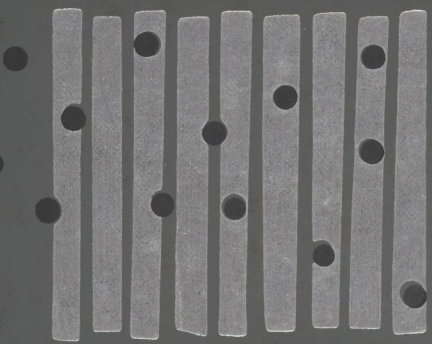
MULTIVARIATE STATISTICAL METHODS



MULTIVARIATE STATISTICAL METHODS

Donald F. Morrison
Professor of Statistics
The Wharton School
University of Pennsylvania

Second Edition

McGRAW-HILL BOOK COMPANY
New York St. Louis San Francisco Auckland Düsseldorf Johannesburg
Kuala Lumpur London Mexico Montreal New Delhi Panama Paris
São Paulo Singapore Sydney Tokoyo Toronto

MULTIVARIATE STATISTICAL METHODS

2 3 4 5 6 7 8 9 0 DODO 7 9 8 7 6

This book was set in Modern by Bi-Comp, Incorporated.
The editors were A. Anthony Arthur and Michael Gardner;
the production supervisor was Leroy A. Young.
R. R. Donnelley & Sons Company was printer and binder.

Library of Congress Cataloging in Publication Data

Morrison, Donald F
Multivariate statistical methods.

(McGraw-Hill series in probability and statistics)
Bibliography: p.
1. Multivariate analysis. I. Title.
QA278.M68 1976 519.5'3 75-14325
ISBN 0-07-043186-8

MULTIVARIATE STATISTICAL METHODS

McGRAW-HILL SERIES IN PROBABILITY AND STATISTICS

DAVID BLACKWELL AND HERBERT SOLOMON, *Consulting Editors*

BHARUCHA-REID: Elements of the Theory of Markov Processes and Their Applications
DE GROOT: Optimal Statistical Decisions
DRAKE: Fundamentals of Applied Probability Theory
EHRENFELD AND LITTAUER: Introduction to Statistical Methods
GIBBONS: Nonparametric Statistical Inference
GRAYBILL: Introduction to Linear Statistical Models
HODGES, KRECH, AND CRUTCHFIELD: StatLab: An Empirical Introduction to Statistics
LI: Introduction to Experimental Statistics
MOOD, GRAYBILL, AND BOES: Introduction to the Theory of Statistics
MORRISON: Multivariate Statistical Methods
RAJ: The Design of Sample Surveys
RAJ: Sampling Theory
STEEL AND TORRIE: Introduction to Statistics
THOMASIAN: The Structure of Probability Theory with Applications
WADSWORTH AND BRYAN: Applications of Probability and Random Variables
WASAN: Parametric Estimation
WOLF: Elements of Probability and Statistics

To My Mother and
the Memory of My Father

PREFACE TO THE SECOND EDITION

This edition was prepared to expand the treatments of certain important methods and to maintain the currency of the citations to the multivariate statistical literature. In this revision emphasis was given to methods in the mainstream of classical multivariate analysis which were concerned with mean structures rather than models for covariance matrices.

Among the major changes in the book are some elementary results on estimates and tests with incomplete data matrices; references have also been provided for more general missing-data techniques. The multivariate analysis of covariance has been given in greater detail, and a new section on the fitting of growth curves has been included in Chapter 5. Linear discrimination has been accorded a whole chapter. Some recent results on the estimation of error rates for the two-sample case have been summarized, and classification rules for several groups have been developed by discriminant functions and minimum-distance rules. The hypothesis tests on covariance matrices have been extended to the single-sample case and to patterns useful in the analysis of repeated measurements. The Appendix charts of the greatest-characteristic-root percentage points have been augmented with tables for the parameter s through twenty.

