

The Survey Process

1.1 ABOUT SURVEYS

We live in an information society. There is an ever-growing demand for statistical information about the economic, social, political, and cultural shape of countries. Such information will enable policy makers and others to make informed decisions for a better future. Sometimes, such statistical information can be retrieved from existing sources, for example, administrative records. More often, there is a lack of such sources. Then, a survey is a powerful instrument to collect new statistical information.

A *survey* collects information about a well-defined 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 to the representatives of the elements in the population. To do this in a uniform and consistent way, a questionnaire is used.

One way to obtain information about a population is to collect data about all its elements. Such an investigation is called a *census* or *complete enumeration*. This approach has a number of disadvantages:

- It is very expensive. Investigating a large population involves a lot of people (e.g., interviewers) and other resources.
- It is very time-consuming. Collecting and processing a large amount of data takes time. This affects the timeliness of the results. Less timely information is less useful.
- Large investigations increase the response burden on people. As many people are more frequently asked to participate, they will experience it more and more as a burden. Therefore, people will be less and less inclined to cooperate.

A survey is a solution to many of the problems of a census. Surveys collect information on only a small part of the population. This small part is called the *sample*.

In principle, the sample provides information only on the sampled elements of the population. No information will be obtained 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 using probability sampling. A random selection procedure uses an element of chance to determine which elements are selected, and which are not. If it is clear how this selection mechanism works and it is possible to compute the probabilities of being selected in the sample, survey results allow making reliable and precise statements about the population as a whole.

At first sight, the idea of introducing an element of uncertainty in an investigation seems odd. It looks like magic that it is possible to say something about a complete population by investigating only a small randomly selected part of it. However, there is no magic about sample surveys. There is a well-founded theoretical framework underlying survey research. This framework will be described in this book.

1.2 A SURVEY, STEP-BY-STEP

Carrying out a survey is often a complex process that requires careful consideration and decision making. This section gives a global overview of the various steps in the process, the problems that may be encountered, and the decisions that have to be made. The rest of the book describes these steps in much more detail. Figure 1.1 shows the steps in the survey process.

The first step in the survey process is *survey design*. Before data collection can start, a number of important decisions have to be made. First, it has to become clear which population will be investigated (the *target population*). Consequently, this is the population to which the conclusions apply. Next, the general research questions must

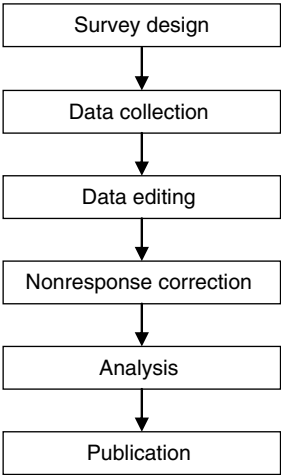


Figure 1.1 The survey process.

be translated into specification of population characteristics to be estimated. This specification determines the contents of the questionnaire. Furthermore, to select a proper sample, a sampling design must be defined, and the sample size must be determined such that the required accuracy of the results can be obtained.

The second step in the process is data collection. Traditionally, in many surveys paper questionnaires were used. They could be completed in face-to-face interviews: interviewers visited respondents, asked questions, and recorded the answers on (paper) forms. The quality of the collected data tended to be good. However, since face-to-face interviewing typically requires a large number of interviewers, who all may have to do much traveling, it was expensive and time-consuming. Therefore, telephone interviewing was often used as an alternative. The interviewers called the respondents from the survey agency, and thus no more traveling was necessary. However, telephone interviewing is not always feasible: only connected (or listed) people can be contacted, and the questionnaire should not be too long or too complicated. A mail survey was cheaper still: no interviewers at all were needed. Questionnaires were mailed to potential respondents with the request to return the completed forms to the survey agency. Although reminders could be sent, the persuasive power of the interviewers was lacking, and therefore response tended to be lower in this type of survey, and so was the quality of the collected data.

Nowadays paper questionnaires are often replaced with electronic ones. *Computer-assisted interviewing* (CAI) allows to speed up the survey process, improve the quality of the collected data, and simplify the work of the interviewers. In addition, computer-assisted interviewing comes in three forms: *computer-assisted personal interviewing* (CAPI), *computer-assisted telephone interviewing* (CATI), and *computer-assisted self-interviewing* (CASI). More and more, the Internet is used for completing survey questionnaires. This is called *computer-assisted web interviewing* (CAWI). It can be seen as a special case of CASI.

