PREFACE

The statistical techniques used in multivariate data analysis (MDA) enable the analyst to detect patterns buried in complex quantitative data. It is therefore not surprising that the techniques have proved to be of great interest and utility to behavioral and social scientists who continually grapple with highly complex phenomena. As a result, some knowledge of MDA techniques is increasingly required of students and practitioners in the behavioral and social sciences, if only to allow them to understand the research literatures in their disciplines. Courses and excellent texts on MDA abound, particularly for graduate students in these areas. What, then, is the justification for a book called *Making Sense of Multivariate Data Analysis?*

The justification for me lies in experiences over many years teaching MDA techniques to psychology graduate students and consulting on the use of these techniques as a dissertation supervisor and a journal reviewer. In my experience, many students are daunted by the prospect of engaging with MDA techniques, seeing them as an impenetrable thicket of technical complexity that they usually confront only because of course requirements. Having engaged with the techniques in some way, most leave with fragmented understandings and often a tendency to follow procedural rules in a semimindless fashion. Most striking of all, very few successfully integrate their grasp of MDA into their broader repertoire of research skills and knowledge. I believe that one major reason for these problems is that many students lack the conceptual frameworks that would enable them to appreciate the *unity* of MDA ideas and to incorporate them within their existing frameworks. This book is intended to help meet this deficiency.

The goal of making sense of MDA can be interpreted in a variety of ways. It can refer to understanding the statistical foundations of the techniques, to using computer packages that perform MDA, or to comprehending technical

articles and books. There are a variety of excellent texts that provide guidance on achieving these goals, and many are cited throughout the present book. However, my objective is complementary to theirs and derives from other interpretations of sense making. In broad terms, MDA techniques rest on a few foundation blocks and gain their power through a small number of ingenious analytic strategies. So, one general way to make sense of MDA techniques is to appreciate their unity both in their foundations and their operations. The goal of this book is to highlight this unity and to do so mainly in a conceptual fashion. Accordingly, there are few statistical symbols and formulae, and there is no assumption of *any* statistical knowledge on the part of the reader.

The book is comparatively short in order to generate and maintain an overall sense of coherence and cohesion. The goal of successfully guiding readers on a rapid journey from simple foundations to complex heights is unlikely to be achieved for all readers at all times. Some will want me to accelerate or decelerate when I do the opposite. Others will balk at simplifications and insufficient qualifications of statements. Others again will want more references to support particular claims. However, a short book does not permit detailed arguments, expositions, and justifications. My aim has been to tell a coherent short story rather than to produce another lengthy textbook that painstakingly charts the MDA territory. Other writers have done this admirably, and my hope is that this book will help to make their work even more accessible.

Although there are simple structures to be found in MDA, it cannot be denied that the techniques themselves are complex and sophisticated. Accordingly, even gaining a broad conceptual overview requires the active engagement of the reader. Careful reading, rereading, and reflection are necessary if the ideas are to be assimilated. Two features of the book are intended to encourage this engagement. First, there are critical reflections on the strengths and limitations of MDA, especially in Chapter 3. The power of MDA techniques can lead the unwary to believe that the techniques are all-powerful, and from there it is a short step to misinterpretation and misuse. Confronting some of these pitfalls in an introductory text may seem premature, but I believe that establishing a critical stance from the outset is important in itself and should encourage the reader to engage in active learning.

Second, sparing use is made of figures. This may seem perverse given the traditional equation of a picture and a thousand words. However, I have found that at the introductory level of MDA, figures can be counterproductive. This

is partly because of individual differences in cognitive processing, but also because the use of figures can lead the naive reader to mistake recognition for comprehension. For example, being able to draw a normal distribution or to explain it in words requires very different levels of understanding. The book therefore relies heavily on verbal exposition in order to encourage the reader to strive for comprehension rather than simple acceptance.

This book is intended as a freestanding, basic *introduction* to MDA with no prerequisites. Accordingly, it should be useful to a variety of readers and for a variety of purposes. In the context of advanced undergraduate or graduate courses, it could be used in conjunction with a set of readings or with a more advanced text. In either case, the present book could serve as a framework within which more detailed knowledge could be located and developed to the desired level. For practitioners who wish to evaluate the research literature that informs their practice, the book could provide an accessible and rapid introduction, which might be sufficient for their purposes or would more likely lead to further technique-specific reading. The book is less likely to be of interest to the experienced researcher, but it might still have some integrating function for those who feel their knowledge is fragmented. Although a psychologist wrote this book, pains have been taken to use examples from a variety of areas. So the intended audience is not limited to psychologists but is meant to encompass any behavioral or social scientist, pure or applied, who needs to become acquainted with MDA techniques.

