CHAPTER ONE

The Nonresponse Problem

1.1 Introduction

We live in an information society. There is an ever growing demand for statistical information about the economic, social, political, and cultural shape of the country. Such information enables policy makers and others to make informed decisions for a better future.

Sometimes it is possible to retrieve such statistical information from existing administrative sources such as public registers. More often there is no such sources. Then a survey is the best instrument to use for collecting new statistical information.

A *survey* collects information about a specific population. This population need not necessarily consist of persons. For example, the elements of the population can be households, farms, companies, or schools. Typically information is collected by asking questions about the elements in the population. To do this in a uniform and consistent way, a questionnaire is developed.

One way to carry out a survey is to collect information on all elements in the population. Such a survey is called a *census* or a *complete enumeration*. This approach has a number of disadvantages. In the first place, it is very expensive. Surveying a large population requires a lot of people (e.g., interviewers) and a lot of computer resources. In the second place, it is very time-consuming. Collecting and processing a large amount of data takes time. This affects the timeliness of the results of the survey, and less timely information is less useful. In the third place, large surveys increase the response burden more than small surveys. As more and more people are asked to participate in surveys, they are less inclined to cooperate.

Handbook of Nonresponse in Household Surveys, By Jelke Bethlehem, Fannie Cobben, and Barry Schouten

A sample survey is a solution to many of the problems of complete enumeration because it collects information on only a small part of the population. This small part is called the *sample*. In principle, the sample only provides information on the sampled elements of the population. There is no information on the nonsampled elements. Still, if the sample is selected in a "clever" way, it is possible to make inference about the population as a whole. In this context, "clever" means that the sample is selected by means of probability sampling. A random selection procedure determines which elements are selected, and which not. If the survey researcher knows how the selection mechanism works and if it is possible to compute the probabilities of being selected in the sample, the results can be used to draw reliable conclusions about the nonsampled elements.

At first sight, the idea of introducing an element of uncertainty in the investigation seems odd. How can a survey researcher say something about a population as a whole by investigating only a small randomly selected part of it? The theory of statistical inference shows that this indeed is possible. Many books about the mathematical and statistical background of survey sampling have been published. Examples are Deming (1950) and Hansen et al. (1953), Cochran (1977), and Bethlehem (2009). The basic principles and concepts of survey sampling are summarized in chapter 2.

The first ideas on survey sampling emerged around the year 1895. See Bethlehem (2009) for an historic overview. The principles of probability sampling have been successfully applied on a regular basis in official and academic statistics since the 1940s, and to a much lesser extent also in commercial market research. Nevertheless, the survey organization does not have full control over the survey process. Practical problems may occur while collecting survey data. One of these problems is nonresponse. *Nonresponse* occurs when elements in the population that are sampled, and that are eligible for the survey, do not provide the requested information, or provide information that is not usable.

When confronted with nonresponse in a survey, a researcher loses control over the sample selection mechanism of the survey. Some groups in the population will be underrepresented in the survey, while other groups will be overrepresented. The immediate consequence of nonresponse is that, without taking special measures, it is not possible to compute reliable estimates of population characteristics. Validity of inference about the population is at stake. Both the composition and size of the sample are affected.

This handbook is about the nonresponse problem. It shows what the effects of incomplete sampling on the outcomes of surveys can be. It also suggests what can be done about the problem. An obvious consideration would be to prevent nonresponse from happening in the first place. This is discussed in more fully in Chapter 3. Practical experience, however, shows that it is impossible to eliminate nonresponse completely. Therefore some corrective action always has to be taken.

This chapter provides a general introduction on the phenomenon of nonresponse and its effect on the usefulness of survey-based estimates. As is shown, nonresponse has become a serious problem.

1.2 Theory

1.2.1 CAUSES AND EFFECTS OF NONRESPONSE

Surveys are often compromised by nonresponse. If the sampled population does not provide the requested information on selected items the collected information is unusable. Two types of nonresponse can be distinguished:

- *Unit nonresponse*. The selected person does not provide any information at all, meaning the questionnaire form remains completely empty.
- *Item nonresponse*. Some questions have been answered but not all questions, especially sensitive questions. So the questionnaire form has been partially completed.

A consequence of unit nonresponse is that the realized sample size is smaller than planned. If nonresponse is random, it will result in increased variances of the estimates, and thus in a lower precision of estimates. Valid estimates can still be obtained, however, because the computed confidence intervals will have the proper confidence level.

If a specific sample size is required, it is important to take into account that nonresponse will occur. For example, if a researcher wants to have at least 1000 completed questionnaires, and the nonresponse rate is expected to be in the order of 60%, the initial sample size should be approximately equal to $1000/0.6 = 1667$.

The main problem of nonresponse is that estimates of population characteristics may be biased. This situation occurs if some groups in the population are over- or underrepresented in the sample, and these groups behave differently with respect to the characteristics to be investigated. This is called *selective nonresponse*.

Indeed estimates must be assumed to be biased unless very convincing evidence to the contrary is provided. Bethlehem and Kersten (1985) mention a number of Dutch surveys where nonresponse is selective:

- A follow-up study of the Dutch Victimization Survey showed that people who are afraid to be home alone at night are less inclined to participate in the survey.
- In the Dutch Housing Demand Survey, it turned out that people who refused to participate have fewer housing demands than people who responded.
- For the Survey of Mobility of the Dutch Population it was obvious that the more mobile people were underrepresented among the respondents.

It will be shown in Chapter 2 that the amount of nonresponse is one of the factors determining magnitude of the bias of estimates. The higher the nonresponse rate, the larger the bias will be.

EXAMPLE 1.1 Nonresponse in the Dutch Housing Demand Survey

The effect of nonresponse is shown in a somewhat simplified example using data from the Dutch Housing Demand Survey. Statistics Netherlands carried out this survey in 1981. The initial sample size was 82,849. The number of respondents was 58,972, which comes down to a response rate of 71.2%.

To obtain more insight in the nonresponse, a follow-up survey was carried out among the nonrespondents. They were also asked whether they intended to move within two years. The results are summarized in the table below:

Do you intend to move within 2 years?	Response	Nonresponse	Total
Yes	17,515	3,056	20,571
No	41,457	20,821	62,278
Total	58,972	23,877	82,849

Based on the response, the percentage of people with the intention to move within two years is 29.7%. However, for the complete sample (response and nonresponse) a much lower percentage of 24.8% is obtained. The reason is clear: there is a substantial difference between respondents and nonrespondents with respect to the intention to move within two years. For nonrespondents this percentage is only 12.8%

Nonresponse can have many causes, and it is important to distinguish these causes. To reduce nonresponse in the field, one needs to know what the underlying reasons and motives are. Moreover different types of nonresponse can have different effects on estimates, and therefore may require different treatment. (For a model of survey participation, see Groves and Couper, 1998.)