In the revision I was encouraged by the reception of the first edition as a reference and text. It has been especially heartening that adoptions have ranged outside the intended behavioral and life sciences audience to include courses in business, economics, and the social sciences. I am indebted to many persons who have used the book for writing about its content, or for calling my attention to errors or ambiguities.

I am grateful to E. S. Pearson and K. C. S. Pillai for permitting the reproduction of Appendix Tables 6 to 14 from *Biometrika*. The inclusion of those critical values has greatly extended the usefulness of the largest-root tests and confidence statements. I am also most appreciative of the investigators and editors who have granted permission for the use of their data in examples and exercises. In the planning of the second edition I was aided by thoughtful and detailed comments from Leon Jay Gleser on the initial outline.

Finally, a special acknowledgment should be made to my wife Phyllis for her support and assistance with the manuscript, and to our son Norman for giving up time that belonged to him.

Donald F. Morrison

PREFACE TO THE FIRST EDITION

Multivariate statistical analysis is concerned with data collected on several dimensions of the same individual. Such observations are common in the social, behavioral, life, and medical sciences: the record of the prices of a commodity, the reaction times of a normal subject to several different stimulus displays, the principal bodily dimensions of an organism, or a set of blood-chemistry values from the same patient are all examples of multidimensional data. As in univariate statistics, we shall assume that a random sample of multicomponent observations has been collected from different individuals or other independent sampling units. However, the common source of each individual observation will generally lead to dependence or correlation among the dimensions, and it is this feature that distinguishes multivariate data and techniques from their univariate prototypes.

This book was written to provide investigators in the life and behavioral sciences with an elementary source for multivariate techniques which appeared to be especially useful for the design and analysis of their experimental data. The book has also been organized to serve as the text for a course in multivariate methods at the advanced undergraduate or graduate level in the sciences. The mathematical and statistical prerequisites are minimal: a semester course in elementary statistics with a survey of the fundamental sampling distributions and an exposure to the calculus for the partial differentiations and integrals required for occasional maximizations and expectations should be sufficient. The review of the essential univariate statistical concepts in the first chapter and a detailed treatment of matrix algebra in the second make the book fairly self-contained both as a reference and a text. The standard results on the multinormal distribution, the estimation of its parameters, and correlation analysis in Chapter 3 are essential background for the developments in the remaining chapters.

The selection of techniques reflects my experiential biases and preferences. Attention has been restricted to continuous observations from multivariate normal populations: no mention has been made of the newer distribution-free tests and the methods for analysis of many-way categorical data tables. It was felt that the implications of the T^2 statistic for repeated-measurements experiments justified a lengthy discussion of tests and confidence intervals for mean vectors. The multivariate general linear hypothesis and analysis of variance has been developed through the Roy union-intersection principle for the natural ease with which simultaneous confidence statements can be obtained. In my experience the Hotelling principal-component technique has proved to be

exceedingly useful for data reduction, analysis of the latent structure of multivariate systems, and descriptive purposes, and its use and properties are developed at length. My approach to factor analysis has been statistical rather than psychometric, for I prefer to think of the initial steps, at least, of a factor analysis as a problem in statistical estimation.

For the preparation of this methods text I wish to acknowledge a considerable debt to those responsible for the theoretical development of multivariate analysis: the fundamental contributions of T. W. Anderson, Harold Hotelling, D. N. Lawley, and the late S. N. Roy are evident throughout. In particular the frequent references to S. N. Roy's monograph "Some Aspects of Multivariate Analysis" are indicative of his influence on the presentation. For the many derivations beyond the level of this book the reader has usually been referred to T. W. Anderson's standard theoretical source "An Introduction to Multivariate Statistical Analysis."

