Chapter 1

INTRODUCTION TO SURVEY SAMPLING

Sample surveys are widely used to provide statistical data on an extensive range of topics for both research and administrative purposes. Numerous surveys have been conducted in disciplines such as sociology, social psychology, demography, political science, economics, education, and public health to describe the characteristics of the populations studied and to develop, test, and refine research hypotheses. Central governments make considerable use of surveys to gather information about the conditions of their populations in terms of employment and unemployment, income and expenditure, housing conditions, education, nutrition, health, travel patterns, and many other subjects. Local governments equally make use of surveys for local planning purposes. Market researchers carry out surveys to identify markets for products, to discover how products are used and how they perform in practice, and to determine consumer reactions. Opinion polls keep track of the popularity of political leaders and their parties, and they measure public opinion on a variety of topical issues. The methods of survey research are also widely used in evaluation studies that aim to assess the impact of government programs. The focus of this book is on methods for sampling human populations. Central governments and other organizations also conduct many surveys of organizations such as manufacturers, retail outlets, farms, schools, and hospitals. While the general sampling approaches for surveys of businesses and other such organizations are similar to those presented here, this text does not address the special techniques used for these types of surveys.

The history of the sample survey as we know it today dates mainly from the early 1900s, and much of the growth in the use of surveys has occurred since the 1930s. At the beginning of the 20th century, statisticians were debating whether anything less than a complete enumeration of a population would suffice, given that this was feasible in principle (Kalton, 2019). When some form of "representative" sampling was accepted, the debate turned to the issue of whether the sample should be obtained by a random or a purposive selection. Then, in 1934, came Neyman's seminal paper that led to the almost universal preference for probability sampling for random selection over purposive selection for the surveys conducted by national statistical agencies and for other major surveys. In that paper, Neyman (1934) also developed a mode of statistical inference for sample surveys using probability sampling—known as *design-based inference*—that is

different from the model-dependent mode of inference used in other fields of statistics. Since that time, probability sampling and design-based inference have become widely accepted for survey research, and an impressive array of sampling methods has been devised to enable efficient and economic samples to be drawn in a variety of practical settings. Among the most widely used methods are systematic sampling, stratification, multistage (cluster) sampling, and probability proportional to size sampling. Initially, for ease of exposition, these methods are discussed in separate chapters of this book, but in practice, they are employed together in what often become complex sample designs.

Although this book is concerned only with issues related to sample design, the sample design must be developed as an integral part of the overall survey design. Survey design involves many interrelated decisions. In addition to sample design, decisions need to be made on such factors as the mode of data collection (e.g., by face-to-face interview, telephone interview, mail questionnaire, or web questionnaire), the framing of the questions to be asked and if any other data need to be collected, and the method of processing the data (see, for instance, Groves et al., 2009). In particular, the economics involved in the data collection process exert a strong influence on the choice of sample design.

A key distinctive aspect of survey sampling is the concept of the finite population of inference. The starting point in survey sample design is to define the population to be studied. Here the term *population* refers to the set of the elements under study, where the "elements" are the units of analysis. The elements may be persons, but they could alternatively be households, farms, schools, or any other unit. The population definition needs to be specified precisely and carefully according to the survey objectives because the survey results will depend on the definition adopted. Consider, for instance, a survey to be carried out in a city to discover the degree of support for the introduction of a new bus system. Should the survey be confined to persons living within the city boundaries? What is the minimum age for the members of the population to be surveyed? Should residents ineligible to vote in city elections be included? Should visitors living temporarily in the city be excluded? If so, how are they to be defined? A number of questions like these arise in defining most populations, making the definitional task less straightforward than it might appear at first.

It is a useful exercise to start by defining the *target population* as the ideal one required to meet the survey objectives. In practice, this definition is then often modified to the *survey population* to take account of practical constraints. For instance, many national surveys in the United States would ideally include some of the following: service members based abroad, people living in hospitals, hotels, prisons, army barracks, and other institutions,

and homeless people. However, the severe problems involved in collecting responses from such persons generally lead to their exclusion from the survey population. The advantage of starting with the ideal target population is that the exclusions are explicitly identified, thus enabling the magnitude and consequences of the restrictions to be assessed.

Once the population has been defined, the question of taking a sample from it can be addressed. One possibility, of course, is to take a complete enumeration of all the elements in the population, but this is seldom appropriate. To collect data from only a part of the population is clearly less costly and, provided that the estimates are sufficiently precise, sampling is thus preferable. A sample inquiry can also be conducted and processed faster, leading to timelier reporting. Furthermore, by concentrating resources on only a part of the population, the quality of the data collected may be superior to that collected in a complete enumeration. As a result, even though a sample survey includes only part of the population, it may, in fact, produce more accurate results than a full census. For these reasons, unless the population is small, sampling is almost always used.

A sample is drawn from the finite survey population, the survey data are collected from the sample, and these data are analyzed to make inferences about characteristics of that specific population at that given point of time. These characteristics are often descriptive statistics, such as the average income of wage earners, the proportion of the population living in poverty, and the numbers and proportions of people in different age groups who are obese. However, they could also include analytic statistics, such as regression and correlation coefficients. The statistics computed from the sample serve as estimates of the population parameters that would have been obtained had the population been fully enumerated.

The subject of sample design is concerned with how to select the part of the population to be included in the survey. A basic distinction to be made is whether the sample is selected by a probability mechanism or not. With a probability sample, each element has a known, nonzero chance of being included in the sample. Consequently, if the sample is selected from the full target population and all sampled elements respond, selection biases are avoided, and survey sampling theory can be used to derive properties of the survey estimators.

Nonprobability sampling covers a variety of procedures, including the use of volunteers, the purposive choice of a "representative" sample of elements, and a range of other methods described in Chapter 15. The weakness of all nonprobability sampling methods is their subjectivity, which precludes the development of a theoretical framework that is free of model assumptions. While this book focuses mainly on probability sampling methods, in recent years the use of nonprobability sampling methods has

increased markedly for several reasons, including the falling response rates that are experienced with probability samples and the large nonprobability samples that can be generated very rapidly and at low cost via the Internet.

The great attraction of design-based inference is that, with a perfectly executed probability sample, statistical inferences about the parameters of the finite population are free of model assumptions. The finite population of inference is the key feature that makes design-based inference with probability sampling different from other modes of statistical inference. Design-based inference assesses the quality of a sample estimate of a finite population parameter by estimating the variability of estimates that could have been generated over all possible samples with the given sample design and estimation method. In design-based inference, the survey variables are fixed values for the elements in the finite population, and the randomness is caused by whether the elements are selected for the sample. If a census of all the population elements is taken, the survey estimates are exactly equal to the finite population parameters. In contrast, in general statistical inference, or model-based inference, the data are assumed to have come from a sample selected randomly and independently from an infinite population represented by some probability distribution, and the survey variables are treated as the random variables. The survey estimates from a census of the finite population are still subject to sampling error for estimating the characteristics of the infinite population.

For reasons of cost efficiency, the sampled elements in survey sample designs are not selected independently of each other, and they are often selected with unequal probabilities. Among the many possible survey sample designs, only in the case of simple random sampling with replacement does a sample design conform to the assumptions made in general statistical inference (see Chapter 2); moreover, that design is seldom, if ever, used in practice.

In practice, survey samples suffer from imperfections, and these imperfections necessarily cause some model dependence. For example, consider the case where 30% of the sampled elements are nonrespondents. Analyses of the respondent data without regard for the nonresponse will provide valid estimates of population parameters only if a model that assumes that the nonrespondents are missing completely at random (MCAR) holds in practice. As discussed in Chapter 11, the MCAR model is rarely realistic, and more realistic models are generally used. It remains the case, however, that the validity of the survey estimates depends on the chosen model. The decline in survey response rates that has been experienced over the past two to three decades means that the degree of model dependence is much higher now than it was in the past.

An essential requirement for any form of probability sampling from a finite population is the existence of a *sampling frame* from which the sampled elements can be selected. In a simple case, when a list of all the population elements is available, the list may be used as the frame. When there is no list, the frame is some equivalent procedure for identifying the elements in the population. Area sampling provides a good illustration of such a frame. With this technique, each element of the population is associated with a particular geographical area (e.g., people or households are associated with the area of their residence, or main residence if they have more than one). Then a sample of areas is drawn, and the sample of elements is drawn only from within the selected areas. The sample of elements in a selected area may be drawn from an existing element frame for the area, if available, or from a list of elements constructed by a listing operation in that area (see Chapters 8 and 14 for examples).

The general organization of the sampling frame and the information the frame contains about the population elements often exert a strong influence on the choice of sample design. In assessing a potential sampling frame, the researcher needs to determine its suitability for the target population of research interest. Since populations can vary over time, with additions and subtractions, the researcher needs to investigate how well the frame has been maintained and updated. Defects in the frame can have harmful effects on the sample unless appropriate countermeasures are taken. Once a potential sampling frame has been identified, the following questions should be asked:

- *How good is the coverage of the sampling frame?* Many frames have some *missing elements* that are part of the target population, but that are not represented on the frame. Missing elements have no chance of selection for the sample, and this noncoverage may cause a distortion in the survey estimates.
- *At the final stage of selection, do the listings represent individual elements, or do they sometimes comprise clusters of elements?* For example, for a survey with households as the elements, the final stage of sample selection may be the selection of addresses from an address list in a given area. Some of the sampled addresses may be found to comprise *clusters* of more than one household. As another example, when the elements of analysis are persons, and the sample is taken from a frame of addresses, the resultant sample of addresses (clusters) will mostly contain more than one person.
- *Are there elements on the frame that are not part of the target population?* Unwanted listings may be classified as either *blanks* or *foreign*

elements. Blanks refer to listings that should not be present on the frame, in particular elements such as demolished dwellings, or persons who have died or emigrated. Those maintaining the frame should have deleted the blanks from the current frame. Foreign elements are elements on the frame that fall outside the defined survey population (e.g., nonsmokers in a survey of the behaviors of smokers) but that are valid listings for the purposes of the frame developers.

- *Do any elements in the survey population appear more than once on the frame?* Unless remedial action is taken, elements that appear more than once on the frame have multiple chances of selection for the sample. For example, this problem of *duplication* occurs when a sample of bank accounts is selected for a survey of the bank's clients because some clients may hold more than one account.
- *Can sampled elements be located so that their survey data can be collected?* It is not sufficient for the frame to provide just a list of the population elements. It also needs to contain up-to-date location information that will enable sampled elements to be located and survey data collected from them.

Since decisions about the sampling frame need to be made at the early stages in planning a survey, it is useful to illustrate them here before discussing sample designs. Suppose that a survey is to be conducted at a particular university about the students' experiences of sexual abuse, with, say, the requirement that the survey should produce separate estimates of adequate precision by sex and broad field of study. What sampling frame should be used? The choice will depend on the mode used to contact sampled students and collect their survey data. One approach could be to send survey questionnaires to the sampled students' addresses, but the address frame may not be up-to-date, and the response rate may be low. An alternative mode of contact could be by email for a web data collection. There are clear benefits to this approach. However, one should first examine the quality of the university's email address frame for coverage, blanks on the frame, and duplication. No sampling frame is perfect, but in this case, the email address frame may be deemed adequate without the need to apply any of the frame-enhancement procedures discussed in Chapter 8. The next question to ask is: What information recorded on the frame—such as sex, department, and year of study—could be useful for improving the efficiency of the sample? Given this information, a sample design can be created to meet the survey objectives in an effective way. Since not all sampled students will respond, this information may also be useful for attempting to compensate for nonresponse.

An important point to note about design-based inference is that it relates only to probability sampling from the population on the sampling frame. In this example, the inferences relate only to this particular university. Any attempt to generalize beyond that, such as to all university students in the country, necessarily depends on untestable model-based assumptions. This point is of particular significance for the many small-scale research surveys that, for reasons of constrained resources, are limited to probability samples of restricted and convenient populations. To the extent that there is room to choose the population with such surveys, that choice should be made with the potential generalizability of the survey results to the larger population in mind.

The next three chapters describe three simple probability designs—simple random sampling, systematic sampling, and stratified sampling—in which the elements are sampled in a single stage from the sampling frame. In general, such designs require a list frame of elements from which the elements can be sampled directly. For face-to-face interview surveys, these designs are suitable only when the population is a compact one so that clustering of sample elements is not needed for efficient data collection; however, compactness is not required for mail, telephone, and Internet surveys.

Chapter 5 introduces complex sample designs in which the sampled elements are selected in clusters, often with several stages of sampling: First, a sample of primary sampling units (PSUs) is selected, and then a sample of second-stage units (SSUs) is selected within the selected PSUs, etc., until elements are selected at the final stage. The techniques of systematic sampling and stratified sampling are commonly used at each stage of a multistage design. With area sampling, for instance, the PSUs are often US counties or combinations of counties, while the SSUs are census blocks, block groups, or combinations of them. A complication with multistage sample designs is that the PSUs, SSUs, etc., vary a great deal in size, where size refers to the number of elements they contain. This complication is often handled by sampling the PSUs, SSUs, etc., with probabilities proportional to size (PPS) or probabilities proportional to estimated size (PPES). PPS and PPES sampling are taken up in Chapter 6. Some other probability sample designs are discussed in Chapter 7.

Chapter 8 describes methods that address deficiencies in sampling frames, and Chapter 9 discusses the effect of the failure to obtain the survey data from some of the sampled elements. The following two chapters describe the methods needed to take account of the survey sample design and the sample deficiencies in the analysis of the survey data. Chapter 10 describes the base weights that are needed to account for the sampled elements' unequal selection probabilities. Chapter 11 describes methods for

adjusting the base weights in an attempt to compensate for element nonresponse and for noncoverage, and the imputation methods that may be used to handle item nonresponses for otherwise responding elements. Chapter 12 describes methods for estimating the variances of survey estimates that take account of a survey's complex sample design. The calculation of the sample size required for a survey based on a probability sample design is finally taken up in Chapter 13, after having identified all the factors that need to be incorporated. Chapter 14 describes four actual examples of complex sample designs that combine many of the sample design features discussed in earlier chapters, together with the weighting and variance estimation methods used.

Chapter 15 is a departure from the rest of the book. It describes various methods of nonprobability sampling that have become popular in recent years. The final chapter provides some suggestions for further reading.