The many ways to classify nonresponse according to cause make it difficult to compare nonresponse for different surveys. Unfortunately, no internationally accepted standardized classification exists. There have been some attempts. The American Association for Public Opinion Research (AAPOR) has published a report with a comprehensive list of definitions of possible survey outcomes (see AAPOR, 2000). However, these definitions only apply to household surveys with one respondent per household, and samples selected by means of Random Digit Dialing (RDD). Lynn et al. (2002) have proposed a more general classification. This is the classification used here. The classification follows probable event outcomes when selected population members are approached in an attempt to obtain cooperation in a survey; see Figure 1.1.

First, contact must be established with the selected person. If this is not successful, there are two explanations. If the selected contact belongs to the target population of the survey (i.e., *eligible* population) and should rightly be

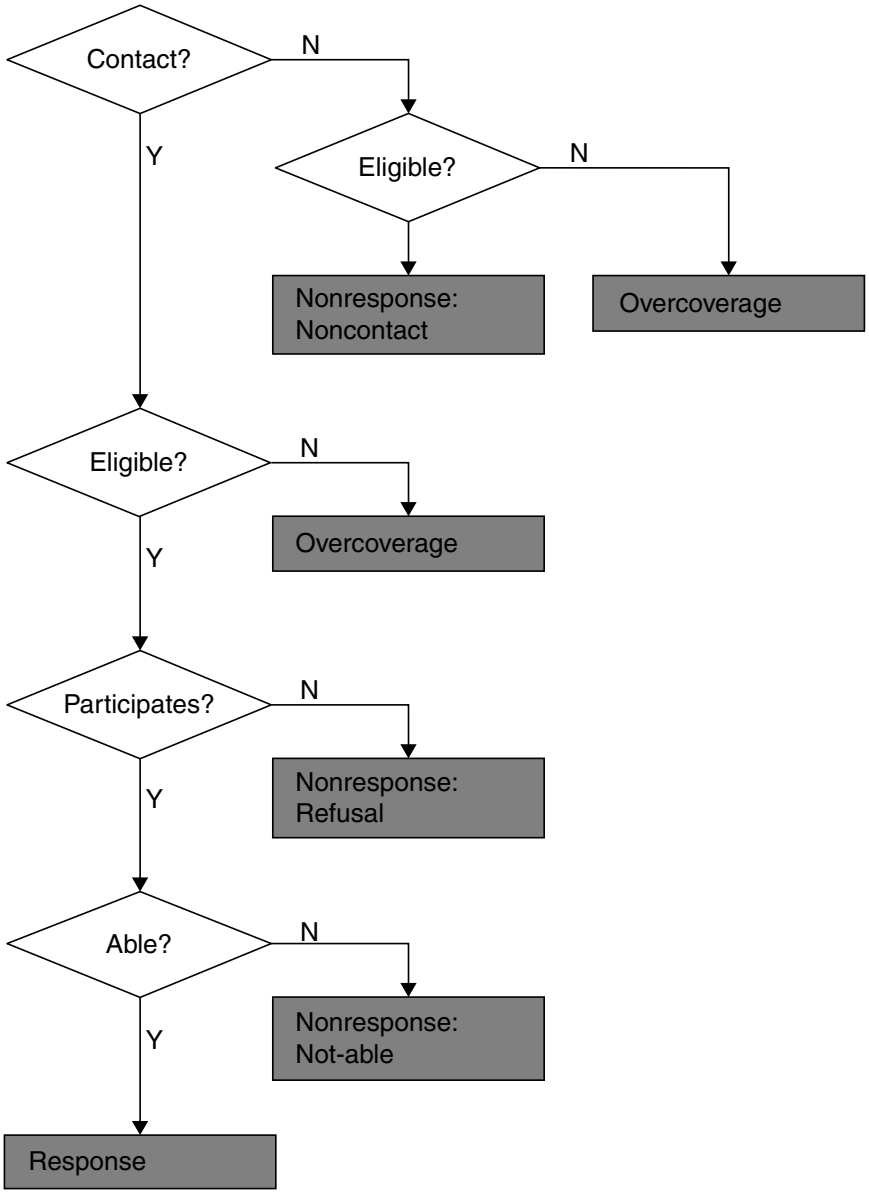


FIGURE 1.1 Possible survey outcomes.

part of the sample, this is nonresponse due to *noncontact*. If the selected contact does not belong to the target population (i.e., is not eligible), and should not be included in the sample, this is a case of overcoverage (see Section 1.3). That contact can therefore be excluded from the survey. Note that it is often not possible to determine in real-life situations whether a noncontact belongs to the target population, and this complicates the calculation of response rates.

Once there is contact with a selected person, we need to establish whether that contact belongs to the target population. If not, that case can be dismissed as an instance of overcoverage.

Once contact is established, that contact's cooperation is needed to get the required information. If the contacted person refuses to cooperate, this is a case of nonresponse due to *refusal*.

Once there is contact with someone who cooperates, there may still be an issue of that person not providing the required information. The reasons may range from illness to language problems. This is an instance of nonresponse due to *not-able*.

Last, say an eligible contact cooperates and is able to provide all the requested information; this is in fact a legitimate case of *response*.

Figure 1.1 shows the three causes of nonresponse: noncontact, refusal, and not-able. Nonresponse is not a permanent situation. In the case of noncontact, another contact may be attempted at some later time. In some surveys as many as six contact attempts are made before a case is closed as a noncontact. Also a refusal may be temporary if an interviewer calls at an inconvenient moment. It may be possible to suggest to a more convenient time for a follow-up. If someone is not able to participate because of illness, a later attempt may be made after the illness has passed. Also a language problem can be re-solved by translating the questionnaire or sending an interviewer capable of speaking the language of the respondent. Nevertheless, many refusals turn out to be nonnegotiable.

EXAMPLE 1.2 Results of the Survey on Well-Being of the Population

The Survey on Well-being of the Population that was carried out by Statistics Netherlands in 1998 had the following results:

Outcome	Frequency
Overcoverage	129
Response	24,008
Noncontact	2,093
Refusal	8,918
Not-able	1,151
Other nonresponse	3,132
Total	39,431

The category "not-able" includes nonresponse due to illness, physical handicap, or language problems. The additional nonresponse category, "other nonresponse," includes cases not followed up by interviewers because of workload. People who had moved and could not be located are also included in this category.

In some surveys the not-able category is split in two subcategories: not-able due to language problems and not-able due to other reasons. This is because these two types of nonresponse pertain to different groups of people and can have different effects on estimates.

The types of nonresponse given in Figure 1.1 do not exhaust the reasons for nonresponse. For example, it may happen that selected persons are not even contacted because of capacity problems of interviewers. This is sometimes called *administrative nonresponse*.

1.2.2 ERRORS IN SURVEYS

Nonresponse is just one thing that can go wrong in a survey. There are many more areas of data collection and data processing that can introduce errors and so affect the quality of the results.

There will always be some error in survey estimates of population characteristics. This error can have many explanations. Bethlehem (2009) gives some possible causes, as are presented in Figure 1.2. The taxonomy derives from a version given by Kish (1967).