It is a pleasure to acknowledge those who have assisted at different stages in the preparation of this book. My thanks are due to Samuel W. Greenhouse for initially encouraging me to undertake the project. I am especially indebted to Karen D. Pettigrew and John J. Bartko for their thoughtful reading of several chapters and for offering suggestions that have improved the clarity of the presentation. George Schink carefully checked the computations of the majority of the examples. However, the ultimate responsibility for the nature and accuracy of the contents must of course rest with the author. Finally, I wish to express my gratitude to the many investigators who graciously permitted the use of their original and published data for the examples and exercises.

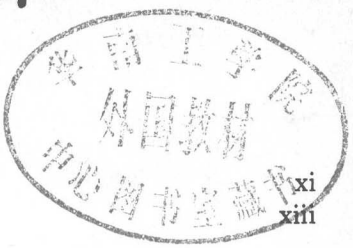
I am indebted to A. M. Mood and the McGraw-Hill Book Company for permission to reproduce Table 1 from the first edition of "Introduction to the Theory of Statistics." Tables 2 and 4 have been abridged from tables originally prepared by Catherine M. Thompson and Maxine Merrington, and have been reproduced with the kind permission of the editor of *Biometrika*, E. S. Pearson. I am also grateful to Professor Pearson and to H. O. Hartley for kindly permitting the reproduction of Charts 1 to 8 from *Biometrika*. I am indebted to the literary executor of the late Sir Ronald A. Fisher, F.R.S., Cambridge, to Dr. Frank Yates, F.R.S., Rothamsted, and to Messrs. Oliver & Boyd Ltd., Edinburgh, for permission to reprint Table 3 from their book "Statistical Tables for Biological, Agricultural, and Medical Research." Charts 9 to 16 have been reproduced from the *Annals of Mathematical Statistics* with the kind permission of D. L. Heck and the managing editor, P. L. Meyer.

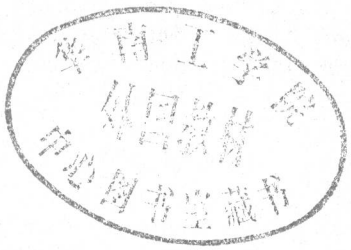
The preparation of parts of an earlier version of the text as class notes was made possible through the enthusiastic cooperation of the Foundation for Advanced Education in the Sciences, Inc., Bethesda, Md.

Support for the use of some chapters in mimeographed form and clerical assistance was kindly provided by Dean Willis J. Winn through funds from a grant to the Wharton School of Finance and Commerce by the New York Life Insurance Co. I am also grateful for the secretarial assistance furnished by the Department of Statistics and Operations Research and the Lecture Note Fund of the University of Pennsylvania.

Donald F. Morrison

CONTENTS





1
SOME ELEMENTARY STATISTICAL CONCEPTS

1.1 **INTRODUCTION.** In this chapter we shall summarize some important parts of univariate statistical theory to which we shall frequently refer in our development of multivariate methods. Certain concepts of statistical inference will be introduced, and some essential univariate distributions will be described. We shall assume that the reader has been exposed to the elements of probability and random variables and has an acquaintance with the basic univariate techniques as applied in some substantive discipline.

1.2 RANDOM VARIABLES

Every statistical analysis must be built upon a *mathematical model* linking observable reality with the mechanism generating the observations. This model should be a parsimonious description of nature: its functional form should be simple, and the number of its parameters and components should be a minimum. The model should be *parametrized* in such a way that each parameter can be interpreted easily and identified with some aspect of reality. The functional form should be sufficiently tractable to permit the sort of mathematical manipulations required for the estimation of its parameters and other inferences about its nature.