Particularly if the data are collected by means of paper questionnaire forms, the completed questionnaires have to undergo extensive treatment. To produce high-quality statistics, it is vital to remove any error. This step of the survey process is called *data editing*. Three types of errors can be distinguished: A *range error* occurs if a given answer is outside the valid domain of answers; for example, a person with an age of 348 years. A *consistency error* indicates an inconsistency in the answers to a set of questions. An age of 8 years may be valid, a marital status “married” is not uncommon, but if both answers are given by the same person, there is something definitely wrong. The third type of error is a *routing error*. This type of error occurs if interviewers or respondents fail to follow the specified branch or skip instructions; that is, the route through the questionnaire is incorrect: irrelevant questions are answered, or relevant questions are left unanswered.

Detected errors have to be corrected, but this can be very difficult if it has to be done afterward, at the survey agency. In many cases, particularly for household surveys, respondents cannot be contacted again, so other ways have to be found out to solve the problem. Sometimes, it is possible to determine a reasonable approximation of a correct value by means of an *imputation* procedure, but in other cases an incorrect value is replaced with the special code indicating the value is “unknown.”

After data editing, the result is a “clean” data file, that is, a data file in which no errors can be detected any more. However, this file is not yet ready for analysis. The collected data may not be representative of the population because the sample is affected by nonresponse; that is, for some elements in the sample, the required information is not obtained. If nonrespondents behave differently with respect to the population characteristics to be investigated, the results will be biased. To correct for unequal selection probabilities and nonresponse, a *weighting adjustment* procedure is often carried out. Every record is assigned some weight. These weights are computed in such a way that the weighted sample distribution of characteristics such as gender, age, marital status, and region reflects the known distribution of these characteristics in the population.

In the case of item nonresponse, that is, answers are missing on some questions, not all questions, an *imputation* procedure can also be carried out. Using some kind of model, an estimate for a missing value is computed and substituted in the record.

Finally, a data file is obtained that is ready for analysis. The first step in the analysis will probably nearly always be tabulation of the basic characteristics. Next, a more extensive analysis will be carried out. Depending on the nature of the study, this will take the form of an exploratory analysis or an inductive analysis. An *exploratory analysis* will be carried out if there are no preset ideas, and the aim is to detect possibly existing patterns, structures, and relationships in the collected data. To make inference on the population as a whole, an *inductive analysis* can be carried out. This can take the form of estimation of population characteristics or the testing of hypotheses that have been formulated about the population.

The survey results will be published in some kind of report. On the one hand, this report must present the results of the study in a form that makes them readable for nonexperts in the field of survey research. On the other hand, the report must contain a sufficient amount of information for experts to establish whether the study was carried out properly and to assess the validity of the conclusions.

Carrying out a survey is a time-consuming and expensive way of collecting information. If done well, the reward is a data file full of valuable information. It is not unlikely that other researchers may want to use these data in additional analysis. This brings up the question of protecting the privacy of the participants in the survey. Is it possible to disseminate survey data sets without revealing sensitive information of individuals? Disclosure control techniques help establish disclosure risks and protect data sets against disclosing such sensitive information.

1.3 SOME HISTORY OF SURVEY RESEARCH

The idea of compiling statistical overviews of the state of affairs in a country is already very old. As far back as Babylonian times, censuses of agriculture were taken. This happened fairly shortly after the art of writing was invented. Ancient China counted its people to determine the revenues and the military strength of its provinces. There are also accounts of statistical overviews compiled by Egyptian rulers long before Christ. Rome regularly took a census of people and of property. The data were used to establish

the political status of citizens and to assess their military and tax obligations to the state. And of course, there was numbering of the people of Israel, leading to the birth of Jesus in the small town of Bethlehem.

In the Middle Ages, censuses were rare. The most famous one was the census of England taken by the order of William the Conqueror, King of England. The compilation of this *Domesday Book* started in the year 1086 AD. The book records a wealth of information about each manor and each village in the country. There is information about more than 13,000 places, and on each county there are more than 10,000 facts. To collect all these data, the country was divided into a number of regions, and in each region, a group of commissioners was appointed from among the greater lords. Each county within a region was dealt with separately. Sessions were held in each county town. The commissioners summoned all those required to appear before them. They had prepared a standard list of questions. For example, there were questions about the owner of the manor, the number of free men and slaves, the area of woodland, pasture, and meadow, the number of mills and fishponds, to the total value, and the prospects of getting more profit. The *Domesday Book* still exists, and county data files are available on CD-ROM or the Internet.

