eBook)
https://doi.org/10.1007/978-3-319-79011-4

Library of Congress Control Number: 2018939028

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This book is a summary of my long and extensive research experiences in many forums, of which the following are probably the most important: Statistics Finland, the statistical office of the European Commission (Eurostat); the Social and Health Research Institute of Finland (Stakes); the University of Southampton; and the University of Helsinki.

At the same time, I have participated in many international networks and projects and have acted as a consultant on statistical issues in several places—most recently in Ethiopia—as noted in a few of the book's comments. One of the important networks is the Household Survey Nonresponse Network, which was founded in Stockholm in 1990 and continues to meet every year. I have often participated in these meetings. The other long-term group that I have been involved in is the expert sampling team of the European Social Survey (ESS); I have been a member since 2001. This book very much focusses on the ESS, which is a good framework for all social surveys, and I recommend that its standards are followed. Another useful network is the Programme for International Student Assessment (PISA). I was a member of the Finnish team that was responsible for the 2006 PISA survey. Thus, I learned quite a lot about it, and since then always have used the PISA data in my teaching and, to some extent, in my research.

For approximately 10 years around the year 2000, I was a Finnish coordinator of several European Union (EU) research projects, many concerning survey methods. In this book, the imputation methodology is based directly on two EU projects: Automatic Imputation Methods (AUTIMP) and Development and Evaluation of New Methods for Editing and Imputation (EUREDIT). The main data in Chap. 12 are derived from the second project, although greatly modified in the examples.

The impact of the EU's project, Data Quality of Complex Surveys within the New European Information Society (DACSEIS), on variance estimation also can be seen in this book, but not as explicitly. Later, around 2010, I was a member of the Finnish team that collated a representative data file on security and crime victimisation. We tested three survey modes, which is not a common approach. The findings were exciting, some of which are included here too.

I have presented short and long survey courses in all the institutions at which I have worked, and in other places as well, although the first comparatively complete version of this methodology was written in 1999 while I was at Stakes. It was the

basis for the first course I taught at Helsinki University in 2002. Since then I have given survey courses there in various forms, including a recent one that covered all the main survey topics. Some of them, however, were at only a rather general level because it was anticipated that certain methods would be too difficult for students with not much background in statistical and/or other quantitative methods.

As a result, I also have given specialised courses on, for example, 'Advanced weighting and reweighting methods', 'Editing', 'Imputation methods' and 'Statistical disclosure limitation methods'. Course participants had a good background in statistical methods and informatics. The structure of this book is derived mostly from these experiences. That is, some parts are expected to be relatively easy for most social scientists, but the more sophisticated parts will be demanding for them, and for specialists as well, because the courses also include new methodologies.

Given that most students in general survey methodology courses have not been statisticians or mathematicians but have tended to be, for example, sociologists, social psychologists, economists, psychologists, demographers, geographers, and political scientists, the course is not formula-focussed. I have found that this can help when attempting to understand the main aspects of surveys. Another goal has been to ensure that students learn survey methodology well at a general level so that they are able to ask advice from specialists as early and knowledgeably as possible.

Often, I have found that a beginner may collect data but forget many important things when doing so. Some eventually understand that they need to contact a specialist for help. Unfortunately, the specialist then finds it difficult to make useful improvements because the mistakes made in the early stages frequently are fatal. This book should help to avoid such awkward situations, even if the reader does not completely understand everything in it. I believe that empirical, real data examples help to establish a basic understanding, if a participant has been paying attention sufficiently, before he or she started a survey. In addition, the students of my courses have done some of their own data handling and have reported on their outcomes. This book can be used for these purposes too.

Some years ago, I decided to write a full survey methodology book in Finnish. Its first version was published as an open access e-book by a Danish publisher, Ventus Publishing, in 2010. A new edition was issued in 2013. These books have helped with my teaching a great deal, although it was soon recommended that I use English more in teaching. When Springer contacted me in 2015, I suggested that I should write an English version as well. I have tested this version twice with my recent Helsinki University students. They have very much inspired my work, and some corrections have been made as a result. The comments of anonymous reviewers naturally have been taken into account as well.