Mathematical models may be divided into three general classes: (1) purely deterministic, (2) static, or deterministic with simple random components, and (3) stochastic. Any observation from a deterministic model is strictly a function of its parameters and such variables as time, space, or inputs of energy or a stimulus. Newtonian physics states that the distance traveled by a falling object is directly related to the

squared time of fall, and if atmospheric turbulence, observer error, and other transient effects can be ignored, the displacement can be calculated exactly for a given time and gravitational constant. In the second kind of model each observation is a function of a strictly deterministic component and a random term ascribable to measurement error or sampling variation in either the observed response or the input variables. The random components are assumed to be independent of one another for different observations. The models we shall encounter in the sequel will be mainly of this class, with the further restriction that the random component will merely be added to the deterministic part. Stochastic models are constructed from fundamental random events or components to explain dynamic or evolutionary phenomena: they range in complexity from the case of a sequence of Bernoulli trials as the model for a coin-tossing experiment to the birth-and-death process describing the size of a biological population. Most stochastic models allow for a "memory" effect, so that each observed response is dependent to some degree upon its predecessors in time or neighbors in space. We shall touch only tangentially on this kind of model.

Now let us define more precisely what is meant by the notions of random variation or the random components in the second and third kinds of models. We shall begin by defining a *discrete random variable*, or one which can assume only a countable number of values. Suppose that some experiment can result in exactly one of k outcomes $E_1, \ldots, E_k$. These outcomes are mutually exclusive, in the sense that the occurrence of one event precludes that of any other. To every event we assign some number p_i between zero and one called the *probability* $P(E_i)$ of that event. p_i is the probability that in a single trial of the experiment the outcome E_i will occur. Within the framework of our experiment we assign a probability of zero to impossible events and a probability of unity to any event which must happen with certainty. Then, by the mutual exclusiveness of the events, in a single trial

$$P(E_i \cap E_j) = 0$$

$$P(E_i \cup E_j) = p_i + p_j$$

where the *intersection* symbol $\cap$ denotes the event "E_i and E_j" and the *union* symbol $\cup$ indicates the event "E_i and/or E_j." By the additive property of the probabilities of mutually exclusive outcomes the total probability of the set of events is

$$P(E_1 \cup \cdots \cup E_k) = p_1 + \cdots + p_k$$

$$= 1$$

Now assign the numerical value x_i to the ith outcome, where for con-

venience the outcomes have been placed in ascending order according to their x_i values. The discrete random variable X is defined as that quantity which takes on the value x_i with probability p_i at each trial of the experiment. As an example, if the experiment consists of the toss of a coin, the score of one might be assigned to the outcome heads, while zero might be the tails score. Then $x_1 = 0$, $x_2 = 1$, and $p_1 = 1 - p$, $p_2 = p$, say. This random variable would be described by its *probability function* $f(x_i)$ specifying the probabilities with which X assumes the values 0 and 1:

X	$f(x_i)$
0	$1 - p$
1	p

We note that the total probability is unity and that we have implicitly assigned a probability of zero to such irrelevant events as the coin's landing on edge or rolling out of sight. We have chosen *not* to assign a numerical value to the single parameter p; this reflects the intrinsic qualities of the coin as well as the manner in which it is tossed. It is only for convenience or for lack of knowledge of the coin's properties that p is ever taken as $\frac{1}{2}$.

The random variables we shall encounter in the sequel will take on values over some continuous region rather than a set of countable events and will be called *continuous random variables* or *continuous variates*. Both terms will be used synonymously. The continuous random variable X defined on the domain of real numbers is characterized by its *distribution function*

$$F(x) = P(X \leq x) \qquad -\infty < x < \infty \tag{1}$$

giving the probability that X is less than or equal to some value x of its domain. Since X is continuous, $P(X = x) = 0$. If $F(x)$ is an absolutely continuous function, the continuous analogue of the discrete probability function is the *density function*

$$f(x) = \frac{dF(x)}{dx} \tag{2}$$

Conversely, by the absolute-continuity property,

$$F(x) = \int_{-\infty}^{x} f(u)\, du \tag{3}$$

and from this integral definition follows the equivalent term *cumulative distribution function* for $F(x)$. Note that these definitions are perfectly general: if the random variable is defined only on some interval of the real line, outside that interval $f(x)$ is defined to be zero, and to the left and right of the interval $F(x)$ is zero and one, respectively. When weighted in proportion to their density function $f(x)$, the values on the