Chapter 1

ASPECTS OF MULTIVARIATE ANALYSIS

1.1 Introduction

Scientific inquiry is an iterative learning process. Objectives pertaining to the explanation of a social or physical phenomenon must be specified and then tested by gathering and analyzing data. In turn, an analysis of the data gathered by experimentation or observation will usually suggest a modified explanation of the phenomenon. Throughout this iterative learning process, variables are often added or deleted from the study. Thus, the complexities of most phenomena require an investigator to collect observations on many different variables. This book is concerned with statistical methods designed to elicit information from these kinds of data sets. Because the data include simultaneous measurements on many variables, this body of methodology is called *multivariate analysis*.

The need to understand the relationships between many variables makes multivariate analysis an inherently difficult subject. Often, the human mind is overwhelmed by the sheer bulk of the data. Additionally, more mathematics is required to derive multivariate statistical techniques for making inferences than in a univariate setting. We have chosen to provide explanations based upon algebraic concepts and to avoid the derivations of statistical results that *require* the calculus of many variables. Our objective is to introduce several useful multivariate techniques in a clear manner, making heavy use of illustrative examples and a minimum of mathematics. Nonetheless, some mathematical sophistication and a desire to think quantitatively will be required.

Most of our emphasis will be on the *analysis* of measurements obtained without actively controlling or manipulating any of the variables on which the measurements are made. Only in Chapters 6 and 7 shall we treat a few experimental plans (designs) for generating data that prescribe the active manipulation of important variables. Although the experimental design is ordinarily the most important part of a scientific investigation, it is frequently impossible to control the

generation of appropriate data in certain disciplines. (This is true, for example, in business, economics, ecology, geology, and sociology.) You should consult [6] and [7] for detailed accounts of design principles that, fortunately, also apply to multivariate situations.

It will become increasingly clear that many multivariate methods are based upon an underlying probability model known as the multivariate normal distribution. Other methods are ad hoc in nature and are justified by logical or commonsense arguments. Regardless of their origin, multivariate techniques must, invariably, be implemented on a computer. Recent advances in computer technology have been accompanied by the development of rather sophisticated statistical software packages, making the implementation step easier.

Multivariate analysis is a "mixed bag." It is difficult to establish a classification scheme for multivariate techniques that is both widely accepted and indicates the appropriateness of the techniques. One classification distinguishes techniques designed to study interdependent relationships from those designed to study dependent relationships. Another classifies techniques according to the number of populations and the number of sets of variables being studied. Chapters in this text are divided into sections according to inference about treatment means, inference about covariance structure, and techniques for sorting or grouping. This should not, however, be considered an attempt to place each method into a slot. Rather, the choice of methods and the types of analyses employed are largely determined by the objectives of the investigation. In Section 1.2, we list a smaller number of practical problems designed to illustrate the connection between the choice of a statistical method and the objectives of the study. These problems, plus the examples in the text, should provide you with an appreciation of the applicability of multivariate techniques across different fields.

The objectives of scientific investigations to which multivariate methods most naturally lend themselves include the following:

1. *Data reduction or structural simplification.* The phenomenon being studied is represented as simply as possible without sacrificing valuable information. It is hoped that this will make interpretation easier.
2. *Sorting and grouping.* Groups of "similar" objects or variables are created, based upon measured characteristics. Alternatively, rules for classifying objects into well-defined groups may be required.
3. *Investigation of the dependence among variables.* The nature of the relationships among variables is of interest. Are all the variables mutually independent or are one or more variables dependent on the others? If so, how?
4. *Prediction.* Relationships between variables must be determined for the purpose of predicting the values of one or more variables on the basis of observations on the other variables.
5. *Hypothesis construction and testing.* Specific statistical hypotheses, formulated in terms of the parameters of multivariate populations, are tested. This may be done to validate assumptions or to reinforce prior convictions.

We conclude this brief overview of multivariate analysis with a quotation from F. H. C. Marriott [19], page 89. The statement was made in a discussion of cluster analysis, but we feel it is appropriate for a broader range of methods. You should keep it in mind whenever you attempt or read about a data analysis. It allows one to maintain a proper perspective and not be overwhelmed by the elegance of some of the theory:

> If the results disagree with informed opinion, do not admit a simple logical interpretation, and do not show up clearly in a graphical presentation, they are probably wrong. There is no magic about numerical methods, and many ways in which they can break down. They are a valuable aid to the interpretation of data, not sausage machines automatically transforming bodies of numbers into packets of scientific fact.

1.2 Applications of Multivariate Techniques

The published applications of multivariate methods have increased tremendously in recent years. It is now difficult to cover the variety of real-world applications of these methods with brief discussions, as we did in earlier editions of this book. However, in order to give some indication of the usefulness of multivariate techniques, we offer the following short descriptions of the results of studies from several disciplines. These descriptions are organized according to the categories of objectives given in the previous section. Of course, many of our examples are multifaceted and could be placed in more than one category.