The ultimate result of all errors is a discrepancy between the survey estimate and the population characteristic to be estimated. This discrepancy is called the *total survey error*. Two broad categories can be distinguished contributing to this total error: sampling errors and nonsampling errors.

Sampling errors are due to the sampling design. They are introduced when estimates are based on a sample and not on a complete enumeration of the population. Sampling errors could be avoided by investigating an entire population. However, only a part of the population is used for computing population characteristics. Because this is not a complete data set, estimates are only an approximation of the values of population characteristics, and some loss of precision results. The sampling error may be one of two types: a selection error or an estimation error.

The *estimation error* can occur when using a sample based on a random selection procedure. Every new selection of a sample will produce different respondents, and thus a different value of the estimator. The estimation error can be controlled through the sampling design. To reduce the error in an estimate,

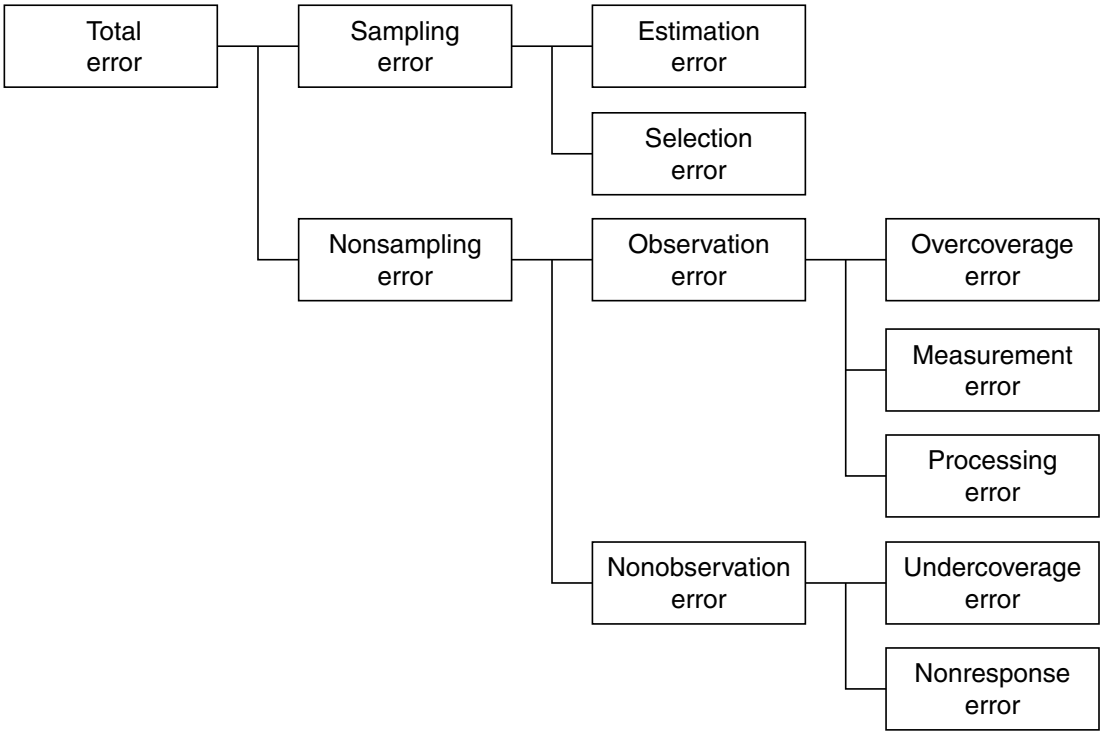


FIGURE 1.2 Types of survey errors.

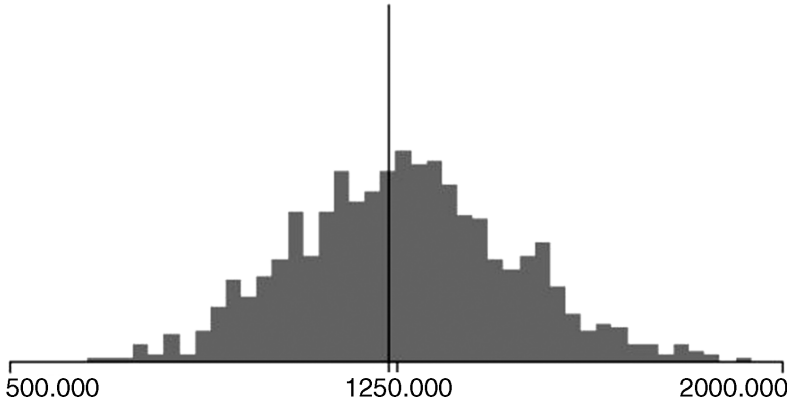
the sample size could be increased, or selection probabilities could be taken proportional to the values of some well-chosen auxiliary variable.

A *selection error* can occur when incorrect selection probabilities are used in an estimation procedure. For example, true selection probabilities may differ from anticipated selection probabilities if elements have multiple occurrences in the sampling frame. Selection errors are hard to avoid without thorough investigation of the sampling frame.

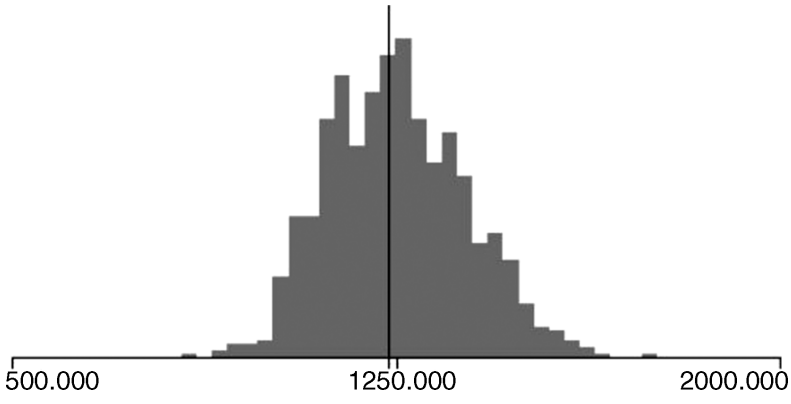
Nonsampling errors may occur even if the whole population is investigated. They denote errors made during the process of obtaining answers to questions asked. Nonsampling errors can arise from both observation and nonobservation errors.

EXAMPLE 1.3 Effect of Sample Size

The effects of selection error can be illustrated by a simulation experiment. From the working population of the small country of Samplonia 1000 samples of size 20 are selected. For each sample the mean income is computed as an estimate of the mean income in the population. The distribution of these 1000 estimates is displayed below:



There is a lot of variation in the estimates around the population mean, as is indicated by the vertical line. This variation can be reduced be increasing the sample size. The figure below shows the distribution of 1000 estimates based on sample of the size 40. Notice that doubling the sample size reduces the magnitude of the error.



Observation errors are one form of nonsampling errors. They are errors that are made during the process of obtaining and recording answers. An *overcoverage error* occurs when elements are included in the survey do not belong to the target population. A *measurement error* occurs when a respondent does not understand a question, or does not want to give the true answer, or when the interviewer makes an error in recording the answer. Also interviewer effects, question wording effects, and memory effects belong to this group of errors. A measurement error results from a discrepancy between the true value and the value processed in the survey. A *processing* error is an error made during data processing, such as in data entry.