I hope that readers will be pleased with my approach to survey methodology, with its focus on handling missing data. The history of sample surveys is not long, the first being implemented in the 1930s. Their use began to expand after the Second World War in the developed countries. Missing data at that time was not a great problem, but this issue has become increasingly worse as time has gone on.

This has led to the development of new research strategies and methods, many of which have been successful. Unfortunately, the strategies have not been able to solve

everything. In particular, it is still difficult to obtain sufficiently accurate information about marginal groups. This is a nuisance that is evident in several examples in this book, the goal of which is to describe the newest methodologies for handling missing data. My focus here is more on post-survey adjustments; nevertheless, all the planning and fieldwork for surveys are just as important.

Helsinki, Finland Seppo Laaksonen

Contents

Introduction

1

From the start road some steps forward

This textbook is on quantitative survey methodology, and it is written in such a way that survey beginners will be able to follow most of it. Nevertheless, they are expected to have some background knowledge about statistics and related issues, and to be interested in learning more about survey methods and practices. What is covered in this book is extensive. It includes fields such as advanced weighting, editing, and imputation, that are not covered well in corresponding survey books (cf. Bethlehem, 2009; Biemer & Lyberg, 2003; De Leeuw, Hox, & Dillman, 2008; Gideon, 2012; Groves et al., 2009; Valliant, Dever, & Kreuter, 2013; Wolf, Joye, Smith, & Fu, 2016). These subjects, naturally, are covered in specialized books and articles—see Chaps. 8, 11, and 12. On the other hand, we do not give much consideration here to statistical tools and methods relating to limitations caused by confidentiality, which are important in practice.

To help the reader without advanced statistical knowledge, we do not use many statistical formulas, but the necessary formulas are still included. This is possible

S. Laaksonen, *Survey Methodology and Missing Data*,
https://doi.org/10.1007/978-3-319-79011-4_1

because it does not take much space to explain the basic ideas in formulas. All of the more detailed and important formulas, however, can be found in the References included in this book. They are cited at the end of each chapter, but the Appendix entitled 'Further Reading' contains other bibliographical references on surveys in journals, books, and articles.

This book also will be useful for more experienced or even sophisticated users because some parts include methodologies that are not generally known even among survey experts. This is very much a result of the book's focus, which is on dealing with missing data. This focus is the result of survey practice, which has been becoming more awkward just because various types of missing data problems have been getting worse during recent decades. At the same time, new tools have been developed. Some of these tools and technologies are valuable, it is clear, but some are not, or their quality is not known. If the recommendations in this book are followed and implemented in a survey process, the outcome definitely will be good, or at least its quality will be known.

It is important to recognize that the book must be considered as a whole. In particular, this means that it will be beneficial to learn the terms used in each chapter, although Chap. 2 includes their core pattern. New terms are introduced later; thus, it might be difficult to go directly to the most sophisticated chapters—Chaps. 8 and 12. It should be noted that these authors' terms are not new, as they have been used earlier; however, terms seem to vary to some extent from one source to the next. It therefore would be useful to understand the terms in this book in order to understand its key points.

Survey Methodology and Missing Data presents many empirical examples that are, in most cases, from real surveys, particularly multinational ones. Two such surveys are used most generally. The first is the European Social Survey (ESS), which is an academically driven biannual survey (europeansocialsurvey.org). Its initial round was conducted in late 2002 or early 2003. The other survey, which has often been applied, is the Organization for Economic Cooperation and Development's (OECD) Program for International Student Assessment (PISA) that has been conducted every three years since 2000. The micro data from both surveys are publicly available and consequently are easy to use around the world, but they require adequate knowledge of survey methodology. We use much the same variables in examples throughout the book, which should help readers follow the methodology. Many examples have not been published elsewhere, as far as we know. The authors think that some results are also interesting from the subject matter point of view (e.g., Chap. 14).

