1 Introduction

1.1 The purpose of this book

There are a number of books available that treat various aspects of survey design, sampling, survey implementation, and so forth (examples include Cochran, 1963; Dillman, 1978, 2000; Groves and Couper, 1998; Kish, 1965; Richardson, Ampt, and Meyburg, 1995; and Yates, 1965). However, there does not appear to be a single book that covers all aspects of a survey, from the inception of the survey itself through to archiving the data. This is the purpose of this book. The reader will find herein a complete treatment of all aspects of a survey, including all the elements of design, the requirements for testing and refinement, fielding the survey, coding and analysing the resulting data, documenting what happened, and archiving the data, so that nothing is lost from what is inevitably an expensive process.

This book concentrates on surveys of human populations, which are both more challenging generally and more difficult both to design and to implement than most surveys of non-human populations. In addition, because of the background of the author, examples are drawn mainly from surveys in the area of transport planning. However, the examples are purely illustrative; no background is needed in transport planning to understand the examples, and the principles explained are applicable to any survey that involves human response to a survey instrument. In spite of this focus on human participation in the survey process, there are occasional references to other types of surveys, especially observational and counting types of surveys.

In writing this book, the author has tried to make this as complete a treatment as possible. Although extensive references are included to numerous publications and books in various aspects of measuring data, the reader should be able to find all that he or she requires within the covers of this book. This includes a chapter on some basic aspects of statistics and probability that are used subsequently, particularly in the development of the statistical aspects of surveys.

In summary, then, the purpose of this book is to provide the reader with an extensive and, as far as possible, exhaustive treatment of issues involved in the design and execution of surveys of human populations. It is the intent that, whether the reader is a student, a professional who has been asked to design and implement a survey, or

someone attempting to gain a level of knowledge about the survey process, all questions will be answered within these pages. This is undoubtedly a daunting task. The reader will be able to judge the extent to which this has been achieved. The book is also designed that someone who has no prior knowledge of statistics, probability, surveys, or the purposes to which surveys may be put can pick up and read this book, gaining knowledge and expertise in doing so. At the same time, this book is designed as a reference book. To that end, an extensive index is provided, so that the user of this book who desires information on a particular topic can readily find that topic, either from the table of contents, or through the index.

1.2 Scope of the book

As noted in the previous section, the book starts with a treatment of some basic statistics and probability. The reader who is familiar with this material may find it appropriate to skip this chapter. However, for those who have already learnt material of this type but not used it for a while, as well as those who are unfamiliar with the material, it is recommended that this chapter be used as a means for review, refreshment, or even first-time learning. It is then followed by a chapter that outlines some basic issues of surveys, including a glossary of terms and definitions that will be found helpful in reading the remainder of the book. A number of fundamental issues, pertinent to overall survey design, are raised in this chapter. Chapter 4 introduces the topic of the ethics of surveys, and outlines a number of ethical issues and proposes a number of basic ethical standards to which surveys of human populations should adhere. The fifth chapter of the book discusses the primary issues of designing a survey. A major underlying theme of this chapter is that there is no such thing as an 'all-purpose survey'. Experience has repeatedly demonstrated that only surveys designed with a clear purpose in mind can be successful.

The next nine chapters deal with all the various design issues in a survey, given that we have established the overall purpose or purposes of the survey. The first of these chapters (Chapter 6) discusses and describes all the current methods that are available for conducting surveys of human populations, in which people are asked to participate in the survey process. Mention is also made of some methods of dealing with other types of survey that are appropriate when the objects of the survey are observed in some way and do not participate in the process. In Chapter 7, the topic of focus groups is introduced, and potential uses of focus groups in designing quantitative and qualitative surveys are discussed. The chapter does not provide an exhaustive treatment of this topic, but does provide a significant amount of detail on how to organise and design focus groups. In Chapter 8, the design of survey instruments is discussed at some length. Illustrations of some principles of design are included, drawn principally from transport and related surveys. Chapters 9 and 10 deal with issues relating to question design and question wording and special issues relating to qualitative and preference surveys. Chapter 11 deals with the design of data collection procedures themselves, including such issues as item and unit nonresponse, what constitutes a

complete response, the use of proxy reporting and its effects, and so forth. The seventh of this group of chapters (Chapter 12) deals with pilot surveys and pretests – a topic that is too often neglected in the design of surveys. A number of issues in designing and undertaking such surveys and tests are discussed. Chapter 13 deals with the topic of sample design and sampling issues. In this chapter, there is extensive treatment of the statistics of sampling, including estimation of sampling errors and determination of sample sizes. The chapter describes most of the available methods of sampling, including simple random samples, stratified samples, multistage samples, cluster samples, systematic samples, choice-based samples, and a number of sampling methods that are often considered but that should be avoided in most instances, such as quota samples, judgemental samples, and haphazard samples.

