Part I

Basic topics

If knowledge can create problems, it is not through ignorance that we can solve them. (Isaac Asimov, 1920–1992)

1 In the beginning

> Education is a progressive discovery of our own ignorance. (Will Durant, 1885–1991)

Why did we *choose* to write this primer? Can it be explained by some inherent desire to seek personal gain, or was it some other less self-centered interest? In determining the reason, we are revealing an underlying objective. It might be one of maximizing our personal satisfaction level or that of satisfying some community-based objective (or social obligation). Whatever the objective, it is likely that there are a number of reasons why we made such a *choice* (between writing and not writing this primer) accompanied by a set of constraints that had to be taken into account. An example of a reason might be to "promote the field of research and practice of choice analysis"; examples of constraints might be the time commitment and the financial outlay.

Readers should be able to think of choices that they have made in the last seven days. Some of these might be repetitive and even habitual (such as taking the bus to work instead of the train or car), buying the same daily newspaper (instead of other ones on sale); other choices might be a one-off decision (such as going to the movies to watch a latest release or purchasing this book). Many choice situations involve *more than one choice*, such as choosing a destination and means of transport to get there, or choosing where to live and the type of dwelling.

The storyline above is rich in information about what we need to include in a study of the choice behavior of individuals. To arrive at a choice, an individual must have considered a set of *alternatives*. These alternatives are usually called the *choice set*. Logically one must evaluate at least two alternatives to be able to make a choice (one of these alternatives may be "not to make a choice" or "not participate at all"). At least one actual choice setting must exist (e.g. choosing where to live), but there may be more than one choice (e.g. what type of dwelling to live in, whether to buy or rent, and how much to pay per week if rented). The idea that an individual may have to consider a number of choices leads to a set of *inter-related choices.*

Determining the set of alternatives to be evaluated in a choice set is a crucial task in choice analysis. Getting this wrong will mean that subsequent tasks in the development of

a choice model will be missing relevant information. We often advise analysts to devote considerable time to the identification of the choices that are applicable in the study of a specific problem. This is known as *choice set generation*. In identifying the relevant choices, one must also consider the range of alternatives and start thinking about what influences the decision to choose one alternative over another. These influences are called *attributes* if they relate to the description of an alternative (e.g. travel time of the bus alternative), but an individual's prejudices (or tastes) will also be relevant and are often linked to socio-economic *characteristics* such as personal income, age, gender, and occupation.

To take a concrete example, a common problem for transportation planners is to study the transport-related choices made by a sample of individuals living in an urban area. Individuals make many decisions related to their transportation needs. Some of these decisions are taken occasionally (e.g. where to live and work) while others are taken more often (e.g. departure time for a specific trip). These examples highlight a very important feature of choice analysis – the *temporal perspective*. Over what time period are we interested in studying choices? As the period becomes longer, the number of possible choices that can be made (i.e. are not fixed or predetermined) are likely to increase. Thus if we are interested in studying travel behavior over a five-year period then it is reasonable to assume that an individual can make choices related to the locations of both living and working, as well as the means of transport and departure time. That is, a specific choice of means of transport may indeed be changed as a consequence of the person changing where they reside or work. In a shorter period such as one year, choosing among modes of transport may be conditional on where one lives or works, but the latter is not able to be changed given the time that it takes to relocate one's employment.

The message in the previous paragraphs is that careful thought is required to define the choice setting so as to ensure that all possible behavioral responses (as expressed by a set of choice situations) can be accommodated when a change in the decision environment occurs. For example, if we increase fuel prices, then the cost of driving a car increases. If one has studied only the choice of mode of transport then the decision maker will be "forced" to modify the choice among a given set of modal alternatives (e.g. bus, car, train). However it may be that the individual would prefer to stay with the car but to change the time of day they travel so as to avoid traffic congestion and conserve fuel. If the departure time choice model is not included in the analysis, then experience shows that the modal choice model tends to force a substitution between modes which in reality is a substitution between travel at different times of day by car.

Armed with a specific problem or a series of associated questions the analyst now recognizes that to study choices we need a set of choice situations (or outcomes), a set of alternatives, and a set of attributes that belong to each alternative. But how do we take this information and convert it to a useful framework within which we can study the choice behavior of individuals? To do this, we need to set up a number of *behavioral rules* under which we believe it is reasonable to represent the process by which an individual considers a set of alternatives and makes a choice. This framework needs to be sufficiently realistic to explain past choices and to give confidence in likely behavioral responses in the future that result in staying with an existing choice or making a new choice. The framework should also be capable of assessing the likely support for alternatives that are not currently available,

be they new alternatives in the market or existing ones that are physically unavailable to some market segments.