Using examples just from one survey, either the ESS or the PISA, would not be reasonable in this textbook because these surveys are different. The ESS is used most commonly in the book's first part and the PISA is used toward the end, where we pay the most attention to survey analysis. The reason for this is that the public ESS files only include the survey weights, not two other important 'instruments'—stratum and cluster. Consequently, we cannot use the ESS in many examples, thus, the PISA is chosen when explaining the importance of other the instruments.

We are able to use examples that require the sampling data, given that the author had access to this material. Unfortunately, we cannot publish the related data file because it is confidential. We have passed over this problem by creating an artificial file that consists of two of the domains that have been used in almost all ESS countries. We include the description of this micro file in Chap. 6.

We use some additional survey data too. The Finnish Security Survey is used for two reasons: (1) some of its questions are special, and (2) it is based on three survey modes, which is not usual. We also use another special data file for imputation because neither the ESS nor the PISA survey is illustrative of this methodology. There are a few small-scale examples outside these main data files. In general, the examples are from social or human surveys, but some comparisons to business surveys are given.

An understanding of surveys and survey terms is demanding and takes time to gain such knowledge. It is not possible to write a book in a way that all terms will be immediately understood when they are explained. The authors use many 'graphical' and other schemes that we hope will facilitate understanding. The empirical examples are, then, for deeper comprehension. It would be beneficial if a user could use real data at the same time. This is possible in most cases since the ESS and the PISA files are publicly available.

Each survey should be conducted in the best way. This book is focused on the methods that help with the process so that the outcome is of as high a quality as possible. Second, the book's purpose is to give reasonable starting tools for using and analyzing a file once a survey has been conducted. This file should be reasonably well cleansed. The core of the book thus is focused on survey data collection and cleaning methods, which cover the following key steps:

- Designing the survey, which includes determining the target population
- Designing the questionnaire
- Designing the sample or samples
- Processing the fieldwork so that the data collection is productive, and the quality is high
- Entering data as much as possible during the fieldwork, or automatically
- Editing the raw data as much as possible during the fieldwork, with final editing afterward
- Inputing missing and implausible values, if this is believed will improve the results
- Including relevant auxiliary variables for the sampling design data file
- Creating the sampling design data file during and after the fieldwork
- Analysing nonresponse and other gaps found
- Weighting the micro data using the most advanced methods and the best possible auxiliary variables available and/or were gathered
- Documenting everything during the process, in a digital form to the extent that this is possible
- Adding other features into the data file that will help a user analyse it as easily as possible.

After creating the well cleaned survey micro file, it is possible to begin survey data analysis. Chapter 14 covers the basic methods for correctly performing the analysis so that the survey data 'instruments' are considered. More demanding survey analysis is not covered in this book.

The penultimate chapter, Chap. 13, contains a summary of all the key terms in the book. We recommend that readers look at this chapter from time to time, maybe after finishing each of the other chapters. It might be a good idea to read this whole chapter quite early, even though many of the concepts would not yet be understood.

The role of the photos at the beginning of each chapter are meant to give an image of the survey terms at a general level. They offer an opportunity to take a break between chapters and to start the next one without any prejudices. All the photos are from nature (not just from Finland) and are without people; they were taken by the author.

References

Bethlehem, J. (2009). *Applied survey methods: A statistical perspective* (p. 392). Hoboken: Wiley Survey Research Methods & Sampling.

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality*. Hoboken: Wiley.

De Leeuw, E., Hox, J., & Dillman, D. (2008). International handbook of survey methodology. Accessed October 2016, from http://joophox.net/papers/SurveyHandbookCRC.pdf

Gideon, L. (2012). *Handbook of survey methodology for the social sciences*. New York: Springer.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Hoboken: Wiley.

Valliant, R., Dever, J. A., & Kreuter, F. (2013). *Practical tools for designing and weighting survey samples*. New York: Springer.

Wolf, C., Joye, D., Smith, T. W., & Fu, Y.-C. (2016). *The SAGE handbook of survey methodology*. Los Angeles: Sage.