Chapter 14 addresses the topic of repetitive surveys. Many surveys are intended to be done as a 'one-off' activity. For such surveys, the material covered in the preceding chapters is adequate. However, there are many surveys that are intended to be repeated from time to time. This chapter deals with such issues as repeated cross-sectional surveys, panel surveys, overlapping samples, and continuous surveys. In particular, this chapter provides the reader with a means to compare the advantages and disadvantages of the different methods, and it also assists in determining which is appropriate to apply in a given situation.

Chapter 15 builds on the material in the preceding chapters and deals with the issue of survey economics. This is one of the most troublesome areas, because, as many companies have found out, it is all too easy to be bankrupted by a survey that is undertaken without a real understanding and accounting of the costs of a survey. While information on actual costs will date very rapidly, this chapter attempts to provide relative data on costs, which should help the reader estimate the costs of different survey strategies. This chapter also deals with many of the potential trade-offs in the design of surveys.

Chapter 16 delves into some of the issues relating to the actual survey implementation process. This includes issues relating to training survey interviewers and monitoring the performance of interviewers, and the chapter discusses some of the danger signs to look for during implementation. This chapter also deals with issues regarding the ethics of survey implementation, especially the relationships between the survey firm, the client for the survey, and the members of the public who are the respondents to the survey. Chapter 17 introduces a topic that is becoming of increasing interest: Web-based surveys. Although this is a field that is as yet quite young, there are an increasing number of aspects that have been researched and from which the reader can benefit. Chapter 18 deals with the process of coding and data entry. A major issue in this topic is the geographic coding of places that may be requested in a survey.

Chapter 19 addresses the topics of data expansion and weighting. Data expansion is outlined as a function of the sampling method, and statistical procedures for expanding each of the different types of sample are provided in this chapter. Weighting relates to problems of survey bias, resulting either from incomplete coverage of the population in the sampling process or from nonresponse by some members of the subject population.

This is an increasingly problematic area for surveys of human populations, resulting from a myriad of issues relating to voluntary participation. Chapter 20 addresses the issue of nonresponse more completely. Here, issues of who is likely to respond and who is not are discussed. Methods to increase response rates are described, and reference is made again to the economics of the survey design. The question of computing response rates is also addressed in this chapter. This is usually the most widely recognised statistic for assessing the quality of a survey, but it is also a statistic that is open to numerous methods of computation, and there is considerable doubt as to just what it really means.

Chapter 21 deals with a range of other measures of data quality, some that are general and some, by way of example, that are specific to surveys in transport. These measures are provided as a way to illustrate how survey-specific measures of quality can be devised, depending on the purposes of the survey. Chapter 22 discusses some issues of the future of human population surveys, especially in the light of emerging technologies and their potential application and misapplication to the survey task.

Chapter 23, the final chapter in the book, covers the issues of documenting and archiving the data. This all too often neglected area of measuring data is discussed at some length. A list of headings for the final report on the survey is provided, along with suggestions as to what should be included under the headings. The issue of archiving data is also addressed at some length. Data are expensive to collect and are rarely archived appropriately. The result is that many expensive surveys are effectively lost soon after the initial analyses are undertaken. In addition, knowledge about the survey is often lost when those who were most centrally involved in the survey move on to other assignments, or leave to work elsewhere.

1.3 Survey statistics

Statistics in general, and survey statistics in particular, constitute a relatively young area of theory and practice. The earliest instance of the use of statistics is probably in the middle of the sixteenth century, and related to the start of data collection in France regarding births, marriages, and deaths, and in England to the collection of data on deaths in London each week (Berntson *et al*., 2005). It was then not until the middle of the eighteenth century that publications began to appear advancing some of the earliest theories in statistics and probability. However, much of the modern development of statistics did not take place until the late nineteenth and early twentieth centuries (Berntson *et al*., 2005):

> Beginning around 1880, three famous mathematicians, Karl Pearson, Francis Galton and Edgeworth, created a statistical revolution in Europe. Of the three mathematicians, it was Karl Pearson, along with his ambition and determination, that led people to consider him the founder of the twentieth-century science of statistics.

It was only in the early twentieth century that most of the now famous names in statistics made their contributions to the field. These included such statisticians as Karl

Pearson, Francis Galton, C. R. Rao, R. A. Fisher, E. S. Pearson, and Jerzy Neyman, among many others, who all made major contributions to what we know today as the science of statistics and probability.

Survey sampling statistics is of even more recent vintage. Among the most notable names in this field of study are those of R. A. Fisher, Frank Yates, Leslie Kish, and W. G. Cochran. Fisher may have given survey sampling its birth, both through his own contributions and through his appointment of Frank Yates as assistant statistician at Rothamsted Experimental Station in 1931. In this post, Yates developed, often in collaboration with Fisher, what may be regarded as the beginnings of survey sampling in the form of experimental designs (O'Connor and Robertson, 1997). His book *Sampling Methods for Censuses and Surveys* was first published in 1949, and it appears to be the first book on statistical sampling designs.

Leslie Kish, who founded the Survey Research Institute at the University of Michigan, is also regarded as one of the founding fathers of modern survey sampling methods, and he published his seminal work, called *Survey Sampling*, in 1965. Close in time to Kish, W. G. Cochran published his seminal work, *Sampling Techniques*, in 1963.

Based on these efforts, the science of survey sampling cannot be considered to be much over fifty years old – a very new scientific endeavour. As a result of this relative recency, there is still much to be done in developing the topic of survey sampling, while technologies for undertaking surveys have undergone and continue to undergo rapid evolution. The fact that most of the fundamental books on the topic are about forty years old suggests that it is time to undertake an updated treatise on the topic. Hence, this book has been undertaken.

2 Basic statistics and probability

2.1 Some definitions in statistics

Statistics is defined by the *Oxford Dictionary of English Etymology* as 'the political science concerned with the facts of a state or community', and the word is derived from the German *statistisch*. The beginning of modern statistics was in the sixteenth century, when large amounts of data began to be collected on the populations of countries in Europe, and the task was to make sense of these vast amounts of data. As statistics has evolved from this beginning, it has become a science concerned with handling large quantities of data, but also with using much smaller amounts of data in an effort to represent entire populations, when the task of handling data on the entire population is too large or expensive. The science of statistics is concerned with providing inputs to political decision making, to the testing of hypotheses (understanding what would happen if …), drawing inferences from limited data, and, considering the data limitations, doing all these things under conditions of uncertainty.

A word used commonly in statistics and surveys is *population*. The *population* is defined as the entire collection of elements of concern in a given situation. It is also sometimes referred to as a *universe*. Thus, if the elements of concern are pre-school children in a state, then the population is all the pre-school children in the state at the time of the study. If the elements of concern are elephants in Africa, then the population consists of all the elephants currently in Africa. If the elements of concern are the vehicles using a particular freeway on a specified day, then the population is all the vehicles that use that particular freeway on that specific day.

It is very clear that statistics is the study of data. Therefore, it is necessary to understand what is meant by *data*. The word *data* is a plural noun from the Latin *datum*, meaning given facts. As used in English, the word means given facts from which other facts may be inferred. Data are fundamental to the analysis and modelling of real-world phenomena, such as human populations, the behaviour of firms, weather systems, astronomical processes, sociological processes, genetics, etc. Therefore, one may state that statistics is the process for handling and analysing data, such that useful conclusions can be drawn, decisions made, and new knowledge accumulated.

Another word used in connection with statistics is *observation*. An *observation* may be defined as the information that can be seen about a member of a subject population. An observation comprises data about relevant characteristics of the member of the population. This population may be people, households, galaxies, private firms, etc. Another way of thinking of this is that an observation represents an appropriate grouping of data, in which each observation consists of a set of data items describing one member of the population.

A *parameter* is a quantity that describes some property of the population. Parameters may be given as numbers, proportions, or percentages. For example, the number of male pre-school children in the state might be 16,897, and this number is a parameter. The proportion of baby elephants in Africa might be 0.39, indicating that 39 per cent of all elephants in Africa at this time are babies. This is also a parameter. Sometimes, one can define a particular parameter as being critical to a decision. This would then be called a *decision parameter*. For example, suppose that a decision is to be made as to whether or not to close a primary school. The decision parameter might be the number of schoolchildren that would be expected to attend that school in, say, the next five years.

A *sample* is some subset of a population. It may be a large proportion of the population, or a very small proportion of the population. For example, a survey of Sydney households, which comprise a population of about 1,300,000 might consist of 130,000, households (a 10 per cent sample) or 300 households (a 0.023 per cent sample).

A *statistic* is a numerical quantity that describes a sample. It is therefore the equivalent of a parameter, but for a sample rather than the population. For example, a survey of 130,000 households in Sydney might have shown that 52 per cent of households own their own home or are buying it. This would be a statistic. If, on the other hand, a figure of 54 per cent was determined from a census of the 1,300,000 households, then this figure would be a parameter.

Statistical inference is the process of making statements about a population based on limited evidence from a sample study. Thus, if a sample of 130,000 households in Sydney was drawn, and it was determined that 52 per cent of these owned or were purchasing their homes, then statistical inference would lead one to propose that this might mean that 676,000 (52 per cent of 1,300,000) households in Sydney own or are purchasing their homes.

2.1.1 Censuses and surveys

Of particular relevance to this book is the fact that there are two methods for collecting data about a population of interest. The first of these is a census, which involves making observations of every member of the population. Censuses of the human population have been undertaken in most countries of the world for many years. There are references in the Bible to censuses taken among the early Hebrews, and later by the Romans at the time of the birth of Christ. In Europe, most censuses began in the eighteenth century, although a few began earlier than that. In the United States of America,

censuses began in the nineteenth century. Many countries undertake a census once in each decade, either in the year ending in zero or in one. Some countries, such as Australia, undertake a census twice in each decade. A census may be as simple as a head count (enumerating the total size of the population) or it may be more complex, by collecting data on a number of characteristics of each member of the population, such as name, address, age, country of birth, etc.

A survey is similar to a census, except that it is conducted on a subset of the population, not the entire population. A survey may involve a large percentage of the population or may be restricted to a very small sample of the population. Much of the science of survey statistics has to do with how one makes a small sample represent the entire population. This is discussed in much more detail in the next chapter. A survey, by definition, always involves a sample of the population. Therefore, to speak of a 100 per cent sample is contradictory; if it is a sample, it must be less than 100 per cent of the population.

2.2 Describing data

One of the first challenges for statistics is to describe data. Obviously, one can provide a complete set of data to a decision maker. However, the human mind is not capable of utilising such information efficiently and effectively. For example, a census of the United States would produce observations on over 300 million people, while one of India would produce observations of over 1 billion people. A listing of those observations represents something that most human beings would be incapable of utilising. What is required, then, is to find some ways to simplify and describe data, so that useful information is preserved but the sheer magnitude of the underlying data is hidden, thereby not distracting the human analyst or decision maker.

Before examining ways in which data might be presented or described, such that the mind can grasp the essential information contained therein, it is important to understand the nature of different types of data that can be collected. To do this, it seems useful to consider the measurement of a human population, especially since that is the main topic of the balance of this book.

In mathematical statistics, we refer to things called *variables*. A variable is a characteristic of the population that may take on differing or varying values for different members of the population. Thus, variables that could be used to describe members of a human population may include such characteristics as name, address, age or date of birth, place of birth, height, weight, eye colour, hair colour, and shoe size. Each of these characteristics provides differing levels of information that can be used in various ways. We can divide these characteristics into four different types of scales, a scale representing a way of measuring the characteristic.

2.2.1 Types of scales

Nominal scales

Each person in the population has a name. The person's name represents a label by which that person can be identified, but provides little other information. Names can

be ordered alphabetically or can be ordered in any of a number of arbitrary ways, such as the order in which data are collected on individuals. However, no information is provided by changing the order of the names. Therefore, the only thing that the name provides is a label for each member of the population. This is called a *nominal* scale. A nominal scale is the least informative of the different types of scales that can be used to measure characteristics, but its lack of other information does not render it of less value. Other examples of nominal data are the colours of hair or eyes of the members of the population, bus route numbers, the numbers assigned to census collection districts, names of firms listed on a country's stock exchange, and the names of magazines stocked by a newsagency.

Ordinal scales

Each person in the population has an address. The address will usually include a house number and a street name, along with the name of the town or suburb in which the house is located. The address clearly also represents a label, just as does the person's name. However, in the case of the address, there is more information provided. If the addresses are sorted by number and by street, in most places in the world this will provide additional information. These sorted addresses will actually help an investigator to locate each home, in that it is expected that the houses are arranged in numerical order along the street, and probably with odd numbers on one side of the street and even numbers on the other side. As a result, there is *order* information provided in the address. It is, therefore, known as an *ordinal* scale. However, if it is known that one person lives at 27 Main Street, and another person lives at 35 Main Street, this does not indicate how far apart these two people live. In some countries, they could be next door to each other, while in others there might be three houses between them or even seven houses between them (if numbering goes down one side of the street and back on the other). The only thing that would be known is that, starting at the first house on Main Street, one would arrive at 27 before one would arrive at 35. Therefore, order is the only additional information provided by this scale. Other examples of ordinal scales would be the list of months in the year, censor ratings of movies, and a list of runners in the order in which they finished a race.

Interval scales

Each person in the population has a shoe size. For the purposes of this illustration, the fact that there are slight inconsistencies in shoe sizes between manufacturers will be ignored, and it will be assumed, instead, that a man's shoe size nine is the same for all men's shoes, for example. Shoe size is certainly a label, in that a shoe can be called a size nine or a size twelve, and so forth. This may be a useful way of labelling shoes for a lot of different reasons. In addition, there is clearly order information, in that a size nine is smaller than a size twelve, and a size seven is larger than a size five. Furthermore, within each of children's, men's, and women's shoes, each increase in a size represents a constant increase in the length of the shoe. Thus, the difference between a size nine and a size ten shoe for a man is the same as the difference between a size eight and a size nine, and so on for any two adjacent numbers. In other words,

there is a constant *interval* between each shoe size. On the other hand, there is no natural zero in this scale (in fact, a size of zero generally does not exist), and it is not true that a size five is half the length of a size ten. Therefore, shoe size may be considered to be an *interval* scale. Women's dress sizes in a number of countries also represent an interval scale, in which each increment in dress size represents a constant interval of increase in size of the dress, but a size sixteen dress is not twice as large as a size eight. In many cases, the sizing of an item of clothing as small, medium, large, etc. also represents an interval scale. Another example of an interval scale is the normal scale of temperature in either degrees Celsius or degrees Fahrenheit. An interval of one degree represents the same increase or decrease in temperature, whether it is between 40 and 41 or 90 and 91. However, we are not able to state that 60 degrees is twice as hot as 30 degrees. There is also not a natural zero on either the Celsius scale or the Fahrenheit scale. Indeed, the Celsius scale sets the temperature at which water freezes as 0, but the Fahrenheit scale sets this at 32, and there is not a particular physical property of the zero on the Fahrenheit scale.

Ratio scale

Each member of the population has a height and a weight. Again, each of these two measures could be used as a label. We might say that a person is 180 centimetres tall, or weighs 85 kilograms. These measures also contain ordinal information. We know that a person who weighs 85 kilograms is heavier than a person who weighs 67 kilograms. Furthermore, we know that these measures contain interval information. The difference between 179 centimetres and 180 centimetres is the same as the difference between 164 centimetres and 165 centimetres. However, there is even more information in these measures. There is *ratio* information. In other words, we know that a person who is 180 centimetres tall is twice as tall as a person who is 90 centimetres tall, and that a person weighing 45 kilograms is only half the weight of a person weighing 90 kilograms. There are two important new pieces of information provided by these measures. First, there is a natural zero in the measurement scale. Both weight and height have a zero point, which represents the absence of weight or the absence of height. Second, there is a multiplicative relationship among the measures on the scale, not just an additive one. Therefore, both weight and height are described as *ratio* scales. Other examples of ratio scales are distance or length measures, measures of speed, measures of elapsed time, and so forth. However, it should be noted that measurement of clock time is interval-scaled (there is no natural zero, and 5 a.m. is not a half of 10 a.m.), while elapsed time is ratio-scaled, because zero minutes represents the absence of any elapsed time, and twenty minutes is twice as long as ten minutes, for example.

Measurement scales

The preceding sections have outlined four scales of measurement: nominal, ordinal, interval, and ratio. They have also demonstrated that these four scales are themselves an ordinal scale, in which the order, as presented in the preceding sentence, indicates increasing information content. Furthermore, each of the scale types, as ordered above,