Nonobservation errors are errors made when the intended measurements are not obtained. *Undercoverage* occurs when elements of the target population do not have a corresponding entry in the sampling frame. These population members cannot ever be contacted. Another type of nonobservation error is *nonresponse*. This phenomenon occurs if the sampled person does not provide the required information.

Figure 1.2 clearly shows that many things can go wrong during the data collection process, and usually they do. Some errors can be avoided by taking preventive measures at the design stage. However, some errors will remain. Therefore it is important to check collected data for errors, and where possible, to correct these errors. This activity is called *data editing*. Data editing procedures do not correct for every type of survey error; they can detect and remove measurement errors, processing errors, and possibly overcoverage errors. Phenomena like selection errors, undercoverage, and nonresponse require the use of adjustment weights in estimation procedures, and not correction of individual values in recorded data.

There are two ways in which nonresponse errors can be minimized. First is by *nonresponse reduction*, whereby every effort is made to prevent nonresponse from occurring in the field. Ideally, if everyone sampled responds, there will be no nonresponse error. Nonresponse reduction measures can include better contact strategies, application of refusal conversion techniques, and deployment of interviewers speaking different languages. Unfortunately, nonresponse can never be eliminated completely. Nonresponse reduction is the topic of Chapter 3. The second way is by *nonresponse correction*, which recognizes that it is not possible to obtain 100% response, then a technique must be applied to reduce the bias of the estimators. An example of such a correction technique is adjustment weighting. This is the topic of Chapter 8.

1.2.3 NONRESPONSE AND UNDERCOVERAGE

From the survey errors shown in Figure 1.2, we see that some people may be missing from the sample for reason of nonresponse or undercoverage. It is important to make a distinction between these two types of missing information.

Nonresponse denotes the situation where a member of the target population (and thus eligible for the survey) does not submit the required information.

Undercoverage is created by the sampling frame from which the sample is selected. Undercoverage denotes the situation where the sampling frame does

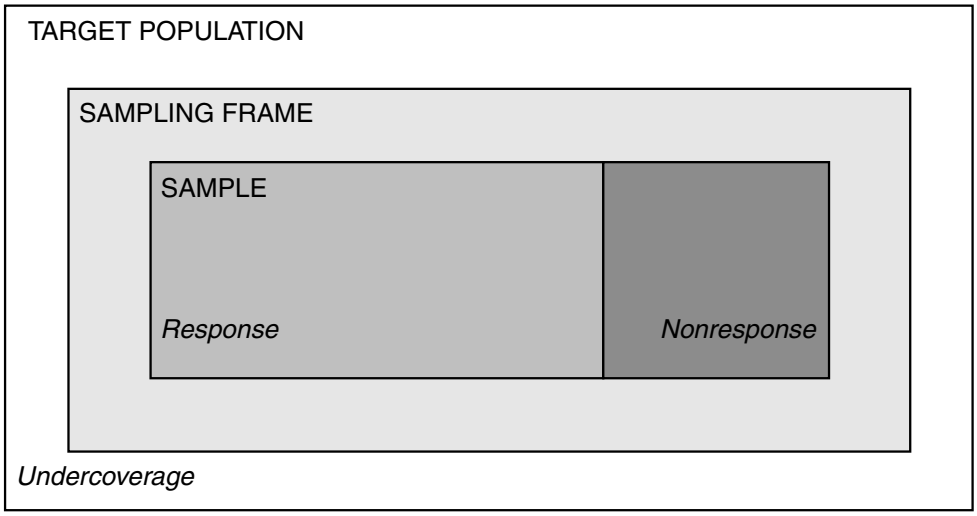


FIGURE 1.3 **Nonresponse and undercoverage.**

not cover completely the target population of the survey. There are persons in the population who do not appear in the sampling frame. They will, as a consequence, never be sampled.

The difference between nonresponse and undercoverage is shown in Figure 1.3. Undercoverage is a defect of the sampling frame that is often difficult to detect in practical survey situations. If people from some subpopulations do not appear in the sample, it may be simply that the sampling mechanism happened not to have selected them.

Undercoverage can also occur if the sample selected is from a different population than the one intended. The consequences for the outcomes of the survey is that the conclusion drawn from the survey does not apply to the original target population but to the population that was contacted through the sampling frame (sometimes called the *frame population*).

EXAMPLE 1.4 Nonresponse and Undercoverage

Suppose that a telephone survey is conducted. The target population consists of all adults in a certain country. The sampling frame is a telephone directory. There is undercoverage, because people without a listed number will never be selected in the sample. There will also be nonresponse because some calls to the selected persons will not be answered. And if calls are answered, persons may refuse to cooperate.

Suppose that a Web survey is conducted. The target population consists of all adults in a certain country. The sampling frame used is the population register. There is no undercoverage because the sample completely covers the population. The frame population is identical to the target population. There will be nonresponse. Not-able may be an important cause of nonresponse, since those people without Internet access will not be able to respond.

1.2.4 RESPONSE RATES

Because of the negative impact that nonresponse may have on the quality of survey results, the *response rate* is regarded as an important indicator of the quality of a survey. Response rates are frequently used to compare the quality of surveys and also to explore the quality of a repeated survey over time.

Presently there is no internationally accepted standard definition for the response rate. The definition we use here is based on one introduced by Lynn et al. (2002): the response rate is defined as the proportion of eligible contacts in the sample who completed the questionnaire. Referring to Figure 1.3, we therefore write the initial sample size n_I as

$$n_I = n_{NC} + n_{OC} + n_{RF} + n_{NA} + n_R, \tag{1.1}$$

where n_{NC} denotes the number of noncontacts, n_{OC} the number of noneligible contacts (i.e., cases of overcoverage), n_{RF} the number of refusers, n_{NA} the number of not-able respondents, and n_R the total number of respondents.

The *response rate* is defined as the total number of respondents divided by the number of n_E eligible contacts in the sample:

$$\textit{Response rate} = \frac{n_R}{n_E}. \tag{1.2}$$

There is a problem in computing the number of eligible elements. This problem arises because the noncontacts consist of eligible noncontacts and noneligible noncontacts. It is not known how many of these noncontacts are eligible. If it is assumed that all noncontacts are eligible, then $n_E = n_{NC} + n_{OC} + n_{RF} + n_{NA} + n_R$. So the response rate is equal to

$$\textit{Response rate} = \frac{n_R}{n_{NC} + n_{RF} + n_{NA} + n_R}. \tag{1.3}$$

This is usually not a realistic assumption. Another assumption is that the proportion of eligibles among the noncontacts is equal to the proportion of eligibles among the contacts. Then the response rate would be equal to

$$\textit{Response rate} = \frac{n_R}{n_{NC}\dfrac{n_{RF} + n_{NA} + n_R}{n_{OC} + n_{RF} + n_{NA} + n_R} + n_{RF} + n_{NA} + n_R}. \tag{1.4}$$

Response rate definitions like (1.3) or (1.4) can be used in a straightforward way for surveys where one person per household is selected. The situation becomes more complicated when the survey population consists of households for which several or all of its members have to provide information. Then there is risk of partial response: some eligible household members may respond, but for other eligible members it may be impossible to obtain response. How to define response at the household level? There are examples of surveys where an outcome is defined as response only if all eligible members respond. So, response rates for households may differ from response rates for persons.

