CHAPTER 1

INTRODUCTION

1.1 WHY MULTIVARIATE ANALYSIS?

Multivariate analysis consists of a collection of methods that can be used when several measurements are made on each individual or object in one or more samples. We will refer to the measurements as *variables* and to the individuals or objects as *units* (research units, sampling units, or experimental units) or *observations*. In practice, multivariate data sets are common, although they are not always analyzed as such. But the exclusive use of univariate procedures with such data is no longer excusable, given the availability of multivariate techniques and inexpensive computing power to carry them out.

Historically, the bulk of applications of multivariate techniques have been in the behavioral and biological sciences. However, interest in multivariate methods has now spread to numerous other fields of investigation. For example, we have collaborated on multivariate problems with researchers in education, chemistry, environmental science, physics, geology, medicine, engineering, law, business, literature, religion, public broadcasting, nursing, mining, linguistics, biology, psychology, and many other fields. Table 1.1 shows some examples of multivariate observations.

Methods of Multivariate Analysis, Third Edition. By Alvin C. Rencher and William F. Christensen

Table 1.1 Examples of Multivariate Data

Units	Variables
1. Students	Several exam scores in a single course
2. Students	Grades in mathematics, history, music, art, physics
3. People	Height, weight, percentage of body fat, resting heart rate
4. Skulls	Length, width, cranial capacity
5. Companies	Expenditures for advertising, labor, raw materials
6. Manufactured items	Various measurements to check on compliance with specifications
7. Applicants for bank loans	Income, education level, length of residence, savings account, current debt load
8. Segments of literature	Sentence length, frequency of usage of certain words and style characteristics
9. Human hairs	Composition of various elements
10. Birds	Lengths of various bones

The reader will notice that in some cases all the variables are measured in the same scale (see 1 and 2 in Table 1.1). In other cases, measurements are in different scales (see 3 in Table 1.1). In a few techniques such as profile analysis (Sections 5.9 and 6.8), the variables must be commensurate, that is, similar in scale of measurement; however, most multivariate methods do not require this.

Ordinarily the variables are measured simultaneously on each sampling unit. Typically, these variables are correlated. If this were not so, there would be little use for many of the techniques of multivariate analysis. We need to untangle the overlapping information provided by correlated variables and peer beneath the surface to see the underlying structure. Thus the goal of many multivariate approaches is *simplification*. We seek to express "what is going on" in terms of a reduced set of dimensions. Such multivariate techniques are *exploratory*; they essentially generate hypotheses rather than test them.

On the other hand, if our goal is a formal hypothesis test, we need a technique that will (1) allow several variables to be tested and still preserve the significance level and (2) do this for any intercorrelation structure of the variables. Many such tests are available.

As the two preceding paragraphs imply, multivariate analysis is concerned generally with two areas, *descriptive* and *inferential* statistics. In the descriptive realm, we often obtain optimal linear combinations of variables. The optimality criterion varies from one technique to another, depending on the goal in each case. Although linear combinations may seem too simple to reveal the underlying structure, we use them for two obvious reasons: (1) mathematical tractability (linear approximations are used throughout all science for the same reason) and (2) they often perform well in practice. These linear functions may also be useful as a follow-up to inferential procedures. When we have a statistically significant test result that compares sev-

eral groups, for example, we can find the linear combination (or combinations) of variables that led to rejection. Then the contribution of each variable to these linear combinations is of interest.

In the inferential area, many multivariate techniques are extensions of univariate procedures. In such cases we review the univariate procedure before presenting the analogous multivariate approach.

Multivariate inference is especially useful in curbing the researcher's natural tendency to read too much into the data. Total control is provided for experimentwise error rate; that is, no matter how many variables are tested simultaneously, the value of α (the significance level) remains at the level set by the researcher.

Some authors warn against applying the common multivariate techniques to data for which the measurement scale is not interval or ratio. It has been found, however, that many multivariate techniques give reliable results when applied to ordinal data.