Concept of Survey and Key Survey Terms

2

Forest Mode and Garden Mode

2.1 What Is a Survey?

We determine the survey in its relatively short form, as follows, but it can be defined in many other forms as well (Laaksonen, 2012):

A survey is a methodology and a practical tool used to collect, handle, and analyse information from individuals in a systematic way. These individuals, or micro units, can be of various types (e.g., people, households, hospitals, schools, businesses, or other corporations). The units of a survey can be simultaneously available from two or more levels, such as from households and their members.

Information in surveys may be concerned with various topics such as people's personal characteristics, their behaviour, health, salary, attitudes and opinions, incomes, impoverishment, housing environments, or the characteristics and performance of businesses. Survey research is unavoidably interdisciplinary, although the role of statistics is extremely influential because the data for surveys is constructed in a quantitative form. Correspondingly, many survey methods are special statistical

S. Laaksonen, *Survey Methodology and Missing Data*,
https://doi.org/10.1007/978-3-319-79011-4_2

applications. Nevertheless, surveys substantially utilize many other sciences such as informatics, mathematics, cognitive psychology, and theories of subject-matter sciences of each survey topic.

A survey is a series of tasks that finally results in a statistical file of numerical units and their characteristics (variables); the units may be:

- Individual people
- Households and dwelling units ('register households' in registered countries)
- Families
- Schools and other public institutions
- Enterprises
- Plants (local units of enterprises)
- Local kinds-of-activity units of enterprises
- Villages, municipalities, and other administrations
- Other areas, including grid squares
- Societies, associations, and corporations

Such a data file may cover basically the entire desired population, or it can be based on a sample (i.e., the terms 'survey sampling' and 'survey statistics' are used, respectively). A survey often is considered to be only a sample survey, but similar methodologies and tools can be applied if a register or another administrative file has been created and handled.

Therefore, the authors use a relatively broad definition for surveys here; however, it is important to recognize that a final survey file may be a combination of various data collections, and several organizations may have participated in conducting it. Naturally, we concentrate mainly on surveys and their methods when collected and handled by one survey organization.

This chapter focusses on determining precisely five cross-sectional populations in surveys. We briefly continue to longitudinal and panel surveys, but our focus in this book is on cross-sectional studies. At the same time, we present other terms needed in later chapters. At this stage, they will not yet be understood thoroughly.

2.2 Five Populations in Surveys

A *population* is a key concept of statistics, as determined by Adolphe Quetelet in the 1820s. It is not just one population in surveys where we need even five. In addition, before the first one, *a target group* that the survey will be concerned with is in mind. That group is usually rather rough, but it may be close to one or more of the following five populations:

1. *Population of interest* is the population that a user would like to get or estimate ideally, but it is not always possible to completely reach; consequently, the researcher determines the second population—target.

2. *Target population* is a population that is realistic. Naturally, it should be exactly determined, including its reference period (i.e., a point in time or a time period).

The following are examples of target populations used in this book. We do not mention any year because it varies in the first two cases.

- The European Social Survey (ESS): 'Persons 15 years or older who are residents within private households in the country on the 1st of November'.
- The European Finnish Security Survey (EFSS): Non-Swedish-speaking 15–74-year-old residents in Finland on the 1st of October.
- The Programme for International Student Assessment (PISA) survey: 15-year-old school students (i.e., specified so that the full calendar year is covered).
- The grid-based study of Finland: People from 25–74 years of age living in south Finland.

► **Discussion of the First Two Populations** We think that first it is better to try to find an ideal population (i.e., a population of interest). This is not possible in most cases. For instance, if the target is to get the voting population, it is not possible. Therefore, the population that is eligible to vote is reasonable, but it can be difficult to achieve as well because the survey institute may be using telephone interviewing with random digit dialing. Thus, it is not known in advance who will be willing to participate (e.g., vote). A good point is that people who do not participate in this survey do not vote well either. Consequently, the quality of such surveys is often satisfactory.