Data reduction or simplification

- Using data on several variables related to cancer patient responses to radiotherapy, a simple measure of patient response to radiotherapy was constructed. (See Exercise 1.15.)
- Track records from many nations were used to develop an index of performance for both male and female athletes. (See [8] and [22].)
- Multispectral image data collected by a high-altitude scanner were reduced to a form that could be viewed as images (pictures) of a shoreline in two dimensions. (See [23].)
- Data on several variables relating to yield and protein content were used to create an index to select parents of subsequent generations of improved bean plants. (See [13].)
- A matrix of tactic similarities was developed from aggregate data derived from professional mediators. From this matrix the number of dimensions by which professional mediators judge the tactics they use in resolving disputes was determined. (See [21].)

Sorting and grouping

- Data on several variables related to computer use were employed to create clusters of categories of computer jobs that allow a better determination of existing (or planned) computer utilization. (See [2].)
- Measurements of several physiological variables were used to develop a screening procedure that discriminates alcoholics from nonalcoholics. (See [26].)
- Data related to responses to visual stimuli were used to develop a rule for separating people suffering from a multiple-sclerosis-caused visual pathology from those not suffering from the disease. (See Exercise 1.14.)

- The U.S. Internal Revenue Service uses data collected from tax returns to sort taxpayers into two groups: those that will be audited and those that will not. (See [31].)

Investigation of the dependence among variables

- Data on several variables were used to identify factors that were responsible for client success in hiring external consultants. (See [12].)
- Measurements of variables related to innovation, on the one hand, and variables related to the business environment and business organization, on the other hand, were used to discover why some firms are product innovators and some firms are not. (See [3].)
- Measurements of pulp fiber characteristics and subsequent measurements of characteristics of the paper made from them are used to examine the relations between pulp fiber properties and the resulting paper properties. The goal is to determine those fibers that lead to higher quality paper. (See [17].)
- The associations between measures of risk-taking propensity and measures of socioeconomic characteristics for top-level business executives were used to assess the relation between risk-taking behavior and performance. (See [18].)

Prediction

- The associations between test scores, and several high school performance variables, and several college performance variables were used to develop predictors of success in college. (See [10].)
- Data on several variables related to the size distribution of sediments were used to develop rules for predicting different depositional environments. (See [7] and [20].)
- Measurements on several accounting and financial variables were used to develop a method for identifying potentially insolvent property-liability insurers. (See [28].)
- cDNA microarray experiments (gene expression data) are increasingly used to study the molecular variations among cancer tumors. A reliable classification of tumors is essential for successful diagnosis and treatment of cancer. (See [9].)

Hypotheses testing

- Several pollution-related variables were measured to determine whether levels for a large metropolitan area were roughly constant throughout the week, or whether there was a noticeable difference between weekdays and weekends. (See Exercise 1.6.)
- Experimental data on several variables were used to see whether the nature of the instructions makes any difference in perceived risks, as quantified by test scores. (See [27].)
- Data on many variables were used to investigate the differences in structure of American occupations to determine the support for one of two competing sociological theories. (See [16] and [25].)
- Data on several variables were used to determine whether different types of firms in newly industrialized countries exhibited different patterns of innovation. (See [15].)

The preceding descriptions offer glimpses into the use of multivariate methods in widely diverse fields.

1.3 The Organization of Data

Throughout this text, we are going to be concerned with analyzing measurements made on several variables or characteristics. These measurements (commonly called *data*) must frequently be arranged and displayed in various ways. For example, graphs and tabular arrangements are important aids in data analysis. Summary numbers, which quantitatively portray certain features of the data, are also necessary to any description.

We now introduce the preliminary concepts underlying these first steps of data organization.

Arrays

Multivariate data arise whenever an investigator, seeking to understand a social or physical phenomenon, selects a number $p \geq 1$ of *variables* or *characters* to record. The values of these variables are all recorded for each distinct *item, individual*, or *experimental unit*.

We will use the notation x_{jk} to indicate the particular value of the kth variable that is observed on the jth item, or trial. That is,

$$x_{jk} = \text{measurement of the } k\text{th variable on the } j\text{th item}$$

Consequently, n measurements on p variables can be displayed as follows:

	Variable 1	Variable 2	$\cdots$	Variable k	$\cdots$	Variable p
Item 1:	x_{11}	x_{12}	$\cdots$	x_{1k}	$\cdots$	x_{1p}
Item 2:	x_{21}	x_{22}	$\cdots$	x_{2k}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
Item j:	x_{j1}	x_{j2}	$\cdots$	x_{jk}	$\cdots$	x_{jp}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
Item n:	x_{n1}	x_{n2}	$\cdots$	x_{nk}	$\cdots$	x_{np}

Or we can display these data as a rectangular array, called $\mathbf{X}$, of n rows and p columns:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}$$

The array $\mathbf{X}$, then, contains the data consisting of all of the observations on all of the variables.