Another interesting example can be found in the Inca Empire that existed between 1000 and 1500 AD in South America. Each Inca tribe had its own statistician, called *Quipucamayoc* (Fig. 1.2). This man kept records of, for example, the number of people, the number of houses, the number of llamas, the number of marriages, and the number of young men who could be recruited to the army. All these facts were recorded on a *quipu*, a system of knots in colored ropes. A decimal system was used for this.



Figure 1.2 The Quipucamayoc, the Inca statistician. Reprinted by permission of ThiemeMeulenhoff.

At regular intervals, couriers brought the quipus to Cusco, the capital of the kingdom, where all regional statistics were compiled into national statistics. The system of Quipucamayocs and quipus worked remarkably well. Unfortunately, the system vanished with the fall of the empire.

An early census also took place in Canada in 1666. Jean Talon, the intendant of New France, ordered an official census of the colony to measure the increase in population since the founding of Quebec in 1608. The enumeration, which recorded a total of 3215 people, included the name, age, gender, marital status, and occupation of every person. The first censuses in Europe were undertaken by the Nordic countries: The first census in Sweden–Finland took place in 1746. It had already been suggested earlier, but the initiative was rejected because “it corresponded to the attempt of King David who wanted to count his people.”

The first known attempt to make statements about a population using only information about part of it was made by the English merchant John Graunt (1620–1674). In his famous tract, Graunt describes a method to estimate the population of London on the basis of partial information (Graunt, 1662). Graunt surveyed families in a sample of parishes where the registers were well kept. He found that on average there were 3 burials per year in 11 families. Assuming this ratio to be more or less constant for all parishes, and knowing the total number of burials per year in London to be about 13,000, he concluded that the total number of families was approximately 48,000. Putting the average family size at 8, he estimated the population of London to be 384,000. Although Graunt was aware of the fact that averages such as the number of burials per family varied in space and time, he did not make any provisions for this phenomenon. Lacking a proper scientific foundation for his method, John Graunt could not make any statement about the accuracy of his method.

Another survey-like method was applied more than a century later. Pierre Simon Laplace (1749–1827) realized that it was important to have some indication of the accuracy of the estimate of the French population. Laplace (1812) implemented an approach that was more or less similar to that of John Graunt. He selected 30 departments distributed over the area of France. Two criteria controlled the selection process. First, he saw to it that all types of climate were represented. In this way, he could compensate for climate effects. Second, he selected departments for which the mayors of the communes could provide accurate information. By using the central limit theorem, he proved that his estimator had a normal distribution. Unfortunately, he overlooked the fact that he used a cluster sample instead of a simple random sample, and moreover communes were selected within departments purposively, and not at random. These problems made the application of the central limit theorem at least doubtful. The work of Laplace was buried in oblivion in the course of the nineteenth century.

In the period until the late 1880s, there were many *partial investigations*. These were statistical inquiries in which only a part of human population was investigated. The selection from the population came to hand incidentally, or was made specifically for the investigation. In general, the selection mechanism was unclear and undocumented. While by that time considerable progress had already been made in the areas of probability theory and mathematical statistics, little attention was paid to applying

these theoretical developments to survey sampling. Nevertheless, gradually probability theory found its way in official statistics. An important role was played by the Dutch/Belgian scientist, Lambert Adolphe Jacques Quetelet (1796–1874). He was involved in the first attempt in 1826 to establish The Netherlands Central Bureau of Statistics. In 1830, Belgium separated from The Netherlands, and Quetelet continued his work in Belgium.

Quetelet was the supervisor of statistics for Belgium (from 1830), and in this position, he developed many of the rules governing modern census taking. He also stimulated statistical activities in other countries. The Belgian census of 1846, directed by him, has been claimed to be the most influential in its time because it introduced careful analysis and critical evaluation of the data compiled. Quetelet dealt only with censuses and did not carry out any partial investigations.

According to Quetelet, many physical and moral data have a natural variability. This variability can be described by a normal distribution around a fixed, true value. He assumed the existence of something called the *true value*. He proved that this true value could be estimated by taking the mean of a number of observations. Quetelet introduced the concept of *average man* (“l’homme moyenne”) as a person of which all characteristics were equal to the true value. For more information, see Quetelet (1835, 1846).