3. *Frame population and the frame* from which the statistical units for the survey can be found. Usually, the frame is not exactly from the same period as data from the target population. The delay in population surveys is rather short (i.e. 1–5 months), but enterprise surveys take much longer, even years. This frame is not always at the element level available, as in the case of the central population register-based surveys. Instead, the frame population can be created from several frames (multiframe), often from three—not all of which may be available when starting the survey fieldwork. Yet, the first frame is necessary to be able to begin. This often consists of the regions, or the areas, or the schools, or the addresses, but later other frames are needed. Fortunately, only the ones of those who were selected in the first stage.

A Multiframe Example

- *Stage 1:* List of the electoral sections (this number might be thousands).
- *Stage 2:* Lists of all household addresses of the units selected during the first stage. The address might be complex because there can be more than one dwelling at one address, and more than one household in one dwelling.
- *Stage 3:* One or more members at the selected household and/or address.

We thus observe that the Stage 1 frame should be available centrally, but the other frames are needed only for those units that are selected. This means that these frames need to be created at this stage. Sometimes this is possible using a local population register; however, from time to time it is created by the survey organization.

► **Comment** To get a realistic target population all important targets are not achieved, but because survey quality is crucial, a certain optimum is good to keep in mind. Respectively, it is advantageous to optimize the target population from the point of view of the availability of the frame or the frames. If the quality of the frames is inferior, the target population might be difficult to determine.

Any frame population is not completely up to date. Fortunately, when the survey fieldwork and the data collection are done some months later, it is possible to get a new frame, which is the fourth population.

4. *Updated frame population* is useful for better estimating the results. Usually, the initial frame population has been used for estimation too. This may lead to biased estimates. Fortunately, this bias is not severe in most human surveys. By contrast, old frames can lead to dramatic biases in business surveys should these concern large business organizations.

Bias

A bias is a systematic error. It is not random like a sampling error. A biased estimate is thus systematically inaccurate. There can be several reasons behind it, due either to an incorrect estimator or more likely to problems in the data.

Finally, we will have the fifth population when we also know how much the fieldwork has succeeded.

5. Survey population or study population. It is ideal if this fifth population corresponds to the target population or even the population of interest. If not, however, the estimates are somewhat biased.

If there are clear gaps in the final data, this should be made known to the users (i.e., how much the survey population differs from the target population). This might be problematic to know exactly, but the main problems should not be difficult to identify.

2.3 The Purpose of Populations

Before continuing with survey terms, it is useful to discuss the purpose of these populations. Naturally, the first point is to approach to the targets of the survey as well as possible, so it is necessary to know all the steps and possible gaps passed or hopefully solved.

The final target is to estimate the desired estimates, such as *averages, standard deviations, medians, distributions, ratios, and statistical model parameters*. This can be done by just calculating in of any kind of way, but such figures cannot be generalized at any population level without using the survey instruments that are explained in this book. If all coverage and related problems are solved, the results can be *generalized at the target population level*. These results are called *point estimates*. This means that they are not 'true values' as in the case of the entire target population without sampling or missingness gaps.

To better understand the quality of these estimates, it is necessary to estimate their uncertainty as well. Indicators for uncertainty are standard errors, confidence intervals (margins of error), and *p*-values, among others. Standard software programs give such figures but it cannot be guaranteed that they are correct unless survey instruments are applied (see Chap. 14).

If this population cannot be achieved satisfactorily, it is best to talk about generalization at *the survey population level*. It is not common to report the surveys in this way, although the reality is that certain groups are not really represented among the respondents. For example, homeless, disabled, and other marginalized people who do not understand the language used in the survey are not well represented in most surveys. It is possible to attempt a generalization in another way, for example using modelling, but this issue is special and cannot be considered in this book. This generalization mainly is concerned with certain connections or explanations found in the data. Thus, it is possible to try to generalize such 'estimates' or other outcomes in some way.

The units of the target population are equal to those of the survey population, but the units of the frame population(s) can be essentially different, except in element-based sampling. The ESS survey designs vary a lot from one country to the next. There are countries where all the units are equal to individuals age 15 and older (i.e., in register countries such as Sweden and Denmark) but many countries have several units (e.g., small areas, addresses, dwellings, individuals 15+ years old).