Another complication arises from self-administered surveys. These are surveys where there are no interviewers. Examples of such surveys are mail surveys (pen-and-paper surveys) or Web surveys. For such surveys it is not possible to distinguish among the different sources of nonresponse. There are only two possible outcomes: response and nonresponse. The response rate simplifies to

$$\textit{Response rate} = \frac{n_R}{n_R + n_{NR}}. \tag{1.5}$$

Self-administered surveys also do not control for variation in the eligible population. The extreme example is the self-selection survey. Internet surveys are often self-selection surveys, as it is a convenient way for an organization to survey a large group of people. No proper sample is selected for such surveys. The survey questionnaire is simply put on the Internet. Respondents are those with Internet access who visit the survey website and decide to complete the questionnaire. The survey researcher has no control over the selection process. It is unclear whether respondents belong to the target population of the survey.

EXAMPLE 1.5 Computing the Response Rate

The Dutch Survey on Well-being of the Population had the following fieldwork results in 1998 (see also Example 1.4):

Outcome	Frequency
Overcoverage	129
Response	24,008
Noncontact	2,093
Refusal	8,918
Not-able	1,151
Other nonresponse	3,132
Total	39,431

The category "Not-able" included nonresponse due to illness, handicap, or language problems. The extra nonresponse category "Other nonresponse" consisted of cases not processed by interviewers due to workload (administrative nonresponse). Also people who had moved and could not be found any more are included in this category.

If it is assumed that the noncontacts are all eligible, the response rate of this survey is

$$100 \times \frac{24008}{24008 + 2093 + 8918 + 1151 + 3132} = 61.09\,\%.$$

If it is assumed that the proportion of eligibles among contacts and noncontacts is the same, the response rate is equal to

$$100 \times \frac{24008}{24008 + 2093 \times \dfrac{39431-2093-129}{39431-2093} + 8918 + 1151 + 3132} = 61.11\%.$$

The differences in both response rates are small. This is due to the small amount of overcoverage.

Yet another complication that can affect the definition of the response rate is the use of sampling designs with unequal selection probabilities. On the one hand, because the response rate is used as an indicator of the quality of survey outcomes, the sizes of the various outcome categories should reflect the structure of the population. Consequently observations should be weighted with inverse selection probabilities. This leads to so-called *weighted response rates*. On the other hand, because the response rate is used as an indicator of the quality of the fieldwork, and more specifically the performance of interviewers, *unweighted response rates* may be more appropriate.

Response rates have declined over time in many countries. Table 1.1 contains (unweighted) response rates for a number of surveys of Statistics Netherlands. The definition of response rates is more or less the same for each survey. It is not easy to explain differences in response rates between surveys. Response rates are determined by a large number of factors, such as the topic of the survey, the target population, the time period, the length of the questionnaire, the quality of the interviewers, and the organization of the fieldwork.

The response rates for different surveys cannot be readily compared. Different surveys may have different target populations. For this reason response rates for interviewers or interviewer regions are usually adjusted for the composition of the population in the interviewer area.

As Table 1.1 shows, nonresponse can be a big problem. Nonresponse has become more of a problem in recent years. It has raised the cost of conducting surveys since more effort has to be expended to obtain estimates with the precision specified in the survey design.

The Labor Force Survey (LFS) is the most important survey of Statistics Netherlands. It has been subjected to many re-designs, the most comprehensive redesign taking place in 1987. Before 1986, data collection was carried out by means of a paper questionnaire (PAPI, paper and pencil interviewing). In 1987 Statistics Netherlands changed to computer-assisted interviewing, which was facilitated by the Blaise System. With this development, the LFS introduced computer-assisted personal interviewing (CAPI). Also, before 1986, the fieldwork for the LFS was carried out by municipal employees who were not professional interviewers. From 1987, each month about 400 interviewers equiped with laptops visited 12,000 addresses.

In 1987 all changed. The questionnaire of the LFS was completely redesigned, and the fieldwork began to be done by professional interviewers. Notice in Table 1.1 the large drop in the response rate of the LFS in 1987. Because this

TABLE 1.1 Response rates of some surveys of Statistics Netherlands

Year	Labor Force Survey	Consumer Sentiments Survey	Survey on Well-being of the Population	Mobility Survey	Holiday Survey
1972		71			
1973	88	77			
1974		75	72		
1975	86	78			86
1976		72	77[a]		87
1977	88	69	70		81
1978		64		67	78
1979	81	63	65[b]	69	74
1980		61	61	68	74
1981	83	65		68	74
1982		60	64[a]	66	71
1983	81	63	58	66	74
1984		65[c]		64	69
1985	77	69		61	68
1986		71	59	59	66
1987	60[c]	71		59	
1988	59	68		55	
1989	61	68	44	58	
1990	61	68	47	55	
1991	60	69	46	57	
1992	58	69	45	57	
1993	58	72	46	56	
1994	59	70	52[c]	55	
1995	60	67	54	54	
1996	58	67	52	52	
1997	56	57	63	50	
1998	54	64	60		
1999	56	62	60		
2000	56	61	57		
2001	58	64	60		
2002	58	65			
2003	59	65	62		

[a] *Young only.*
[b] *Elderly only.*
[c] *Change in survey design.*

was when so many things were changed in the survey design and the survey fieldwork, no single cause could explain this drop.

Another important survey of Statistics Netherlands is the Survey of Well-being of the Population (SWP). It is a survey that samples every month a group of 3000 selected persons. The survey has a modular in structure; a base module contains questions for all sampled persons, and in addition there are a number of modules about specific themes (e.g., employment situation, health, and justice). The sampled persons are selected for one of the thematic modules; the base module is answered by everyone. The SWP was created in 1997; before that year all the modules were separate surveys.

The Consumer Sentiments Survey (denoted by CSS) measures consumer confidence (e.g., in changing economic circumstances). Since April 1986, it is performed monthly by means of CATI (computer-assisted telephone interviewing). Before 1984, the survey was conducted by telephone interviewers using pen and paper (PAPI) to record responses. Every month 1500 households are selected in a simple random sample. Telephone numbers are obtained for the selected addresses from listed numbers of landline telephones. This is only possible for about two-thirds of the addresses. The phone numbers are then passed to the CATI interviewers. Only one person in every household is interviewed. The response rates of these three major surveys are graphically presented in Figure 1.4.

Notice that from 1972 to 1983 the response percentages of the CSS and the SWP show similar, falling trends. After 1983 the response percentage for the CSS stabilized, whereas for the SWP it kept on falling. Both rates start to converge in 1993 and settle into a similar pattern over the last six years. The two breakpoints coincide with re-designs of these surveys (CSS in 1984 and SWP in 1997). The re-design of the CSS in 1984 caused a temporary increase in response rates. The same is true for the re-design of the SWP in 1997.