In the second half of the nineteenth century, so-called *monograph studies* or surveys became popular. They were based on Quetelet’s idea of the average man (see Desrosières, 1998). According to this idea, it suffices to collect information only on typical people. Investigation of extreme people was avoided. This type of inquiry was still applied widely at the beginning of the twentieth century. It was an “officially” accepted method.

Industrial revolution was also an important era in the history of statistics. It brought about drastic and extensive changes in society, as well as in science and technology. Among many other things, urbanization started from industrialization, and also democratization and the emerging social movements at the end of the industrial revolution created new statistical demands. The rise of statistical thinking originated partly from the demands of society and partly from work and innovations of men such as Quetelet. In this period, the foundations for many principles of modern social statistics were laid. Several central statistical bureaus, statistical societies, conferences, and journals were established soon after this period.

The development of modern sampling theory started around the year 1895. In that year, Anders Kiaer (1895, 1997), the founder and first director of Statistics Norway, published his *Representative Method*. It was a partial inquiry in which a large number of persons were questioned. This selection should form a “miniature” of the population. Persons were selected arbitrarily but according to some rational scheme based on general results of previous investigations. Kiaer stressed the importance of *representativeness*. His argument was that if a sample was representative with respect to variables for which the population distribution was known, it would also be representative with respect to the other survey variables.

Kiaer was way ahead of his time with ideas about survey sampling. This becomes clear in the reactions on the paper he presented at a meeting of the International

Statistical Institute in Bern in 1895. The last sentence of a lengthy comment by the influential Bavarian statistician von Mayr almost became a catch phrase: “Il faut rester ferme et dire: pas de calcul là où l’observation peut être faite.” The Italian statistician Bodio supported von Mayr’s views. The Austrian statistician Rauchberg said that further discussion of the matter was unnecessary. And the Swiss statistician Milliet demanded that such incomplete surveys should not be granted a status equal to “la statistique serieuse.”

A basic problem of the representative method was that there was no way of establishing the accuracy of estimates. The method lacked a formal theory of inference. It was Bowley (1906) who made the first steps in this direction. He showed that for large samples, selected at random from the population, the estimate had an approximately normal distribution.

From this moment on, there were two methods of sample selection. The first one was Kiaer’s representative method, based on purposive selection, in which representativeness played a crucial role, and for which no measure of the accuracy of the estimates could be obtained. The second was Bowley’s approach, based on simple random sampling, for which an indication of the accuracy of estimates could be computed. Both methods existed side by side for a number of years. This situation lasted until 1934, when the Polish scientist Jerzy Neyman published his now famous paper (see Neyman, 1934). Neyman developed a new theory based on the concept of the confidence interval. By using random selection instead of purposive selection, there was no need any more to make prior assumptions about the population.

Neyman’s contribution was not restricted to the confidence interval that he invented. By making an empirical evaluation of Italian census data, he could prove that the representative method based on purposive sampling failed to provide satisfactory estimates of population characteristics. The result of Neyman’s evaluation of purposive sampling was that the method fell into disrepute in official statistics.

Random selection became an essential element of survey sampling. Although theoretically very attractive, it was not very simple to realize this in practical situations. How to randomly select a sample of thousands of persons from a population of several millions? How to generate thousands of random numbers? To avoid this problem, often systematic samples were selected. Using a list of elements in the population, a starting point and a step size were specified. By stepping through this list from the starting point, elements were selected. Provided the order of the elements is more or less arbitrary, this systematic selection resembles random selection. W.G. and L.H. Madow made the first theoretical study of the precision of systematic sampling only in 1944 (see Madow and Madow, 1944). The use of the first tables of random numbers published by Tippet (1927) also made it easier to select real random samples.

In 1943, Hansen and Hurvitz published their theory of multistage samples. According to their theory, in the first stage, primary sampling units are selected with probabilities proportional to their size. Within selected primary units, a fixed number of secondary units are selected. This proved to be a useful extension of the survey sampling theory. On the one hand, this approach guaranteed every secondary unit to have the same probability of selection in the sample, and on the other, the

sampled units were distributed over the population in such a way that the fieldwork could be carried out efficiently.

The classical theory of survey sampling was more or less completed in 1952. Horvitz and Thompson (1952) developed a general theory for constructing unbiased estimates. Whatever the selection probabilities are, as long as they are known and positive, it is always possible to construct a reliable estimate. Horvitz and Thompson completed the classical theory, and the random sampling approach was almost unanimously accepted. Most of the classical books about sampling were also published by then: Cochran (1953), Deming (1950), Hansen et al. (1953), and Yates (1949).

Official statistics was not the only area where sampling was introduced. Opinion polls can be seen as a special type of sample surveys, in which attitudes or opinions of a group of people are measured on political, economic, or social topics. The history of opinion polls in the United States goes back to 1824, when two newspapers, the *Harrisburg Pennsylvanian* and the *Raleigh Star*, attempted to determine political preferences of voters before the presidential election. The early polls did not pay much attention to sampling. Therefore, it was difficult to establish the accuracy of results. Such opinion polls were often called *straw polls*. This expression goes back to rural America. Farmers would throw a handful of straws into the air to see which way the wind was blowing. In the 1820s, newspapers began doing straw polls in the streets to see how political winds blew.

It took until the 1920s before more attention was paid to sampling aspects. At that time, Archibald Crossley developed new techniques for measuring American public's radio listening habits. And George Gallup worked out new ways to assess reader interest in newspaper articles (see, for example, Linehard, 2003). The sampling technique used by Gallup was *quota sampling*. The idea was to investigate groups of people who were representative for the population. Gallup sent out hundreds of interviewers across the country. Each interviewer was given quota for different types of respondents: so many middle-class urban women, so many lower class rural men, and so on. In total, approximately 3000 interviews were carried out for a survey.

Gallup's approach was in great contrast with that of the *Literary Digest* magazine, which was at that time the leading polling organization. This magazine conducted regular "America Speaks" polls. It based its predictions on returned ballot forms that were sent to addresses obtained from telephone directories and automobile registration lists. The sample size for these polls was very large, something like 2 million people.

The presidential election of 1936 turned out to be decisive for both approaches (see Utts, 1999). Gallup correctly predicted Franklin Roosevelt to be the new President, whereas *Literary Digest* predicted that Alf Landon would beat Franklin Roosevelt. How could a prediction based on such a large sample be so wrong? The explanation was a fatal flaw in the *Literary Digest*'s sampling mechanism. The automobile registration lists and telephone directories were not representative samples. In the 1930s, cars and telephones were typically owned by the middle and upper classes. More well-to-do Americans tended to vote Republican and the less well-to-do were inclined to vote Democrat. Therefore, Republicans were overrepresented in the *Literary Digest* sample.

As a result of this historic mistake, the *Literary Digest* magazine ceased publication in 1937. And opinion researchers learned that they should rely on more scientific ways of sample selection. They also learned that the way a sample is selected is more important than the size of the sample.

1.4 THIS BOOK

This book deals with the theoretical and practical aspects of sample survey sampling. It follows the steps in the survey process described in Section 1.1.

Chapter 2 deals with various aspects related to the design of a survey. Basic concepts are introduced, such as population, population parameters, sampling, sampling frame, and estimation. It introduces the Horvitz–Thompson estimator as the basis for estimation under different sampling designs.

Chapter 3 is devoted to questionnaire designing. It shows the vital importance of properly defined questions. It also discusses various question types, routing (branching and skipping) in the questionnaire, and testing of questionnaires.

Chapters 4 and 5 describe a number of sampling designs in more detail. Chapter 3 starts with some simple sampling designs: simple random sampling, systematic sampling, unequal probability sampling, and systematic sampling with unequal probabilities. Chapter 4 continues with composite sampling designs: stratified sampling, cluster sampling, two-stage sampling, and sampling in space and time.

Chapter 6 presents a general framework for estimation. Starting point is a linear model that explains the target variable of a survey from one or more auxiliary variables. Some well-known estimators, such as the ratio estimator, the regression estimator, and the poststratification estimator, emerge as special cases of this model.

Chapter 7 is about data collection. It compares traditional data collection with paper questionnaire forms with computer-assisted data collection. Advantages and disadvantages of various modes of data collection are discussed. To give some insight into the attractive properties of computer-assisted interviewing, a software package is described that can be seen as the de facto standard for CAI in official statistics. It is the Blaise system.

Chapter 8 is devoted to the quality aspects. Collected survey data always contain errors. This chapter presents a classification of things that can go wrong. Errors can have a serious impact on the reliability of survey results. Therefore, extensive error checking must be carried out. It is also shown that correction of errors is not always simple. Imputation is discussed as one of the error correction techniques.

Nonresponse is one of the most important problems in survey research. Nonresponse can cause survey estimates to be seriously biased. Chapter 9 describes the causes of nonresponse. It also incorporates this phenomenon in sampling theory, thereby showing what the effects of nonresponse can be. Usually, it is not possible to avoid nonresponse in surveys. This calls for techniques that attempt to correct the negative effect of nonresponse. Two approaches are discussed in this chapter: the follow-up survey and the Basic Question Approach.

Adjustment weighting is one of the most important nonresponse correction techniques. This technique assigns weights to responding elements. Overrepresented groups get a small weight and underrepresented groups get a large weight. Therefore, the weighted sample becomes more representative for the population, and the estimates based on weighted data have a smaller bias than estimates based on unweighted data. Several adjustment weighting techniques are discussed in Chapter 10. The simplest one is poststratification. Linear weighting and multiplicative weighting are techniques that can be applied when poststratification is not possible.

Chapter 11 is devoted to online surveys. They become more and more popular, because such surveys are relatively cheap and fast. Also, it is relatively simple to obtain cooperation from large groups of people. However, there are also serious methodological problems. These are discussed in this chapter.

Chapter 12 is about the analysis of survey data. Due to their special nature (unequal selection probabilities, error correction with imputation, and nonresponse correction by adjustment weighting), analysis of such data is not straightforward. Standard software for statistical analysis may not interpret these data correctly. Therefore, analysis techniques may produce wrong results. Some issues are discussed in this chapter. Also, attention is paid to the publication of survey results. In particular, the advantages and disadvantages of the use of graphs in publications are described.

The final chapter is devoted to statistical disclosure control. It is shown how large the risks of disclosing sensitive information can be. Some techniques are presented to estimate these risks. It becomes clear that it is not easy to reduce the risks without affecting the amount of information in the survey data.

1.5 SAMPLONIA

Examples will be used extensively in this book to illustrate concepts from survey theory. To keep these examples simple and clear, they are all taken from an artificial data set. The small country of Samplonia has been created, and a file with data for all inhabitants has been generated (see Fig. 1.3). Almost all examples of sampling designs and estimation procedures are based on data taken from this population file.

Samplonia is a small, independent island with a population of 1000 souls. A mountain range splits the country into the northern province of Agria and the southern province of Induston. Agria is rural province with mainly agricultural activities. The province has three districts. Wheaton is the major supplier of vegetables, potatoes, and fruits. Greenham is known for growing cattle. Newbay is a fairly new area that is still under development. Particularly, young farmers from Wheaton and Greenham attempt to start a new life here.

The other province, Induston, is for a large part an industrial area. There are four districts. Smokeley and Mudwater have a lot of industrial activity. Crowdon is a commuter district. Many of its inhabitants work in Smokeley and Mudwater. The small district of Oakdale is situated in the woods near the mountains. This is where the rich and retired live.

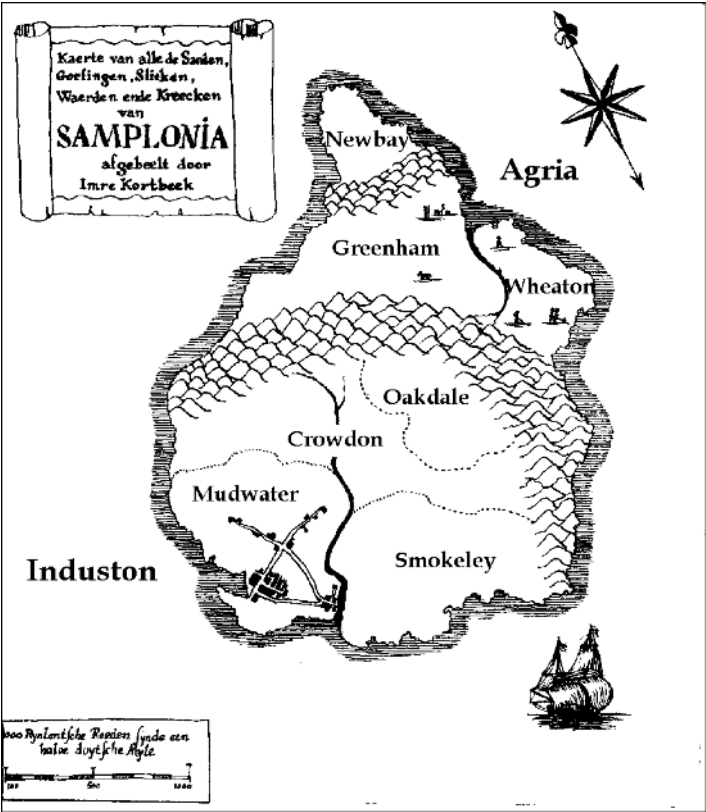


Figure 1.3 The country of Samplonia. Reprinted by permission of Imre Kortbeek.

Samplonia has a central population register. This register contains information such as district of residence, age, and gender for each inhabitant. Other variables that will be used are employment status (has or does not have a job) and income (in Samplonian dollars). Table 1.1 contains the population distribution. Using an

Table 1.1 The Population of Samplonia by Province and District	
Province/District	Inhabitants
Agria	293
Wheaton	144
Greenham	94
Newbay	55
Induston	707
Oakdale	61
Smokeley	244
Crowdon	147
Mudwater	255
Total	1000

Table 1.2 Milk Production by Dairy Farms in Samplonia

	Mean	Standard Deviation	Minimum	Maximum
Milk production (liters per day)	723.5	251.9	10.0	1875.0
Area of grassland (hectares)	11.4	2.8	4.0	22.0
Number of cows	28.9	9.0	8.0	67.0

artificial data file has the advantage that all population data are exactly known. Therefore, it is possible to compare computed estimates with true population figures. The result of such a comparison will make clear how well an estimation procedure performs.

Some survey techniques are illustrated by using another artificial example. There are 200 dairy farms in the rural part of Samplonia. Surveys are regularly conducted with as objective estimation of the average daily milk production per farm. There is a register containing the number of cows and the total area of grassland for each farm. Table 1.2 summarizes these variables.

Included in the book is the software package SimSam. This is a program for simulating samples from finite populations. By repeating the selection of a sample and the computation of an estimate a large number of times, the distribution of the estimates can be characterized in both graphical and numerical ways. SimSam can be used to simulate samples from the population of Samplonia. It supports several of the sampling designs and estimation procedures used in this book. It is a useful tool to illustrate the behavior of various sampling strategies. Moreover, it is also possible to generate nonresponse in the samples. Thus, the effect of nonresponse on estimation procedures can be studied.

EXERCISES

- 1.1** The last census in The Netherlands took place in 1971. One of the reasons to stop it was the concern about a possible refusal of a substantial group of people to participate. Another was that a large amount of information could be obtained from other sources, such as population registers. Which of statements below about a census is correct?
- a.** In fact, a census is a sample survey, because there are always people who refuse to cooperate.
 - b.** A census is not a form of statistical research because the collected data are used only for administrative purposes.
 - c.** A census is a complete enumeration of the population because, in principle, every member of the population is asked to provide information.
 - d.** The first census was carried out by John Graunt in England around 1662.

- 1.2** The authorities in the district of Oakdale want to know how satisfied the citizens are with the new public swimming pool. It is decided to carry out a survey. What would be the group of people to be sampled?
- a.** All inhabitants of Oakdale.
 - b.** All adult inhabitants of Oakdale.
 - c.** All inhabitants of Oakdale who have visited the swimming pool in a specific week.
 - d.** All inhabitants of Oakdale who have an annual season ticket.
- 1.3** No samples were selected by national statistical offices until the year 1895. Before that data collection was mainly based on complete enumeration. Why did they not use sampling techniques?
- a.** The idea of investigating just a part of the population had not yet emerged.
 - b.** They considered it improper to replace real data by mathematical manipulations.
 - c.** Probability theory had not been invented yet.
 - d.** National statistical offices did not yet exist.
- 1.4** Arthur Bowley suggested in 1906 to use random sampling to select a sample from a population. Why was this idea so important?
- a.** It made it possible to introduce the “average man” (“l’homme moyenne”) in statistics.
 - b.** It was not important because it is too difficult to select probability samples in practice.
 - c.** It made it possible to carry out partial investigations.
 - d.** It made it possible to apply probability theory to determine characteristics of estimates.
- 1.5** Why could Gallup provide a better prediction of the outcome of the 1936 Presidential election than the poll of the *Literary Digest* magazine?
- a.** Gallup used automobile registration lists and telephone directories.
 - b.** Gallup used a much larger sample than *Literary Digest* magazine.
 - c.** Gallup used quota sampling, which resulted in a more representative sample.
 - d.** Gallup interviewed people only by telephone.