PISA and other student surveys typically use two units: (1) 'PISA' schools (or school classes) and (2) students themselves who are needed from those classes sampled.

2.4 Cross-Sectional Survey Micro Data

We next present three schemes to illustrate the nature of the cross-sectional survey data. The first scheme is the simplest and is never found in practice (Scheme 2.1). Nevertheless, it is good to consider because it starts by presenting the concepts and symbols used in this book.

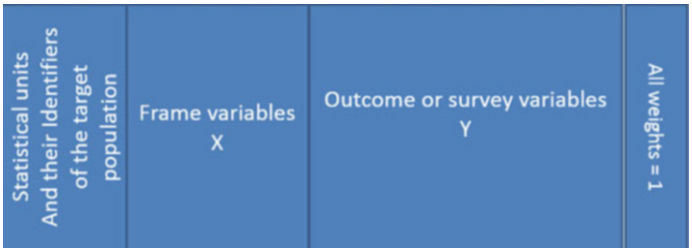
If the whole target population has been examined and no missingness's occur, there are four groups of concepts:

1. Statistical units that are often identifiers in surveys. Our general symbol of the statistical unit is k. The identifiers are of two types:
   - Ones known to be needed in survey institutions for several purposes.
   - Anonymous ones that are given for outsiders in order to protect the individuals
2. Frame variables X or x that are used in collecting the data.
3. Outcome or survey variables Y or y that are obtained by the fieldwork.
4. The sampling weights that are all equal to one because all are included in the survey and all are replied to. Such weights are not needed in the analysis.

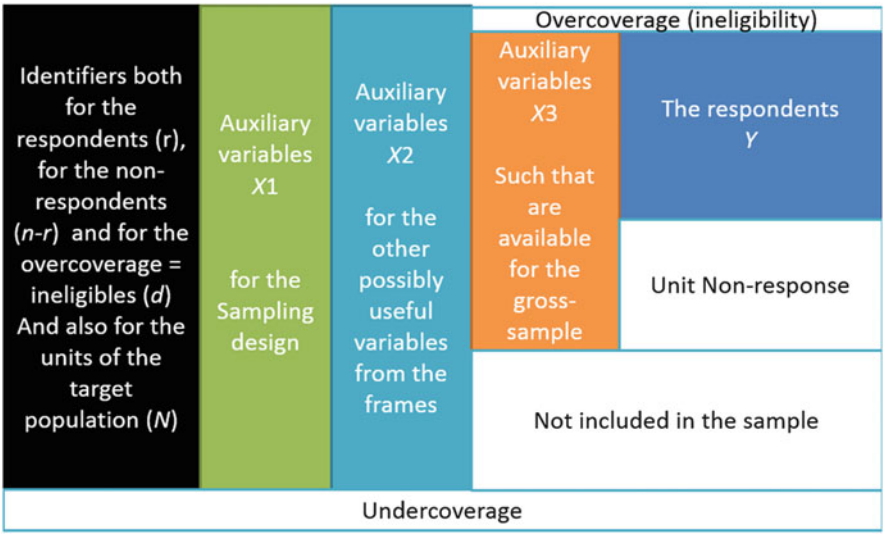
The simplest scheme is thus never found in real life but is a micro-survey file based on a sample. This means that only a certain proportion of the frame population and of the target population is fully achieved. On the other hand, missingness because of unit non-response occurs. Other gaps are appearing as well and new concepts are needed. The following scheme illustrates these and are explained, respectively. See also Scheme 2.3.

The measures of this scheme are not the same as in real life because the sample fraction is not as big as it is here but may be 1–5% of the target population size N. (Note that the symbol U is used as the reference population of N.) The sample size, respectively, is symbolized by n, and the number of the respondents by r. The symbol for overcoverage units or ineligibles is D in the frame and d in the sample. These two concepts are helpful to distinguish because the sample ineligibles are often known if they are contacted; however, the entire D population is not necessarily well known. This may cause biases in estimates.