For many years the applications lagged behind the theory because the computations were beyond the power of the available desk-top calculators. However, with modern computers, virtually any analysis one desires, no matter how many variables or observations are involved, can be quickly and easily carried out. Perhaps it is not premature to say that multivariate analysis has come of age.

1.2 PREREQUISITES

The mathematical prerequisite for reading this book is matrix algebra. Calculus is not used [with a brief exception in equation (4.29)]. But the basic tools of matrix algebra are essential, and the presentation in Chapter 2 is intended to be sufficiently complete so that the reader with no previous experience can master matrix manipulation up to the level required in this book.

The statistical prerequisites are basic familiarity with the normal distribution, t-tests, confidence intervals, multiple regression, and analysis of variance. These techniques are reviewed as each is extended to the analogous multivariate procedure.

This is a multivariate methods text. Most of the results are given without proof. In a few cases proofs are provided, but the major emphasis is on heuristic explanations. Our goal is an intuitive grasp of multivariate analysis, in the same mode as other statistical methods courses. Some problems are algebraic in nature, but the majority involve data sets to be analyzed.

1.3 OBJECTIVES

We have formulated three objectives that we hope this book will achieve for the reader. These objectives are based on long experience teaching a course in multivariate methods, consulting on multivariate problems with researchers in many fields, and guiding statistics graduate students as they consulted with similar clients.

The first objective is to gain a thorough understanding of the details of various multivariate techniques, their purposes, their assumptions, their limitations, and so

on. Many of these techniques are related, yet they differ in some essential ways. These similarities and differences are emphasized.

The second objective is to be able to select one or more appropriate techniques for a given multivariate data set. Recognizing the essential nature of a multivariate data set is the first step in a meaningful analysis. Basic types of multivariate data are introduced in Section 1.4.

The third objective is to be able to interpret the results of a computer analysis of a multivariate data set. Reading the manual for a particular program package is not enough to make an intelligent appraisal of the output. Achievement of the first objective and practice on data sets in the text should help achieve the third objective.

1.4 BASIC TYPES OF DATA AND ANALYSIS

We will list four basic types of (continuous) multivariate data and then briefly describe some possible analyses. Some writers would consider this an oversimplification and might prefer elaborate tree diagrams of data structure. However, many data sets can fit into one of these categories, and the simplicity of this structure makes it easier to remember. The four basic data types are as follows:

1. A single sample with several variables measured on each sampling unit (subject or object).
2. A single sample with two sets of variables measured on each unit.
3. Two samples with several variables measured on each unit.
4. Three or more samples with several variables measured on each unit.

Each data type has extensions, and various combinations of the four are possible. A few examples of analyses for each case will now be given:

1. A single sample with several variables measured on each sampling unit:
   - **a.** Test the hypothesis that the means of the variables have specified values.
   - **b.** Test the hypothesis that the variables are uncorrelated and have a common variance.
   - **c.** Find a small set of linear combinations of the original variables that summarizes most of the variation in the data (principal components).
   - **d.** Express the original variables as linear functions of a smaller set of underlying variables that account for the original variables and their intercorrelations (factor analysis).
2. A single sample with two sets of variables measured on each unit:
   - **a.** Determine the number, the size, and the nature of relationships between the two sets of variables (canonical correlation). For example, we may

wish to relate a set of interest variables to a set of achievement variables. How much overall correlation is there between these two sets?

b. Find a model to predict one set of variables from the other set (multivariate multiple regression).

3. Two samples with several variables measured on each unit:

a. Compare the means of the variables across the two samples (Hotelling's T^2-test).

b. Find a linear combination of the variables that best separates the two samples (discriminant analysis).

c. Find a function of the variables that will accurately allocate the units into the two groups (classification analysis).

4. Three or more samples with several variables measured on each unit:

a. Compare the means of the variables across the groups (multivariate analysis of variance).

b. Extension of 3b to more than two groups.

c. Extension of 3c to more than two groups.