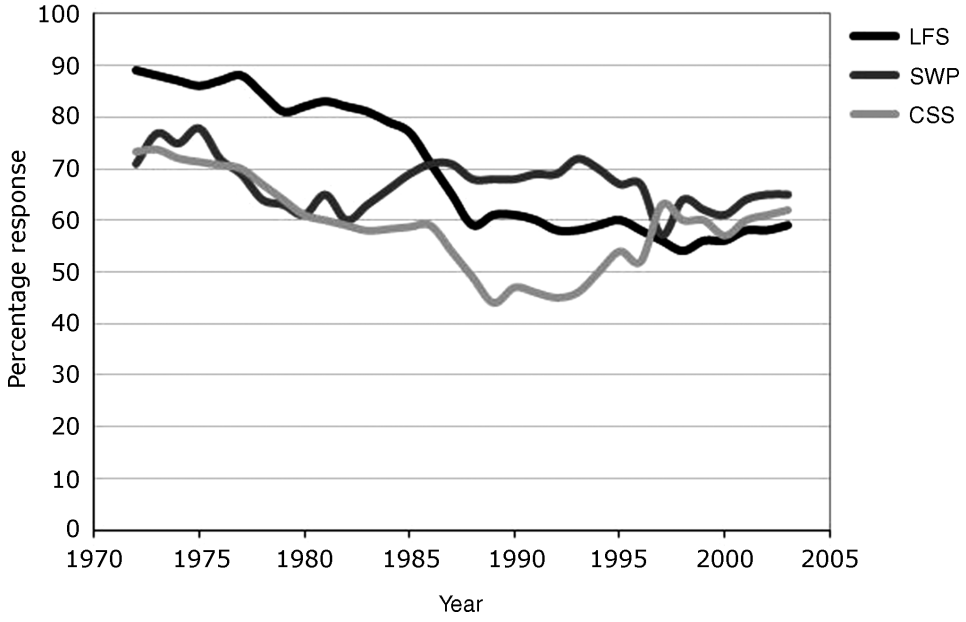


FIGURE 1.4 Response percentages for three Dutch Surveys: LFS, SWP, and CSS.

The response percentage of the LFS was initially higher than that of the other two surveys, but around 1983 and 1984 it decreased and reached the same level as the rates of the CSS and SWP. From 1987, responses took more or less stable paths. As mentioned before, 1987 was the year of a comprehensive re-design of the LFS.

1.2.5 REPRESENTATIVITY

Conducting a survey means two selection processes take place. First, a sampling design is chosen for selecting the sample and so is based on some form of probability sampling. Next, as nonresponse occurs in the fieldwork, only answers obtained from respondents (the survey response) can be used for analysis. The question is whether this data set allows for proper inference with respect to the population. Is it possible to draw reliable conclusions?

It is often said that the survey response must be *representative*, but what does it mean? Kruskal and Mosteller (1979a, b, c) present an extensive review of what representative is supposed to mean in nonscientific literature, in scientific literature excluding statistics, and in the current statistical literature. They compiled the following ways to consider if a sampling is truly "representative":

1. Validation for data. It means not much more than a general assurance, without evidence, that the data are supportable. This meaning of "representative" is typically used by the media, without explaining what it exactly means.
2. Absence of bias. No elements or groups of elements were favored in the selection process, either consciously or unconsciously.
3. Miniature of the population. The sample is a small-scale model of the population. The sample has the same characteristics as the population. The sample proportions are in all respects similar to population proportions.
4. Typical or ideal subject(s). The persons sampled are "typical" of a certain population. They are representative in terms of the idea of *l'*homme moyenne ("average man"), which was introduced by the Dutch/Belgian statistician Adolphe Quetelet (1835, 1846).
5. Allowances for a population's heterogeneity. The variation that exists in a population should be provided for in the sample by including atypical subjects.
6. A vague term that is used without describing what it means.
7. Equal probability sampling. A form of probability sampling was used giving equal selection probabilities to each member of a population.
8. As permitting good estimation. All characteristics of a population and its variability must be present in the sample so that the estimates computed are reliable.
9. Suitability for a particular purpose. Sample should show that a phenomenon thought to be very rare or absent occurs with some frequency.

Because the term "representative" can have many different interpretations, it is best not to use it in practice unless it is made clear what meaning is intended. In this book the term "representative" is used in two ways.

First, we say that the survey response is *representative with respect to* a variable if its relative distribution in the survey response is equal to its relative distribution in the population. For example, a sample is representative with respect to the variable "gender" if the percentages of males and females in the survey response are equal to the percentage of males and females in the population. Chapter 8 covers weighting adjustment techniques to correct for nonresponse problems. These techniques help make the survey responses representative of as many variables as possible. Of course the survey responses should be representative with respect to all other survey variables as well.

Second, we say that the response mechanism is representative if each member of a population in the sample would have the same probability of response. This implies that there are no biases active in the selection process. Chapter 7 covers indicators used for measuring representativity. These indicators estimate response probabilities and attempt to determine whether or not they are the same.

1.3 Application

Throughout this book theoretical concepts are applied to real survey data. The data are derived from a Dutch survey that was carried out by Statistics Netherlands. To avoid the disclosure of sensitive individual information, the data set has been anonymized. It is called the General Population Survey (GPS).

The fieldwork of the GPS covered a period of two months. In the first month, selected persons where approached by means of CAPI (computer-assisted personal interviewing). For persons that could not be contacted or refused and who had a listed phone number, a second attempt was made in the second month now using CATI (computer-assisted telephone interviewing). Table 1.2 shows the fieldwork

TABLE 1.2 Fieldwork results of the GPS

Result	Frequency	Percentage
Sample size	32,019	100.0%
Response	18,792	58.7%
First month	14,870	46.4%
Second month	3,922	12.2%
Nonresponse	13,227	41,3%
Unprocessed cases	2,456	7.7%
Noncontact	1,847	5.8%
Refusal	7,890	24.6%
Not able	1,034	3.2%

results. Note that there is a nonresponse category "Unprocessed." This denotes nonresponse due to unprocessed cases. Such cases were assigned to interviewers but were not undertaken by interviewers because of capacity problems or illness of the interviewer.

The selection of participants was by a stratified two-stage sampling process. In the first stage, municipalities were selected within regions with probabilities proportional to the number of inhabitants. In the second stage, an equal probability sample was drawn in each selected municipality. Sampling frames for the persons were the population registers of the municipalities. The sampling design was such that each person had the same probability of being selected (a so-called self-weighting sample). The sample of the GPS consisted of 32,019 persons. The number of respondents was 18,792.

Statistics Netherlands has an integrated system of social statistics which is called the *Social Statistics Database* (SSD). The SSD contains a wide range of information on each person who lives in The Netherlands. There are data on demography, geography, income, labor, education, health, and social protection. These data are obtained by combining data from registers and other administrative data sources. For more information about the SSD, see Everaers and Van Der Laan (2001).

SSD records can be linked to survey data records by way of personal identification numbers. This can be done for both respondents and nonrespondents, so demographic variables like gender, age, province of residence, and ethnicity are available for all sampled persons, and also socioeconomic variables like employment and various types of social security benefits.

The Netherlands is divided in approximately 420,000 postal code areas. A postal code area contains, on average, 17 addresses. These areas are homogeneous with respect to social and economic characteristics of its inhabitants. Using information from the population register, Statistics Netherlands has computed some demographic characteristics for these postal code areas. Since postal codes are included in the survey data file for both respondents and nonrespondents, these characteristics can be linked to the survey data file. Examples of such variables are degree of urbanization, town size, and percentage of people with a foreign background (nonnatives). From another source average house values can be included.

During their fieldwork the interviewers kept records of all contact attempts. For each attempt the contact result was recorded (successful, or not). If contact was established, the result of the cooperation request was recorded—response or nonresponse, and in case of nonresponse the reason of nonresponse. Other information was included, like the mode of contact used in the fieldwork attempt (CAPI or CATI). All the fieldwork information is available for analysis.

Table 1.3 gives an overview of all variables in the survey data file of the GPS. The values of the target variables are only available for the respondents. The auxiliary variables are available for both respondents and nonrespondents.

There were 150 cases in the initial sample that did not belong to the target population of the GPS. These cases are not contained in GPS survey data file. The amount of overcoverage is needed to be able to compute the response rate of

TABLE 1.3 Variables in the GPS survey data file

Variable	Description	Cats	Type	Source
PC	PC in household	2	Target variable	Survey
Move	Wants to move within a year	6	Target variable	Survey
Health	General health condition	5	Target variable	Survey
Newspap	Has newspaper subscription	2	Target variable	Survey
Club	Is active in a club	2	Target variable	Survey
Politics	Is interested politics	4	Target variable	Survey
Employed	Employment situation	3	Target variable	Survey
Joblevel	Job level	6	Target variable	Survey
Educat	Level of education	8	Target variable	Survey
Ownhouse	Owns house	2	Target variable	Survey
Religion	Religious denomination	5	Target variable	Survey
Gender	Gender	2	Auxiliary variable	SSD
Marstat	Marital status	4	Auxiliary variable	SSD
Married	Is married	2	Auxiliary variable	SSD
Age3	Age in 3 age groups	3	Auxiliary variable	SSD
Age13	Age in 13 age groups	13	Auxiliary variable	SSD
Nonnativ	Is nonnative	2	Auxiliary variable	SSD
Ethnic	Type of nonnative	5	Auxiliary variable	SSD
HHSize	Size of the household	5	Auxiliary variable	SSD
HHType	Type of household	5	Auxiliary variable	SSD
Children	Children in household	2	Auxiliary variable	SSD
Phone	Has listed phone number	2	Auxiliary variable	SSD
Hasjob	Has a job	2	Auxiliary variable	SSD
Socall	Has social allowance	2	Auxiliary variable	SSD
Disaball	Has disability allowance	2	Auxiliary variable	SSD
Unempall	Has unemployment allowance	2	Auxiliary variable	SSD
Allowan	Has an allowance	2	Auxiliary variable	SSD
Region	Region of the country	5	Auxiliary variable	SSD
Urban	Degree of urbanization	5	Auxiliary variable	SSD
Houseval	Average house value in neighborhood	12	Auxiliary variable	SSD
Pnonnat1	Percentage nonnatives in neighborhood	8	Auxiliary variable	SSD
Pnonnat2	Percentage nonwestern nonnatives in neighborhood	7	Auxiliary variable	SSD
Resp1	Response in first month of fieldwork	2	Fieldwork variable	SSD
Result	Final result of the fieldwork attempt	5	Fieldwork variable	SSD
Response	Response	2	Fieldwork variable	SSD

TABLE 1.4 Percentage of people with a job in GPS survey data file

	Job	No job	Total
Response	52.6%	47.4%	100.0%
Nonresponse	48.8%	51.2%	100.0%
Total	51.0%	49.0%	100.0%

the GPS. As was discussed in Section 1.2.4, the response rate can only be computed if the number of eligible cases among the noncontacts is known. If it is assumed that all noncontacts are eligible, the response rate is equal to

$$100 \times \frac{18792}{18792+2456+1847+7890+1034} = 100 \times \frac{18792}{32019} = 58.69\%.$$

If it is assumed that the proportion of eligibles among the contacted persons is the same as the proportion among the noncontacts, the response rate becomes

$$100 \times \frac{18792}{18792+2456+1847 \times \frac{32019-1847}{32019-1847+150}+7890+1034} = 58.71\%.$$

The differences are minimal. This is because the amount of overcoverage is very small. So rounded to one decimal, the response rate is 58.7%.

The nonresponse of the GPS is selective. For example, Table 1.4 contains the distribution of the (register) variable *Hasjob*. This variable records whether a person has a job.

Among the respondents 52.6% of the people have a job, and among the nonrespondents only 48.8%. Apparently, those without jobs are less inclined to respond. Table 1.5 gives more detail. It shows the different causes of nonresponse. As is clear from this table, the nonresponse is mostly due to persons not able to respond. Among those persons only 17.2% have a job. This low percentage is not surprising as people unable to respond often are so disabled that they are unable to work.

TABLE 1.5 Types of job-related responses and nonresponse of the GPS survey date file

	Job	No job	Total
Response	52.6%	47.4%	100.0%
Unprocessed	53.5%	46.5%	100.0%
Noncontact	51.4%	47.4%	100.0%
Refusal	50.9%	49.1%	100.0%
Not-able	17.2%	82.8%	100.0%
Total	51.0%	49.0%	100.0%

1.4 Summary

A survey is an instrument to collect information about a specific populations. Such populations may consist of persons, households, companies, or other elements. Typically not all people are asked to participate in a survey, only a sample.

With a good survey design accurate estimates of population characteristics can be computed. Also the accuracy of estimates can be computed. A good design suggests also that the sample is selected by means of probability sampling.

Nonresponse is a phenomenon that may affect the quality of the survey outcomes. It occurs when the people who are selected as eligible for the sample do not provide the requested information, or when the provided information is not usable. Nonresponse can cause estimators of the population characteristics to be biased. This occurs when specific groups are over- or underrepresented, and these groups may behave differently with respect to the survey variables.

Nonresponse is mainly due to noncontact, refusal to answer, and not-able to answer. It is important to distinguish among these different causes because they may have different impacts on estimates.

Unfortunately, computation of the response rate is not straightforward. This is because the proportion of eligible elements among the noncontacts is not known. An estimate of the response rate can be obtained if assumptions are made about this component.

Response rates have been decreasing in The Netherlands over the last few decades. So nonresponse has become a serious problem. It not only affects the quality of the survey outcomes, but attempts to reduce the problem also increases survey costs.

1.5 KEY TERMS

Eligible To be eligible, the sample elements selected must belong to the target population of the survey.

Item nonresponse Some questions have been answered, but no answer is given for other, possibly sensitive, questions. So the questionnaire form has been only partially completed.

Noncontact A type of nonresponse where it is not possible to establish contact with a sampled population member.

Nonresponse The selected person for the sample is eligible for the survey but does not provide the requested information, or provides information that is not usable.

Nonresponse correction An attempt to compensate for the nonresponse problem by adjusting estimates using survey responses and other information about the population.

Nonresponse reduction An attempt to compensate for the nonresponse problem by reducing the amount of nonresponse in the field.

Nonsampling error The difference between the estimate and the true value caused by other phenomena than sampling. Such errors may also occur if the complete population is investigated. Nonresponse is one type of nonsampling error.

Not-able A cause of nonresponse where contact is established with a sampled person who is not able to cooperate for reason of illness or language problems.

Overcoverage Happens if the sampling frame includes persons who do not belong to the target population of the survey. These persons should be excluded from the survey.

Refusal A cause of nonresponse where contact is established with a sampled persons who refuses to cooperate.

Response rate The number of responding eligible persons in the sample divided by the total number of eligible persons in the sample. Their response rate can be weighted or unweighted.

Representative with respect to a variable The distribution of this variable in the survey response is equal to the distribution of the variable in the population.

Representative response All members of the population have the same probability of response.

Sampling error The difference between the estimate and the true value caused by estimates that are based on a sample and not on the complete enumeration of the population.

Selective nonresponse Nonresponse that is selective if, due to nonresponse, some groups in the population are over- or underrepresented in the sample, and these groups behave differently from the characteristics being sampled.

Undercoverage Happens if a sampling frame does not cover completely the target population of the survey. There are persons in the population who do not appear in the sampling frame. They will never be included in the sample.

Unit nonresponse A type of nonresponse that occurs when a selected person does not provide any information at all, meaning the questionnaire form remains empty.

1.6 REFERENCES

AAPOR (2000), *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. American Association for Public Opinion Research, Ann Arbor, MI.

Bethlehem, J.G. (2009), *Applied Survey Methods: A Statistical Perspective*. Wiley, Hoboken, NJ.

Bethlehem, J.G., and Kersten, H.M.P. (1985), On the Treatment of Non-response in Sample Surveys. *Journal of Official Statistics* 1, pp. 287–300.

Cochran, W.G. (1977), *Sampling Techniques*, 3rd ed. Wiley, New York.

Deming, W.E. (1950), *Some Theory of Sampling*. Wiley, New York.

Everaers, P., and Van Der Laan, P. (2001), *The Dutch Virtual Census*. E-Proceedings of the 53th Session. International Statistical Institute, Seoul, Korea.

Groves, R.M., and Couper, M.P. (1998), *Nonresponse in Household Interview Surveys*. Wiley, New York.

Hansen, M.H., Hurvitz, W.N., and Madow, W.G. (1953), *Survey Sampling Methods and Theory*. Wiley, New York.

Kish, L. (1967), *Survey Sampling*. Wiley, New York.

Kruskal, W., and Mosteller, F. (1979a), Representative Sampling I: Non-scientific Literature. *International Statistical Review* 47, pp. 13–24.

Kruskal, W., and Mosteller, F. (1979b), Representative Sampling II: Scientific Literature Excluding Statistics. *International Statistical Review* 47, pp. 111–123.

Kruskal, W., and Mosteller, F. (1979c), Representative Sampling III: Current Statistical Literature, *International Statistical Review* 47, pp. 245–265.

Lynn, P., Beerten, R., Laiho, J., and Martin, J. (2002), Towards Standardisation of Survey Outcome Categories and Response Rate Calculations. *Research in Official Statistics* 1, pp. 63–86.

Quetelet, L.A.J. (1835), *Sur l'*Homme et le Développement de ses Facultés, Essai de Physique Sociale. Paris, France.

Quetelet, L.A.J. (1846), *Lettre à S.A.R. le Duc Régant de Saxe Coburg et Gotha sur la Théorie des Probabilités, Appliquée aux Sciences Morales at Politiques*. Brussels, Belgium.

1.7 EXERCISES

Exercise 1.1 Which phenomenon makes it difficult, if not impossible, to computer the response rate of a survey?

(a) Overcoverage

(b) Undercoverage

(c) Noncontact

(d) Refusal

Exercise 1.2 A survey is usually carried out to measure the state of a target population at a specific point in time (the reference date). The survey outcomes are supposed to describe the population at this date. Ideally the fieldwork of the survey should take place at that date. This is not possible in practice. Interviewing usually takes place in a period of a number of days or weeks around the reference date.

Suppose that a sample of persons is selected from a sampling frame (a population register) two weeks before the reference date. Interviewing takes place in the period of four weeks: the two weeks between sample selection and the reference date and the two weeks after the reference date.

Explain for each of the situations described below what kind of problem has occurred. You can choose between nonresponse, undercoverage, overcoverage, and any other sampling frame errors.

- The contact attempt takes place between the sample selection date and the reference date. It turns out the person died before the sample selection date.
- The contact attempt takes place between the sample selection date and the reference date. It turns out the person died after the sample selection date.

- The contact attempt takes place after the reference date. It turns out the person died before the sample selection date.
- The contact attempt takes place after the reference date. It turns out the person died between the sample selection date and the reference date.
- The contact attempt takes place after the reference date. It turns out the person died after the reference date.

Exercise 1.3 A town council wants to do something about the traffic problems in the center of town. There is a plan to turn it into a pedestrian area. Because cars will not be permitted into to center any more, the town council wants to know what companies think of this plan. A simple random sample of 1000 companies is selected. A representative of each selected company is invited to participate in the survey. They are asked whether they are in favor of the plan. Also the location of the company is recorded (town center or suburb). The results of the survey are summarized in the table below:

	Suburbs	Town center
In favor	120	80
Not in favor	40	240

(a) Determine the response percentage.

(b) Determine the percentage respondents in favor of the plan.

(c) Compute a lower bound and an upper bound for the percentage in favor of the plan in the complete sample.

Exercise 1.4 The local authorities in a town want to know how satisfied citizens are with the public transport facilities. They conduct a survey. The target population is defined as all citizens that used public transport at least once in the last year. A sample is selected from the population register of the town. The results of the survey are summarized in the table below:

Result	Frequency
Overcoverage	320
Refusal	240
Noncontact	80
Not able	40
Response	440
Total	1120

Compute the response rate of the survey. Clearly show how the response rate was computed and the assumptions made.

Exercise 1.5 Suppose that a telephone directory is used as a sampling frame and a survey researcher is not aware of the fact that some people appear more than once in this directory. What type of error could this phenomenon cause?

(a) Overcoverage

(b) Estimation error

(c) Selection error

(d) Processing error

Exercise 1.6 Analyze the dataset of the GPS that is provided on the website of the book. Use Tables 1.4 and 1.5 in Section 1.3 to find other variables that show a large difference between respondents and nonrespondents.