KEY TERMS

In order to discuss the structure of the book, and given its lack of prerequisite knowledge, it is first necessary to comment on how some fundamental key terms are used. Throughout the book key terms are shown in bold when they are first discussed. Statistical analysis typically focuses on a set of **cases.** In the behavioral and social sciences, a case may be a person, a rat, a social group, an organization, a country, or any identifiable entity of interest to the analyst. The terms "case" and "individual" are used throughout the book only for consistency. Statistical information or **data** for a set of cases consist of numerical values for one or more of their attributes. Since these values typically vary across cases, a quantified attribute is commonly known as a **variable.** Some variables may be seen as somehow accounting for other variables.

The former are called **independent variables,** and the latter, **dependent variables.** For example, in Chapter 1 we will conduct a simple analysis to see whether women are happier than men. In this example, happiness is seen as a dependent variable in that it is proposed as being dependent on the independent variable of gender.

The numerical value of a variable for a case is produced by some sort of measurement process. Throughout the book we will distinguish between measurement processes that result in a **score** and those that result in a **category.** A score, for our purposes, is a number that has been produced using what is called an **interval scale of measurement.** Imagine that in the happiness example cases have given themselves a happiness score between 1 and 10. For this to count as an interval scale, we would have to assume that all of the ten possible scores are mutually exclusive, the ten scores can be meaningfully rank ordered, and the "distance" between any two adjacent scores is the same. As we will see, if these assumptions can be made, we will have a considerable head start in our attempts at statistical analysis.

Classifying the cases as men and women is an example of **categorical or nominal scaling.** To achieve this, all we have to assume is that the categories are mutually exclusive and comprehensive. (Strictly speaking, this is not measurement in the sense of assigning numbers to attributes. We could refer to men as the "1"category and women as the "2" category, but this would be arbitrary.) Distinguishing between scores and categories is important because, as we will see, their analysis requires different statistical techniques. There will be occasional reference to other sorts of measurement scales, but generally the interval and categorical scales will be sufficient for our purposes.

If one variable is analyzed on its own, the analysis is called **univariate.** If the relationship between two variables becomes the focus, the analysis is **bivariate.** When relationships involving three or more variables are analyzed, the analysis is **multivariate.** Actually, the word "multivariate" is more correctly used when there is more than one *dependent* variable, and some writers prefer the term "multivariable analysis" when this is not the case. However, in this book we will adopt the more common and looser usage of "multivariate analysis."

It is instructive to note that in the happiness example the implicit research question can be couched in two ways: Do men and women *differ* in happiness, or is there a *relationship* between happiness and gender? These sound different, but they are logically equivalent. The reason this is instructive is that it provides flexibility in how we frame an analysis. Conversely, it enables us to

appreciate what apparently diverse analyses have in common. In Part 1 we will introduce simple analyses more in terms of concrete differences than of abstract relationships among variables. In particular, we will emphasize the distinction between **individual (case) differences** and **group differences.** In the happiness example, this is reflected in the distinction between how much individuals differ in happiness versus how much men and women differ. Grasping this distinction at an early stage will help to avoid confusion later.

The final term that requires preliminary comment is the notion of **accounting,** which recurs throughout the book. The objective of most multivariate analyses is to account for patterns of differences on one or more variables. Different research objectives give different meanings to "accounting." The research goal may be to predict differences, to give a causal explanation, or to reduce the set of differences to a smaller set, and this does not exhaust the possibilities. The word "accounting" is used as a generic term that is intended to cover all of these specific meanings and to carry fewer problematic connotations. As we will see in Chapter 3, confusion between the theoretical and statistical usages of a term like "predictor" is one source of misconceptions about the capabilities of MDA.

STRUCTURE OF THE BOOK

The book is divided into two parts: The three chapters in Part 1 introduce the core ideas of MDA, and the five chapters in Part 2 explore the techniques themselves. Chapter 1 introduces the univariate and bivariate statistical building blocks that form the foundation of MDA, using small, contrived data sets. As noted earlier, the discussion assumes no prior statistical knowledge. In Chapter 2 we review the variety of factors that can influence the trustworthiness of results from any statistical analysis, in particular the role of chance. Then, in Chapter 3 we examine the reasons why MDA is needed and the strategy that lies at the heart of any MDA technique. This is the notion of the composite variable, whereby multiple variables can be packaged into one. Finally, in Chapter 3 we stand back from the statistical details and reflect on the strengths and limitations of MDA techniques. Together, these three chapters provide the foundation on which Part 2 is built and should therefore ideally be read first and in order. Even the reader with a good grasp of basic statistics should find material worth pondering on.

The five chapters in Part 2 introduce six MDA techniques: multiple regression, logistic regression, discriminant analysis, multivariate analysis of variance (MANOVA), factor analysis, and log-linear analysis. There are a number of issues dealt with in detail in Chapter 4 on multiple regression to which frequent and briefer reference is made in later chapters. Accordingly, while the chapters in Part 2 could be read in any order, it is probably easier to read Chapter 4 first. Most of the technique chapters share a broad common structure. This begins with a discussion of how the composite variable strategy is used within a particular technique. Then the particular statistical tools for the technique are introduced and their use is illustrated with actual data and examples from the research literature. After this, the issues affecting the trustworthiness of results produced by the particular technique are reviewed. The main exception to this structure is in Chapter 6, where it is necessary to provide a fairly lengthy preamble on the analysis of variance before introducing its multivariate counterpart MANOVA.

The techniques covered in Part 2 are a selection of those found in the MDA area. Since one of the primary aims of the book is to demonstrate the continuity between basic statistical building blocks and multivariate analysis, more advanced techniques such as path analysis and structural equation modeling have been excluded. These techniques sit on the next story up in the building that is statistical analysis and are better left for more advanced and comprehensive texts. Even on the first level of MDA techniques, there are too many to be included in a short introduction. Those selected are a variety in a number of senses. Four of the six techniques are used to examine how multiple independent variables account for one or more dependent variables. But in the last two chapters, the independent/dependent variable distinction disappears, and other analytic objectives emerge. The techniques also vary in the ways in which they deal with scores and categories, as signaled earlier. Finally, the techniques vary in their frequency of usage in the behavioral and social sciences. Multiple regression and factor analysis are the workhorses of countless nonexperimental studies, while MANOVA is frequently used in experimental studies. In contrast, logistic regression is rapidly increasing in popularity and overtaking discriminant analysis, which is used for the same analytic purposes but is more problematic. Finally, log-linear analysis, which focuses on categorical data, is probably the least used but deserves wider appreciation and adoption.

All of the examples in the book are concerned with what psychologists call "subjective well-being" or, more loosely, happiness. This choice was partly

motivated by the desire to accentuate the positive in a context that is often associated with negative feelings. But the focus was also chosen because this topic is an intrinsically interesting and burgeoning one in many parts of the behavioral and social sciences. One piece of research to which we return repeatedly in Part 2 is a study of the effects of workplace characteristics on the well-being of nurses. This is a published study (Budge, Carryer & Wood, 2003), but the authors have also kindly provided me with their data for detailed analyses using all of the MDA techniques covered in Part 2. All of these analyses were conducted with the SPSS for Windows package, Version 11.5, which is one of the most popular statistical packages among behavioral and social scientists.

At the end of each chapter are suggestions for further reading. To facilitate the flow of the text, references within chapters have been kept to a minimum. Almost all of the references deliberately guide the reader to sources that are accessible for the beginner. Those sources in turn provide guidance toward the more technical literature for those who wish it.

ACKNOWLEDGMENTS

I am very grateful for the contributions made by a number of people in the preparation of this book. The suggestions made by six anonymous reviewers at different stages were much appreciated. Inevitably, I did not always follow their advice, and any remaining errors or patches of fog are my responsibility. I am also very grateful to the literally thousands of students who over the years have pushed me to find new ways of understanding and communicating complex ideas. In particular, I thank those who encouraged me to write a book on multivariate data analysis. Turning to those contributors who can be named, I have been very fortunate to benefit from the outstanding editorial support provided by Lisa Cuevas Shaw. Her enthusiasm and professional acumen throughout the process have been invaluable. A debt of gratitude is also due to Claire Budge, Jenny Carryer, and Sue Wood, who very generously allowed me to dissect their published work and data as a running example in the book. Finally, more than tradition leads me to reserve my deepest appreciation for my wife, Claire. As a research psychologist and methods enthusiast, she has acted as a critical sounding board during countless hours of reading and discussion. As a partner, she has continually provided the many types of support that authors need to succeed in their solitary projects. This book is because of Claire, and for Claire.