The following sections of this primer will introduce the main rules that are needed to start understanding the richness of methods available to study. We will start right from the beginning and learn to "walk before we run." We will be pedantic in the interest of clarity, since what is taken for granted by the long-established choice analyst is often gobbledy-gook to the beginner. Intolerance on the part of such "experts" has no place in this primer.

We have found in our courses that the best way to understand the underlying constructs that are the armory of choice analysis is to select one specific choice problem and follow it through from the beginning to the end. We will do this here. Selecting such a single case study is always fraught with problems since, no matter what we select, it will not be ideal for every reader. To try and offer a smorgasbord of case studies would (in our view) defeat the purpose of this primer. While readers will come from different disciplinary backgrounds such as economics, geography, environmental science, marketing, health science, statistics, engineering, transportation, logistics, and so forth and will be practicing in these and other fields, the tools introduced through one case study are universally relevant.

A reader who insists that this is not so is at a disadvantage; she is committing the sin of assuming uniqueness in behavioral decision making and choice response. Indeed the great virtue of the methods developed under the rubric of choice analysis is their universal relevance. Their portability is amazing. Disciplinary boundaries and biases are a threat to this strength. While it is true that specific disciplines have a lot to offer to the literature on choice analysis, we see these offerings as contributions to the bigger multi-disciplinary effort.

The case study focuses on a transport choice – the choice between a number of public and private modes of transport for travel within an urban area. The data were collected in 1994 as part of a larger study that resulted in the development of an environmental impact simulator to assess the impact of policy instruments on the demand for travel. We have selected this specific context because we have a real data set (provided on the primer website) that has all of the properties we need to be able to illustrate the following features of choice analysis:

1. There are *more than two alternatives* (in particular, car drive alone, car ride share, train, and bus). This is important because a choice situation involving more than two alternatives introduces a number of important behavioral conditions that do not exist when studying a binary choice.
2. It is possible to view the set of alternatives as *more than one choice* (e.g. choosing between public and private modes, choosing among the private modes, and choosing among the public modes). This will be important later to show how to set up a choice problem with more than one (inter-related) choice decision.
3. Two types of choice data have emerged as the primary sources of choice response. These are known as *revealed preference* (RP) and *stated preference* (SP) data. RP data refer to situations where the choice is actually made in real market situations; in contrast, SP data refer to situations where a choice is made by considering hypothetical

situations (which are typically the same alternatives in the RP data set, but are described by different levels of the same attributes to those observed in actual markets as well as additional attributes not in the data collected from actual markets). SP data are especially useful when considering the choice among existing and new alternatives since the latter are not observed in RP data. The case study data have both RP and SP choice data with the SP choice set comprising the exact same four alternatives in the RP data set plus two "new" alternatives – light rail and a dedicated busway system.

4. Often in choice modeling we *over- and under-sample* individuals observed to choose specific alternatives. This is common where particular alternatives are dominant or popular (in this data it is use of the car compared to public transport). The case study data have over-sampled existing choosers of bus and train and under-sampled car users. In establishing the relative importance of the attributes influencing the choice among the alternatives we would want to correct for this over- and under-sampling strategy by *weighting the data* to ensure reproduction of the population choice shares. These weighted choice shares are more useful than the sample choice shares.
5. The data have a large number of *attributes* describing each alternative and *characteristics* describing the socio-economic profile of each sampled trip maker (e.g. personal income, age, car ownership status, occupation). This gives the analyst plenty of scope to explore the contributions of attributes of alternatives and characteristics of individuals to explaining choice behavior.
6. The alternatives are *well-defined modes of transport* that are described by labels such as bus, train, car drive alone, and car ride share. A data set with labeled alternatives is preferred over one where the alternatives are not well defined in terms of a label such as abstract alternatives that are only defined by combinations of attributes. Labeled alternatives enable us to study the important role of alternative-specific constants.
7. Finally, most analysts have had *personal experience* in choosing a mode of transport for the journey to work. Thus the application should be very familiar.

The following chapters set out the process of choice analysis in a logical sequence consistent with what researchers and practitioners tend to do as they design their study and collect all the necessary inputs to undertake data collection, analysis, and reporting. We begin with a discussion on what we are seeking to understand in a study of choice; namely the role of an individual's preferences and the constraints that limit the ability to choose alternatives that are the most preferred in an unconstrained setting. Having established the central role of preferences and constraints, we are ready to formalize a framework within which a set of behavioral rules can be introduced to assist the analyst in accommodating these individual preferences, recognizing that the analyst does not have as much information about these individual preferences as the individual decision maker being studied. The behavioral rules are used to develop a formal *model of choice* in which we introduce the sources of individual preferences (i.e. attributes), constraints on such preferences (i.e. characteristics of individuals, peer influences, and other contextual influences), and the available set of alternatives to choose from. This is where we introduce choice models such as multinomial logit and nested logit.

With a choice-modeling framework set out, we are ready to design the data stage. Important issues discussed are survey design and administration, data paradigms, data collection strategies, and data preparation for model estimation. Many analysts have difficulties in preparing their data in a format suitable for model estimation. Although there are a number of software options available for model estimation, we have selected NLOGIT for two reasons – it is the most popular software package for choice model estimation and it is the package that the authors have greatest expertise in using (William Greene and David Hensher are the developers of NLOGIT). We set out, step by step, what the analyst must do to run a simple choice model and then introduce more advanced features. The results are interpreted in a way that ensures that the main outputs are all considered and reported as appropriate. Estimating models is only one critical element of the choice-modeling process. The findings must be used in various ways such as forecasting, scenario analysis, valuation (willingness to pay or WTP), and understanding of the role of particular attributes and characteristics. We discuss the most common ways of applying the results of choice models such as simulating the impact of changes in levels of attributes, deriving marginal rates of substitution (or values) of one attribute relative to another (especially if one attribute is measured in monetary units), and in constructing empirical distributions of specific attributes and ratios of attributes. Throughout the book we add numerous hints under the heading of "**As an aside**". This format was chosen as a way of preserving the flow of the argument but placing useful tips where they would best be appreciated. Before we can delve into the foundations of choice analysis we need to set out the essential statistical concepts and language that readers new to statistics (or very rusty) will need to make the rest of the book easier to read.

2 Basic notions of statistics

> If scientific reasoning were limited to the logical processes of arithmetic, we should not get very far in our understanding of the physical world. One might as well attempt to grasp the game of poker entirely by the use of the mathematics of probability. (Vannevar Bush 1890–1974)

2.1 Introduction

This chapter is intended to act as a review of the basic statistical concepts, knowledge of which is required for the reader to fully appreciate the chapters that follow. It is not designed to act as a substitute for a good grounding in basic statistics but rather as a summary of knowledge that the reader should already possess. For the less confident statistician, we recommend that in reading this and subsequent chapters, that they obtain and read other books on the subject. In particular, we recommend for the completely statistically challenged *Statistics without Tears: A Primer for Non-Mathematicians* (Rowntree 1991). More confident readers may find books such as those by Howell (1999) and Gujarati (1999, chapters 2–5) to be of particular use.

2.2 Data

Data are fundamental to the analysis and modeling of real world phenomena such as consumer and organizational behavior. Understanding data are therefore critical to any study application and nowhere more than to studies involving discrete choice analysis. The data sets which we use, whether collected by ourselves or by others, will invariably be made up of numerous observations on multiple variables (an object that can take on many different values). Only through understanding the qualities possessed by each variable will the analyst be capable of deriving the most benefit from their data.

We may define variables on a number of different dimensions. Firstly, variables may be *qualitative* or *quantitative*. A qualitative variable is one in which the "true" or naturally

occurring levels or categories taken by that variable are not described as numbers but rather by verbal groupings (e.g. the levels or categories that hair color may take might include red, blond, brown, black). For such variables, comparisons are based solely on the *qualities* possessed by that particular variable. Quantitative variables on the other hand are those in which the natural levels take on certain quantities (e.g. price, travel time). That is, quantitative variables are measurable in some *numerical* unit (e.g. dollars, minutes, inches, etc.).

A second dimension is whether the variable is *continuous* or *non-continuous* in nature. A continuous variable is one in which it can theoretically assume any value between the lowest and highest points on the scale on which it is being measured (e.g. speed, price, time, height). For continuous variables, it is common to use a scale of measure such as minutes to quantify the object under study; however, invariably, such scale measures are potentially infinitely divisible (e.g. one could measure time in seconds, or thousandths of seconds, and so on). As such, continuous-level data will only be an approximation of the true value taken of the object under study, with the precision of the estimate dependent upon the instrument of measure. Non-continuous variables, sometimes referred to as *discrete variables*, differ from continuous variables in that they may take on a relatively few possible distinct values (e.g. male and female for gender).

A third dimension used in describing data are that of *scales of measurement*. Scales of measurement describe the relationships between the characteristics of the numbers or levels assigned to objects under study. Four classificatory scales of measurement were developed by Stevens (1951) and are still in use today. These are nominal, ordinal, interval, and ratio.

Nominal scaled data

A nominal scaled variable is a variable in which the levels observed for that variable are assigned unique values – values which provide classification but which do not provide any indication of order. For example, we may assign the values zero to represent males and one to represent females; however, in doing so, we are not saying that females are better than males or males are better than females. The numbers used are only to *categorize objects*. As such, it is common to refer to such variables as categorical variables (note that nominal data must be discrete). For nominal scaled data, all mathematical operations are meaningless (i.e. addition, subtraction, division and multiplication).

Ordinal scaled data

Ordinal scaled data are data in which the values assigned to levels observed for an object are (1) unique and (2) provide an indication of order. An example of this is ranking of products in order of preference. The highest-ranked product is more preferred than the second-highest-ranked product, which in turn is more preferred than the third-ranked product, etc. While we may now place the objects of measure in some order, we cannot determine *distances between the objects.* For example, we might know that product A is preferred to product B; however, we do not know by how much product A preferred to product B. Thus, addition, subtraction, division and multiplication are meaningless in terms of ordinal scales.

Interval scaled data

Interval scaled data are data in which the levels of an object under study are assigned values which are (1) unique, (2) provide an indication of order, and (3) have an equal distance between scale points. The usual example is temperature (Centigrade or Fahrenheit). In either scale, 41 degrees is higher than 40 degrees, and the increase in heat required to go from 40 degrees to 41 degrees is the same as the amount of heat to go from 20 degrees to 21 degrees. However, zero degrees is an arbitrary figure – it does not represent an absolute absence of heat (it does represent the temperature at which water freezes however when using the Centigrade scale). Because of this, you may add and subtract interval scale variables meaningfully but ratios are not meaningful (that is 40, degrees is not strictly twice as hot as 20 degrees).

Ratio scaled data

Ratio scaled data are data in which the values assigned to levels of an object are (1) unique, (2) provide an indication of order, (3) have an equal distance between scale points, and (4) the zero point on the scale of measure used represents an absence of the object being observed. An example would be asking respondents how much money was spent on fast food last week. This variable has order ($1 is less than $2), has equal distances among scale points (the difference between $2 and $1 is the same as the difference between $1000 and $999), and has an absolute zero point ($0 spent represents a lack of spending on fast food items). Thus, we can add, subtract and divide such variables meaningfully ($1 is exactly half of $2).

2.2.1 The importance of understanding data

Information such as that given in the previous section can be found in any of a number of statistics books. This fact alone suggests that understanding the types of data one has is an important (perhaps the most important) element in conducting any study. It is sometimes lost on the practitioner that the type of data she has (whether she collected it herself or not) dictates the type of analysis that can be undertaken. All statistical analysis makes assumptions with regard to the data used. For example, if one has collected data on income to use as the dependent variable in a linear regression analysis, the income variable must be collected in a format that meets the data requirements of the analytical technique for which it was collected to be used. Thus, in collecting data (i.e. in writing surveys), one must always be cognizant of the *types of analysis* one intends to conduct, even if the analysis is not to be conducted for another six months. Statistics, like a game of chess, requires the players to always be thinking several moves ahead.

2.3 A note on mathematical notation

In this section we outline the mathematical notion that is used throughout the book.

2.3.1 Summation

The Greek capital letter sigma, Σ, indicates summation or addition. For example:

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + \cdots + X_n$$

where i is an index of summation indicating that for some variable X, we take the first value of X ($i = 1$) and add to each subsequent value taken by X up to the nth appearance of X. The subscript i is important in mathematical notation as it is used to denote a *variable* as opposed to a constant term. A variable, as the name suggests, is a quantity that is able to assume any set of values (i.e. has no fixed quantitative value). A constant is a quantity which is fixed at some level and as such does not vary. Constant terms are generally denoted without a subscript i (e.g. k).

In practice, one can abbreviate the summation as follows:

$$\sum_{i=1}^{n} X_i \text{ or simply } \sum X_i$$

The summation operator has several useful properties:

1. The *summation of a constant term* (note we drop the subscript i for constant terms):

$$\sum_{i=1}^{n} k = k + k + \cdots + k = nk$$

For example, where $n = 3$ and $k = 4$, $\sum_{i=1}^{3} 4 = 4 + 4 + 4 = 3 \times 4 = 12$

2. The *summation of a constant term multiplied by a variable*:

$$\sum_{i=1}^{n} kX_i = kX_1 + kX_2 + \cdots + kX_n = k(X_1 + X_2 + \cdots + X_n)$$
$$= k\left(\sum_{i=1}^{n} X_i\right)$$

For example, where $k = 3$ and $X_1 = 2$, $X_2 = 3$, and $X_3 = 4$

$$\sum_{i=1}^{n} 3X_i = (3 \times 2) + (3 \times 3) + (3 \times 4) = 3(2 + 3 + 4) = 3\left(\sum_{i=1}^{n} X_i\right) = 27$$

3. *Summing two variables*:

$$\sum_{i=1}^{n} (X_i + Y_i) = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i$$

For example, assume $X_1 = 2$, $X_2 = 3$, and $X_3 = 4$ and $Y_1 = 4$, $Y_2 = 3$, and $Y_3 = 2$

$$\sum_{i=1}^{n} (X_i + Y_i) = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i = (2 + 3 + 4) + (4 + 3 + 2) = 18$$