Scheme 2.2 includes three groups of auxiliary variables. All their symbols are X, but we now have more such variables. The auxiliary group $X1$ corresponds to the



Scheme 2.1 Micro data for the entire target population



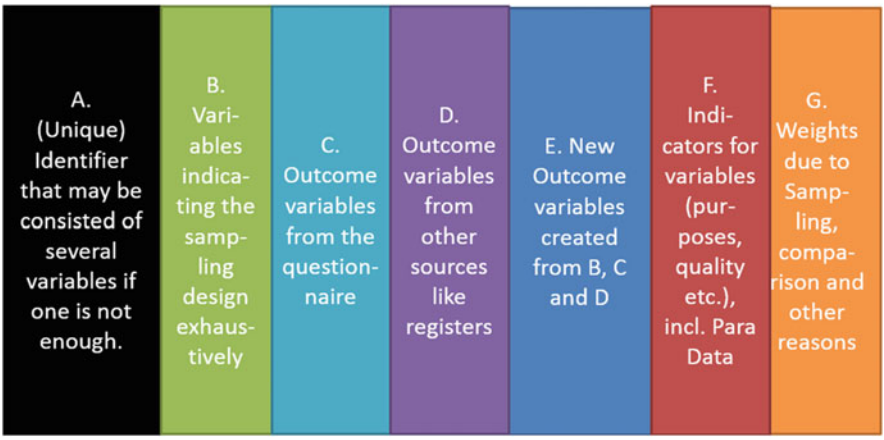
Scheme 2.2 General structure of a micro-level cross-sectional survey data file (The weights variables are not given here)

frame variables, *X*, in Scheme 2.1. These are used in sampling and should be available for the entire population. They are considered in more detail in the chapters on sampling—Chaps. 4, 5, and 6.

The frame is not usually complete, one reason being *undercoverage.* There can be other types of auxiliary variables than those for sampling. Variables *X2* often are available from the same source as sampling frame variables, but also from other registers if a general register is available. Some auxiliary variables can be obtained for the gross sample. We go into detail with examples of auxiliary variables in Chap. 6, in particular. The following explains what is behind the other concepts of the Scheme 2.2.

2.4.1 Specific Examples of Problems in the Data File

- *Unit non-response:* not contacted, unable to participate, refusals (hard and soft), fieldwork mistakes.
- *Item non-response:* These are missing values for survey variables. There can be many reasons for this, such as 'do not know', 'too confidential to answer', 'refusal to answer', 'not applicable'. These are considered further in other chapters and, in particular, in empirical examples.
- *Overcoverage (ineligible):* Examples—died, emigrants, living outside the target population, errors in the frame. Some of these can be observed during the fieldwork, although not all. This is a worsening problem nowadays because, if the unit (person) is not contacted, it is difficult to know whether a unit is ineligible or a unit non-respondent.



Scheme 2.3 General structure of a micro-level cross-sectional survey data file that consists of *r* respondents (That is, rows in a matrix)

- *Undercoverage:* Examples—new-born, new immigrants, illegally living in a country, errors in the frame. The updated frame helps to discover them. If it is not available, an effort should be made to assess its importance using external statistical sources.

A *real survey file* is not the same as the scheme in Scheme 2.2, except in some special cases such as methodological experiments using simulations. There are two real files:

- A sampling design (data) file that covers the gross sample units and auxiliary variables. This file is considered in detail in Chap. 6. From this file we usually create the sampling weights and other sampling design variables and merge these into the following.
- The file of the respondents, which is used in the analysis shown in Chap. 14, in particular. The scheme of this file is given next (Scheme 2.3).

It is possible that there are other data outside this scheme; for example, para data and content data. Good meta data should be available for all variables. If the file is released to outsiders, the identifiers should be anonymous. Initial variables rarely can be used as such in analysis.

We present examples in subsequent chapters of how either the new variables can be created from each initial variable using a different scaling or another transformation, or a new variable can be combined from several initial variables. Two larger examples of such transformations are given at the end of this chapter, thus this relates to the *E* variables.

Sampling weights are of two types: