

CHAPTER 2

The Survey and Its Imperfections

2.1. THE SURVEY OBJECTIVE

The objective of a survey is to provide information about unknown characteristics, called parameters, of a finite collection of elements, called a population, such as a population of individuals, of households, or of enterprises. A typical survey involves many study variables and produces estimates of different types of parameters. Simple parameters are the total or the mean of a study variable, or the ratio of the totals of two study variables. Different types of elements are sometimes measured in the same survey, as when both individuals and households are observed.

Many surveys are conducted periodically, for example monthly or yearly. As a consequence, an important objective is to measure the change in the level of a variable between two survey occasions. The objectives of estimation of change and estimation of level often coexist in a survey, but they may require somewhat different techniques.

A survey usually originates in an expressed need for information about a social or economic issue, a need which existing data sources are incapable of filling. The first step in the planning process is to determine the survey objectives as clearly and unambiguously as possible. The next step, referred to as *survey design*, is to develop the methodology for the survey.

Survey design involves making decisions on a number of future survey operations. The data collection method must be decided on, a questionnaire must be designed and pretested, procedures must be set out for minimizing or controlling response errors, the sampling method must be decided on, interviewers must be selected and trained, questionnaires must be constructed and tested, techniques for handling nonresponse must be agreed on, and procedures for tabulation and analysis must be settled.

A survey will usually encounter various technical difficulties. No survey is perfect in all regards. The statistics that result from the survey are not error-free. The *frame* from which the sample is drawn is hardly ever perfect, so there will be *coverage errors*. There will be *sampling error* whenever observation is limited

to a sample of elements, rather than to the entire population. No matter how carefully the survey is designed and conducted, some of the desired data will be missing, because of refusal to provide information or because contact cannot be established with a selected element. Since nonresponding elements may be systematically different (for example, have larger or smaller variable values, on average) from responding elements, there will be *nonresponse error*.

These three types of error – sampling error, nonresponse error and coverage error – are discussed at length in this book, especially the first two. It is true that a survey will usually also have other imperfections, such as measurement error and coding error. These errors are not discussed.

Subpopulations of interest are called *domains*. If the survey is required to give accurate information about many domains, a complete enumeration within these domains may become necessary, especially if they are small.

The survey planner is likely to first consider whether statistics derived from available *administrative registers* could satisfy the need for information. This avenue can be followed in countries well endowed with high-quality administrative registers. If not, a *census* (a complete enumeration of the population) may have to be conducted. If all domains of interest are at least moderately large, a *sample survey* may give statistics of sufficient accuracy.

These three different types of survey – based on administrative registers, census survey and sample survey – differ not only in the extent to which they can produce accurate information for domains, but also in other important respects. For example, sample surveys have the advantage of yielding diverse and timely data on specified variables, whereas statistics derived from administrative registers, although perhaps less expensive, may give information of limited relevance, because except in fortunate circumstances, available registers are not designed to meet specific information needs. On the other hand, a census might provide the desired information with great accuracy, but is very expensive to conduct. For a discussion of these issues, see Kish (1979).

Most of the issues raised in the following apply to all three types of survey. But most often, we will have in mind a sample survey. Therefore, the term ‘survey’ will usually refer to a ‘sample survey’. We will now review some frequently used survey terminology.

A survey seeks to provide information about a *target population*. The delimitation of the target population must be clearly stated at the planning stage of the survey. The statistician’s interest does not lie in publishing information about individual elements of the target population (such disclosure is often ruled out by law), but in measuring quantities (totals or functions of totals) for aggregates of elements, the whole population or domains. These targeted quantities are called *parameters* or *parameters of interest*. For example, three important objectives of a labour force survey (as conducted in most industrialized countries) are to obtain information about the number of unemployed, the number of employed and the unemployment rate. These are examples of parameters. The first two parameters are *population totals*. The third is a *ratio of population totals*, namely, the number of unemployed persons divided by the total number of persons in the labour force.

Examples of other population parameters are *population means* – for example, mean household income – and *regression coefficients*, say, the regression coefficient of income (dependent variable) regressed on number of years of formal education (independent variable), for a population of individuals. We can estimate any of these parameters with the aid of data on a sample of elements.

The sample is a selection from the *frame population*. The frame population is a list or other device that identifies and represents all elements that could possibly be drawn. Ideally, the frame population represents exactly the set of physically existing elements that make up the target population. In reality, the frame population and the target population differ more or less, as we discuss in more detail later.

Sampling design is used as a generic term for the (usually probabilistic) rule that governs the sample selection. Commonly used sampling designs are: simple random sampling (SI), stratified simple random sampling (STSI), cluster sampling, two-stage sampling, and probability-proportional-to-size (π ps) sampling, of which Poisson sampling is a special case. With the possible exception of SI, these designs require planning before sampling can be carried out. STSI requires a set of well-defined strata. Cluster sampling requires a decision on what clusters to use. Sampling in two or more stages requires a clear definition of the first-stage sampling units, the second-stage units, and so on.

Every sampling design involves two other important general concepts: *inclusion probabilities* and *design weights*. The inclusion probability of an element is the known probability with which it is selected under the given sampling design. The design weight of an element is computed as the inverse of its inclusion probability, assumed to be greater than zero for all elements. Examples of designs where the inclusion probabilities are equal for all elements are SI and STSI with proportional allocation. Many sampling designs used in practice do not give the same inclusion probability to all population elements. In STSI, the inclusion probabilities are equal within strata, but they can differ widely between strata.

The inclusion probability can never exceed one. Consequently, a design weight is greater than or equal to one. The inclusion probability (and the design weight) is equal to one for an element that is selected with certainty. Many business surveys include a number of elements (usually very large elements) that are ‘certainty elements’. These form a subgroup often called a *take-all stratum*.

A majority of the elements have inclusion probabilities strictly less than one. For example, in an STSI design, an element belonging to a stratum from which 200 elements are selected out of a total of 1600 has an inclusion probability equal to the sampling rate in the stratum, $200/1600 = 0.125$, and its design weight is $1/0.125 = 8$. One often heard interpretation is that ‘an element with a design weight equal to 8 represents itself and seven other (non-sampled, non-observed) population elements as well’. When it comes to estimation, the observed value for this element is given the weight 8. Another stratum in the same survey may have 100 sampled elements out of a total of 200. Each element in this stratum has the inclusion probability $100/200 = 0.5$, and its design weight is then $1/0.5 = 2$.

STSI is a widely used design. It is very well suited for *surveys of individuals and households* in countries that can rely on a frame in the form of a total population register (see Example 2.1). Such a register lists the country's population and contains a number of variables suitable for forming strata, such as age, sex and geographical area. It is often of interest to measure households as well as individuals in the same survey. One way to obtain a sample of households from the sample of individuals is to identify the households to which the selected individuals belong. Household variables such as household expenditure can be observed, and variables on individuals, some or all of those residing in an identified household, can also be observed. We can obtain statistics on households as well as statistics on individuals.

The reverse order of selection is also possible. Practical considerations may necessitate drawing first a sample of households, with a specified sampling design, then selecting some or all of the individuals in the selected households. Again, both household variables and variables on individuals can be measured in the same survey. The selection of households can, for example, proceed by drawing a stratified sample of city blocks from a city map and then enumerating all the households in the selected city blocks.

In *business surveys*, the distribution of many variables of interest is highly skewed. The 'industry giants' account for a major share of the total for typical study variables related to production and output. The largest elements (enterprises) must be given a high inclusion probability (probability one or very near to one). Many business surveys use coordinated sampling for small enterprises to distribute the response burden. This entails some control over the frequency with which an enterprise is asked to provide information over a designated period of time, say a year. A number of countries have (to some extent different) systems for coordinated sampling. Statistics Sweden, for example, uses the system referred to as SAMU, described in Atmer *et al.* (1975). Another early reference for coordinated sampling is Brewer *et al.* (1972).

Coordinated sampling techniques are based on the concept of permanent random numbers: a uniformly distributed random number is attached at birth to a statistical element (an enterprise), and it remains with that element for the duration of its life; in that sense, it is permanent. The permanent random numbers play a crucial role in realizing both the desired inclusion probabilities and the desired degree of coordination of samples.

2.2. SOURCES OF ERROR IN A SURVEY

In this section we discuss frames, sampling and nonresponse. Figure 2.1 gives the background.

Coverage errors

We define the *target population* to be the set of elements that the survey aims to encompass at the point in time when the data are collected, by the completion of a

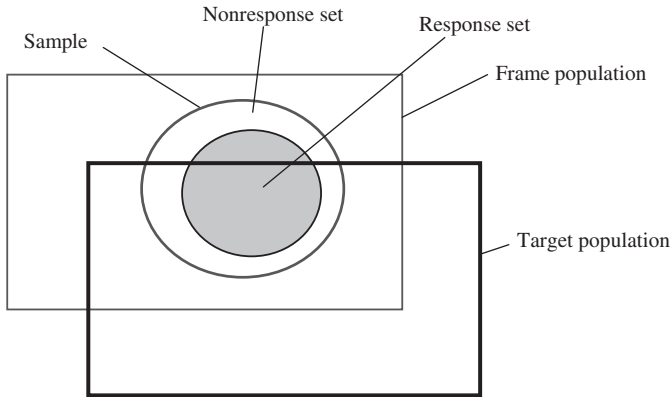


Figure 2.1 Representation of selected sample and response set, seen as subsets of the frame population, and their intersections with the target population.

questionnaire or in some other way. This point in time is called the *reference time point for the target population*. The sampling frame, on the other hand, is usually constructed at an earlier date, sometimes as much as 12 months earlier. This time point is referred to as the *reference time point for the frame population*. The lag between the two time points should be as short as possible, because the risk of coverage errors increases with the time lag. Three types of coverage error are commonly distinguished: *undercoverage*, *overcoverage* and *duplicate listings*. We comment on the first two of these in particular. As the name suggests, duplicate listings refer to errors occurring when a target population element is listed more than once in the frame.

Elements that are in the target population but not in the frame population constitute undercoverage. Especially in business surveys, a significant part of the undercoverage is made up of elements that are new to the target population and therefore absent from the frame population. These are commonly referred to as ‘births’. Undercoverage may have other causes as well.

Elements that are in the frame population but not in the target population constitute overcoverage. Elements that have ceased to exist somewhere between the two reference time points can be a significant source of overcoverage. They are commonly referred to as ‘deaths’.

It follows that undercoverage elements have zero probability of being selected for any sample drawn from the frame population. This is undesirable, because of the bias that can result if the study variable values for the undercoverage elements differ systematically from those of other population elements.

Bias from overcoverage can usually be avoided as long as it is possible to correctly identify the sample elements that belong to the overcoverage. But this is not always possible. For example, elements that are selected into the sample but become nonrespondents can often not be correctly classified as either ‘in the target population’ or ‘in the overcoverage’. Biased estimates can be an undesirable consequence.

Although attempts may be made to minimize the lag between the frame population reference time point and the target population reference time point, the time lag is often considerable. It may be a practical necessity. One reason may be slow updating of the frame. As time goes by, events occur that motivate a change or update of the frame information. An example is a change in a variable value for an element existing in the frame, as when updated information is received about the number of employees or the gross business income of an enterprise. Such changes are sometimes recorded only with considerable delay.

It follows that the values recorded in the frame, at a given point in time and for a specific frame variable, may refer to different points in time for different elements. All elements are not necessarily updated at the same moment in time. This is not ideal, but it is a reality that has to be accepted.

Births and deaths are examples of events that need to be recorded. These events cause a change in the set of elements in the frame.

The frame population for a planned survey is sometimes created from a larger, more extensive collection or list of elements, each having recorded values for a number of variables. A frame population deemed appropriate for a particular survey may then be constructed from this larger collection, using some of the recorded variables to delimit the frame population. Imperfections in the recorded variable values, because of unequal reference times or other causes, may harm the effectiveness of the delimitation.

Imperfect frame variable values are undesirable for other reasons. For example, frame variables are often used before sampling to stratify the population and/or after sampling to poststratify the sample. Errors in the frame variables make these practices less efficient.

Example 2.1. The total population register

A total population register (TPR) exists in all Scandinavian countries (Denmark, Finland, Iceland, Norway and Sweden) and in the Netherlands. Such registers are being developed in other countries. A TPR aims at a complete listing of the country's population. Among the register variables recorded for every person are the person's unique personal identity number, name and address. This makes it possible to access, by mail or otherwise, every person for a wide range of surveys. The addresses are usually classified into administrative regions, such as counties and municipalities. The TPR may have every piece of real estate identified by coordinates, making it possible to construct regions other than counties and municipalities. Other important register variables are date of birth, sex, civil status (married, etc.), country of birth and taxable income. If we take Sweden as an example, information about births, deaths, immigration, emigration and changes of other register variables is received continuously by Statistics Sweden. The register can be kept almost perfectly up to date. Persons arriving from abroad and intending to stay at least 1 year are entered in the register, after the necessary permission has been granted. Since this is not instantaneous, the TPR at any given point in time have a minor degree of undercoverage, consisting of persons who properly belong to the Swedish population but who are not yet entered in the register. The information about births is almost error-free. There

is also some overcoverage, that is, persons in the TPR but not (or no longer) in the Swedish population. This overcoverage, estimated to be around 0.4 % of the entire population, is made up essentially of persons who have emigrated without proper notification for removal from the TPR. Since the personal identity number is unique, duplicates do not occur in the TPR. For STSI, the sizes of strata based on variables such as sex, age group, civil status and many others are determined by a simple count in the TPR. □

Example 2.2. The business register

A business register (BR) that lists enterprises and work establishments (local units) is maintained by the statistical agencies in many countries. Although countries differ in the specifics, many principles are common or similar between countries. Let us again use Sweden as an illustration. Every week, Statistics Sweden's BR receives information from the National Tax Board about births and deaths of enterprises. The births fall into three categories: (i) pure births, that is, enterprises generated by new business activity; (ii) births occurring because of reorganization, as when an existing enterprise is split into several entities; (iii) births arising because of a registration of an enterprise under new legal form. A large percentage of the births may belong to (ii) and (iii). Among the BR variables, two important ones for an enterprise are the Standard Industrial Classification (SIC) code and the number of employees. The information for updating these variables comes from several different sources, specializing in different subsets of the BR. At any given point in time, the most recently recorded variable values do not necessarily refer to the same point in time.

Only at a limited number of times each year is the BR an ideal source for establishing a sampling frame. Those are the times just after the completion of an important updating activity. For updating, the BR unit at Statistics Sweden receives information partly from external sources, partly from its own specially designed surveys. At Statistics Sweden, new sampling frames for business surveys are produced in March, May, August and November, based on the BR as it exists at those different points in time. These frames are of good quality, without being perfect in every regard. □

Sampling error and nonresponse error

When survey statisticians speak about *sampling error*, they mean the error caused by the fact that values of a study variable are solicited only for a sample of elements, not for the whole population. If the whole population were indeed observed, the sampling error would be zero. (There could be other errors, for example, measurement error and nonresponse error, but the sampling error would be zero.) This situation is exceptional. Statisticians often measure 'error' using the concept of variance. Hence, the sampling error is measured by the variance of the estimator in use, *assuming that there are no other errors*.

The usual technique for estimating a population total consists in summing appropriately weighted variable values for the responding elements in the sample. Different weighting systems can come into consideration.

We can use the design weights, given by inverting the inclusion probabilities. If all sampled elements respond (the case of complete response), this gives an unbiased estimator for the population total of the variable in question. This is known as the Horvitz–Thompson (HT) estimator and is discussed further in Section 4.2.

A more sophisticated (and usually more efficient) weighting is the one used in the generalized regression (GREG) estimator, discussed in Section 4.3. For complete response, a GREG point estimate of a population total is also a sum of weighted observed values, in such a way that the weight of an element is the product of the design weight and the *g-weight*. The latter, explained in more detail in Section 4.3, is computed with the aid of the available auxiliary information. *Poststratification* weights are a simple type of *g-weight* familiar to many statisticians.

We have already stated that many parameters of interest in a survey are more complex than a population total or a population mean of a variable. A number of parameters of interest are expressible as a function of two or more population totals. A simple principle is used for estimating a function of totals: replace each unknown population total by its HT estimator or by a GREG estimator. For example, to estimate wages per hour worked (in an industry, for example), we compute the (HT or GREG) estimate of total wages and divide it by the (HT or GREG) estimate of total hours worked.

The *variance* of an estimator is the average of the square of the deviation of the estimator from its central value (mean). This average is with respect to all possible samples that can be drawn using the given sampling design. Since each of these samples has a known probability, determined by the sampling design, we can derive a formula for the variance. It is important to note that variance is measured as ‘variability over all possible samples’. But in practice we never draw all possible samples; we draw just one single sample. Nevertheless, we seek to quantify the variance by a computation based on the data we do have. This is what *variance estimation* sets out to do.

The variance of an estimator is an unknown quantity. It depends on data for the whole population. We *estimate* the variance with the aid of the values observed for the sampled elements. When this is possible the sampling design is said to be *measurable*. The objective is to do this so that the variance estimator is (almost) unbiased.

The estimated standard deviation is defined as the square root of the estimated variance of the estimator. It is used in confidence interval calculation. The familiar procedure for obtaining a confidence interval at (roughly) the 95 % level is to compute the endpoints of the interval as: point estimate plus or minus 1.96 times the estimated standard deviation. It is important that an unbiased point estimate is used. If not, the interval will be off-centre, and it will not carry the desired 95 % confidence.

In the last two decades, computer software has been designed in order to handle estimation in surveys. Such software will typically handle weight computation, point estimation and the corresponding variance estimation. Some software handles the computation of weights for survey data affected by nonresponse. Software

of this general kind, with more or less extensive features, has been produced in a number of countries. It includes BASCULA from the Netherlands (see Nieuwenbroek and Boonstra, 2002), CALMAR and POULPE from France (see Deville *et al.*, 1993; Caron *et al.*, 1998; Caron, 1998), CLAN97 from Sweden (see Andersson and Nordberg, 1998) and GES from Canada (see Estevao *et al.*, 1995). A number of other countries could be mentioned, including Belgium, Italy and the United States. Software development is ongoing. Existing software is updated. New products appear. For particulars about a given software, the reader is referred to the references mentioned and to related material.

Considerable resources are spent in many countries on improving data collection procedures, so as to keep the *nonresponse* as low as possible. Nevertheless, once data collection is concluded, one has to accept some, perhaps even considerable, nonresponse. It causes *nonresponse error* in the estimates. A 20% nonresponse rate is common, and in many surveys it is much higher. Nonresponse rates are on the increase in many surveys and many countries.

For example, for the year 2000, Statistics Sweden reports nonresponse rates between 25% and well over 30% for several surveys, including those relating to consumer buying expectations, income distribution, national and international road goods transport, and transition from upper secondary school to higher education. The last mentioned survey is analysed in detail in Section 14.4.

Internationally, one finds that response rates vary greatly between countries as well as between types of survey. An analysis of patterns in household survey nonresponse is reported in de Leeuw and de Heer (2002). They collected time series data for 10 different surveys from producers of official statistics in 16 developed countries. Not all countries had provided data for each survey. All 16 countries had supplied data for the labour force survey, a majority of them for the family expenditures survey, and some of them for surveys related to health, to travel and to income. The length of the time series varied between countries and between surveys.

The data were carefully analysed by multilevel logistic models. Three dependent variables were response rate, noncontact rate and refusal rate. The latter two are the traditionally recognized major components of nonresponse. The explanatory variables were year (as a fixed effect) and country, survey and year (as random effects). The main results are summarized in the following points. The effects mentioned were found statistically significant.

- Response rates have declined over the years. Countries as well as surveys differ with regard to response rates.
- Noncontact rates have increased over the years, and they differ between countries.
- Refusal rates have increased over the years, and they vary between countries and between surveys.

The noncontact rates were not found to differ significantly between surveys, something which may be explained by a similarity of field work between surveys in one and the same country. The overall noncontact rate was estimated at 6%,

the overall refusal rate at 9%. The overall increase per year in the noncontact rate was estimated at 0.2%, that of the overall noncontact rate at 0.3%.

It is worth noting that the surveys analysed are still, at this point in time, well respected, and generally viewed as being in the public interest. Comparatively speaking, the response rates in those surveys are good. Considerably lower response rates are encountered in many other surveys, in national statistical agencies as in private survey institutes. The theory and the methods in this book, and in important references that we have relied on, may be applied to surveys with 'typical nonresponse rates', considering what the conditions are at this point in time. Nevertheless, the methods in this book are not a panacea; the objective of a survey should always be to achieve as favourable a response rate as possible, under the given constraints of cost and other matters.

For surveys on individuals, the existing vast literature contains much information about nonresponse distributed with respect to basic variables such as age, sex, and region. Experience gathered from these *nonresponse analyses* shows that, for surveys on individuals, lower response rates are usually expected for metropolitan residents, single persons, members of childless households, older persons, divorced or widowed persons, persons with lower educational attainment, and self-employed persons; see Groves and Couper (1998), Holt and Elliot (1991), Lindström (1983). A review of issues in unit nonresponse and item nonresponse for business surveys is given in Willimack *et al.* (2002).

Since variables such as age, sex and region often covary with many social survey study variables, the nonresponding elements are likely to be atypical with respect to these variables. This causes *nonresponse bias* in the estimates, unless proper action can be taken at the estimation stage. A distinct possibility is that some, or even considerable, bias remains in the estimates even after a carefully executed estimation.

Another effect of nonresponse is an *increase in the variance* of estimates, because the effective sample size is reduced. If an increased variance were the only problem, it could be fairly easily counteracted by some degree of 'oversampling', so that the sample size is fixed at the design stage at an appropriately 'higher than normal' rate. The only negative effect of nonresponse might then be some increase in administrative burden and in data collection cost, through higher postage fees, for example.

But it is the nonresponse bias that constitutes by far the most important obstacle to correct statistical conclusions in a survey. Compared to the bias, the increased variance must be considered only a minor disturbance. In the presence of a significant bias, a computed confidence interval will be centred on the wrong value and thus misleading. It does not carry the level of confidence required.

A related disturbance, in the case of a stratified design, is that if the sample is allocated to strata in an optimal fashion, the nonresponse that occurs may cause the responding elements to distribute themselves in a far from optimal manner over the strata. The normally important gains of stratification can be severely compromised by the nonresponse.

Table 2.1 Illustration of data present (marked \times) and data missing (marked nr) in a hypothetical data collection.

Identity	Register variables		Questionnaire variables		
	1	2	1	2	3
1	\times	\times	\times	\times	\times
2	\times	\times	\times	\times	nr
3	\times	\times	\times	nr	\times
4	\times	\times	\times	\times	\times
5	\times	\times	\times	\times	\times
6	\times	\times	nr	\times	nr
7	\times	\times	nr	nr	nr
8	\times	\times	nr	nr	nr

A survey may contain many study variables. In some surveys, it is possible to obtain data on some of these variables from available registers. We call them *register variables*. These sources of information will typically show data on all variables and all elements; no data are missing. However, for the other study variables, data are solicited from a sample selection of elements by telephone or mail or electronically, using a questionnaire or similar data transfer device. The data for these *questionnaire variables* are invariably affected, more or less, by nonresponse. In the following, ‘study variable’ will ordinarily mean ‘questionnaire variable.’ It is customary to distinguish two types of nonresponse: *unit nonresponse* and *item nonresponse*. We shall use the following definitions. A unit nonresponse element is one for which information is missing on all the questionnaire variables. An item nonresponse element is one for which information is missing on at least one, but not all, of the questionnaire variables. The set of elements with a recorded response on at least one questionnaire item will be called the *response set*. These concepts are illustrated by the following example.

Example 2.3. Unit nonresponse, item nonresponse and response set

Table 2.1 illustrates the result of a (hypothetical) data collection in a survey with eight sampled elements. The symbol \times indicates a presence of data, and nr indicates that data are missing. Elements 1, 4 and 5 have complete response. Elements 2, 3 and 6 have some item nonresponse, that is, one or more values are missing among the three questionnaire variables. Although all eight sample elements have data for the two register variables, we shall say that elements 7 and 8 constitute the unit nonresponse, because both of these have no response at all to the questionnaire part of the survey. Elements 1–6, all of which have values recorded for at least one questionnaire item, form the response set in this example. \square

